Experimentology

An Open Science Approach to Experimental Psychology Methods

Michael C. Frank Mika Braginsky Julie Cachia
Nicholas A. Coles Tom E. Hardwicke
Robert D. Hawkins Maya B. Mathur
Rondeline Williams

2024-01-25

Contents

Preface

As scientists and practitioners, we often want to create generalizable, causal theories of human behavior. As it turns out, experiments – in which we use random assignment to measure a causal effect – are an unreasonably effective tool to help with this task. But how should we go about doing good experiments?

This book provides an introduction to the workflow of the experimental researcher working in psychology or the behavioral sciences more broadly. The organization of the book is sequential, from the planning stages of the research process through design, data gathering, analysis, and reporting. We introduce these concepts via narrative examples from a range of sub-disciplines, including cognitive, developmental, and social psychology. Throughout, we also illustrate the pitfalls that led to the "replication crisis" in psychology.

To help researchers avoid these pitfalls, we advocate for an open-science based approach in which transparency is integral to the entire experimental workflow. We provide readers with guidance for preregistration, project management, data sharing, and reproducible report writing.

The story of this book

Experimental Methods (Psych 251) is the foundational course for incoming graduate students in the Stanford psychology department. The course goal is to orient students to the nuts and bolts of doing behavioral experiments, including how to plan and design a solid experiment and how to avoid common pitfalls regarding design, measurement, and sampling.

Almost all student coursework both before and in graduate school deals with the content of their research, including theories and results in their areas of focus. In contrast, our course is sometimes the only one that deals with the *process* of research, from big questions about why we do experiments and what it means to make a causal inference, all the way

to the tiny details of project organization, like what to name your directories and how to make sure you don't lose your data in a computer crash.

This observation leads to our book's title. "Experimentology" is the set of practices, findings, and approaches that enable the construction of robust, precise, and generalizable experiments.

The centerpiece of the Experimental Methods course is a replication project, reflecting a teaching model first described in Frank and Saxe (2012)[1] and later expanded on in Hawkins, Smith et al. (2018).[2] Each student chooses a published experiment in the literature and collects new data on a pre-registered version of the same experimental paradigm, comparing their result to the original publication. Over the course of the quarter, we walk through how to set up a replication experiment, how to pre-register confirmatory analyses, and how to write a reproducible report on the findings. The project teaches concepts like reliability and validity, which allow students to analyze choices that the original experimenters made – often choices that could have been made differently in hindsight!

[1] Frank, Michael C, and Rebecca Saxe. 2012. "Teaching Replication." Perspectives on Psychological Science 7: 595–99.

[2] Hawkins, Robert D, Eric N Smith, Carolyn Au, Juan Miguel Arias, Rhia Catapano, Eric Hermann, Martin Keil, et al. 2018. "Improving the Replicability of Psychological Science Through Pedagogy." Advances in Methods and Practices in Psychological Science 1 (1): 7–18.

At the end of the course, we reap the harvest of these projects. The project presentations are a wonderful demonstration of both how much the students can accomplish in a quarter and also how tricky it can be to reproduce (redo calculations in the original data) and replicate (recover similar results in new data) the published literature. Often our replication success rate for the course hovers just above 50%, an outcome that can be disturbing or distressing for students who assume that the published literature reports the absolute truth.

This book is an attempt to distill some of the lessons of the course (and students' course projects) into a textbook. We'll tell the story of the major shifts in psychology that have come about in the last ten years, including both the "replication crisis" and the positive methodological reforms that have resulted from it. Using this story as motivation, we will highlight the importance of transparency during all aspects of the experimental process from planning to dissemination of materials, data, and code.

What this book is and isn't about

This book is about psychology experiments. These will be typically be short studies conducted online or in a single visit to a lab, often – though certainly not always – with a convenience sample. When we say "experiments" here we mean **randomized experiments** where some aspect

of the participants' experience is **manipulated** by the experimenter and then some outcome variable is **measured**.[3]

[3] We use **bold** to indicate the introduction of new technical terms.

The central thesis of the book is that:

> Experiments are intended to make maximally unbiased, generalizable, and precise estimates of specific causal effects.

We'll explore the implications of this thesis for a host of topics, including causal inference, experimental design, measurement, sampling, preregistration, data analysis, and many others.

Because our focus is on experiments, we won't be talking much about observational designs, survey methods, or qualitative research; these are important tools and appropriate for a whole host of questions, but they aren't our focus here. We also won't go into depth about the many fascinating methodological and statistical issues brought up by single-participant case studies, longitudinal research, field studies, or other methodological variants. Many of the concerns we raise are still important for these types of studies, but some of our advice won't transfer to these less common designs.

Even for students who are working on non-experimental research, we expect that a substantial part of the book content will still be useful, including chapters on replication (Chapter 3), ethics (Chapter 4), statistics (Chapters 5, 6, 7), sampling (Chapter 10), project management (Chapter 13), and reporting (Chapters 14, 15, 16).

In our writing, we presuppose that readers have some background in psychology, at least at an introductory level. In addition, although we introduce a number of statistical topics, readers might find these sections more accessible with an undergraduate statistics course under their belt. Finally, our examples are written in the R statistical programming language, and for chapters on statistics and visualization especially (Chapters 5, 6, 7, 15, 16), some familiarity with R will be helpful for understanding the code. None of these prerequisites are necessary to read the book, but we offer them so that readers can calibrate their expectations.

How to use this book

The book is organized into five main parts, mirroring the timeline of an experiment: 1) Foundations, 2) Statistics, 3) Planning, 4) Execution, and 5) Reporting. We hope that this organization makes it well-suited for teaching or for use as a reference book.[4]

[4] If you are an instructor who is planning to adopt the book for a course, you might be interested in our resources for instructors, including sample course schedules, in Appendix A.

The book is designed for a course for graduate students or advanced undergraduates, but the material is also suitable for self-study by anyone interested in experimental methods, whether in academic psychology or any other context – in our out of academia – in which behavioral experimentation is relevant. We also hope that some readers will come to particular chapters of the book because of an interest in specific topics like measurement (Chapter 8) or sampling (Chapter 10) and will be able to use those chapters as standalone references. And finally, for those interested in the "replication crisis" and subsequent reforms, Chapters 3, 11, and 13 will be especially interesting.

Ultimately, we want to give you what you need to plan and execute your own study! Instead of enumerating different approaches, we try to provide a single coherent – and often quite opinionated – perspective, using marginal notes and references to give pointers to more advanced materials or alternative approaches. Throughout, we offer:

- **Case studies** that illustrate the central concepts of a chapter,
- **Accident reports** describing examples where poor research practices led to issues in the literature, and
- **Depth boxes** providing simulations, linkages to advanced techniques, or more nuanced discussion.

While case studies are often integral to the chapters, the other boxes can typically be skipped without issue.

Themes

We highlight four major cross-cutting themes for the book: TRANSPARENCY, MEASUREMENT PRECISION, BIAS REDUCTION, and GENERALIZABILITY.[5]

[5] Themes are noted using SMALL CAPS.

- TRANSPARENCY: For experiments to be reproducible, other researchers need to be able to determine exactly what you did. Thus, every stage of the research process should be guided by a primary concern for transparency. For example, preregistration creates transparency into the researcher's evolving expectations and thought processes; releasing open materials and analysis scripts creates transparency into the details of the procedure.
- MEASUREMENT PRECISION: We want researchers to start planning an experiment by thinking "what causal effect do I want to measure" and to make planning, sampling, design, and analytic choices that maximize the precision of this measurement. A downstream consequence of this mindset is that we move away

from a focus on dichotomized inferences about statistical significance and towards analytic and meta-analytic models that focus on continuous effect sizes and confidence intervals.
- BIAS REDUCTION: While precision refers to random error in a measurement, measurements also have systematic sources of error that bias them away from the true quantity. In our samples, analyses, experimental designs, and in the literature, we need to think carefully about sources of bias in the quantity being estimated.
- GENERALIZABILITY: Complex behaviors are rarely universal across all settings and populations, and any given experiment can only hope to cover a small slice of the possible conditions where a behavior of interest takes place. Psychologists must therefore consider the generalizability of their findings at every stage of the process, from stimulus selection and sampling procedures, to analytic methods and reporting.

Throughout the book, we will return to these four themes again and again as we discuss how the decisions made by the experimenter at every stage of design, data gathering, and analysis bear on the inferences that can be made about the results. The introduction of each chapter highlights connections to specific themes.

The software toolkit for this book

We introduce and advocate for an approach to reproducible study planning, analysis, and writing. This approach depends on an ecosystem of open-source software tools, which we introduce in the book's appendices.[6]

- The R statistical programming language and the RStudio[7] integrated development environment,
- Version control using `git` and GitHub[8], allowing collaboration on text documents like code, prose, and data, storing and integrating contributions over time (Appendix B),
- The `RMarkdown` and `Quarto` tools for creating reproducible reports that can be rendered to a variety of formats (Appendix C),
- The `tidyverse` family of R packages, which extend the basic functionality of R with simple tools for data wrangling, analysis, and visualization (Appendix D), and
- The `ggplot2` plotting package, which makes it easy to create flexible data visualizations for both confirmatory and exploratory data analyses (Appendix E).

[6] These appendices are available online at https://experimentology.io but not in the print version of the book, since their content is best viewed in the web format.

[7] https://posit.co/download/rstudio-desktop/

[8] https://github.com/

Where appropriate, we provide **code boxes** that show the specific `R` code used to create our examples.

Onward!

Thanks for joining us for Experimentology! Whether you are casually browsing, doing readings for a course, or using the book as a reference in your own experimental work, we hope you find it useful. Throughout, we have tried to practice what we preach in terms of reproducibility, and so the full source code for the book is available at https://github.com/langcog/experimentology. We encourage you to browse, comment, and log issues or suggestions.[9]

[9] The best way to give us specific feedback is to create an issue on our github page at https://github.com/langcog/experimentology/issues.

Acknowledgments

Thanks first and foremost to the many generations of students and TAs in Stanford Psych 251, who have collectively influenced the content of this book through their questions and interests.

Thanks to the staff at MIT Press, especially Philip Laughlin and Amy Brand, for embracing a vision of a completely open web textbook that is also reviewed and published through a traditional press. We appreciate your support and flexibility.

We adapt the Contributor Roles (CRediT) Taxonomy[10] to describe our contributions to this manuscript, and we recommend that you do so in your work as well.

- Preface
  - Primary writer: Michael C. Frank
  - Editor: Tom E. Hardwicke
- Chapter 1
  - Primary writer: Michael C. Frank
  - Co-writer: Nicholas Coles
  - Editor: Tom E. Hardwicke
- Chapter 2
  - Primary writer: Michael C. Frank
  - Editor: Nicholas Coles, Tom E. Hardwicke
- Chapter 3
  - Primary writer: Michael C. Frank

[10] Learn more at https://credit.niso.org/.

    - Editor: Maya B. Mathur, Tom E. Hardwicke, Nicholas Coles
- Chapter 4
    - Primary writer: Rondeline Williams
    - Co-writer: Michael C. Frank
    - Editor: Tom E. Hardwicke, Julie Cachia
- Chapter 5
    - Co-writer: Maya B. Mathur, Nicholas Coles, Michael C. Frank
    - Editor: Julie Cachia, Tom E. Hardwicke
- Chapter 6
    - Primary writer: Michael C. Frank
    - Co-writer: Maya B. Mathur
    - Editor: Julie Cachia, Tom E. Hardwicke
- Chapter 7
    - Co-writer: Maya B. Mathur, Michael C. Frank
    - Editor: Tom E. Hardwicke, Robert D. Hawkins
- Chapter 8
    - Primary writer: Michael C. Frank
    - Editor: Robert D. Hawkins, Tom E. Hardwicke, Rondeline Williams
- Chapter 9
    - Primary writer: Michael C. Frank
    - Editor: Nicholas Coles, Tom E. Hardwicke
- Chapter 10
    - Primary writer: Michael C. Frank
    - Editor: Julie Cachia, Tom E. Hardwicke, Maya B. Mathur
- Chapter 11
    - Primary writer: Tom E. Hardwicke
    - Editor: Michael C. Frank
- Chapter 12
    - Co-writer: Rondeline Williams, Michael C. Frank
    - Editor: Tom E. Hardwicke
- Chapter 13

    - Primary writer: Michael C. Frank
    - Editor: Tom E. Hardwicke
- Chapter 15
    - Primary writer: Robert D. Hawkins
    - Editor: Michael C. Frank, Tom E. Hardwicke, Mika Braginsky
- Chapter 14
    - Primary writer: Tom E. Hardwicke
    - Editor: Michael C. Frank
- Chapter 16
    - Co-writer: Nicholas Coles, Maya B. Mathur
    - Editor: Michael C. Frank, Tom E. Hardwicke
- Conclusion
    - Primary writer: Michael C. Frank
    - Editor: Tom E. Hardwicke
- Appendix A
    - Primary writer: Julie Cachia
    - Editor: Michael C. Frank
- Appendix B
    - Primary writer: Julie Cachia
    - Editor: Michael C. Frank
- Appendix C
    - Primary writer: Michael C. Frank
    - Editor: Julie Cachia
- Appendix D
    - Primary writer: Michael C. Frank
    - Editor: Julie Cachia, Mika Braginsky
- Appendix E
    - Primary writer: Michael C. Frank
    - Editor: Julie Cachia, Mika Braginsky
- Technical infrastructure
    - Developer: Mika Braginsky, Natalie Braginsky

Part I

Foundations

1 Experiments

Learning goals

- Define what an experiment is
- Contrast observational and experimental studies using causal graphs
- Understand the role of randomization in experiments
- Consider constraints on the generalizability of experiments

Welcome to Experimentology! This is a book all about the art of running experiments in psychology. Throughout, we will be guided by a simple idea:

> The purpose of experiments is to estimate the magnitude of causal effects.[1]

Starting from our core idea, we'll provide advice about how to navigate things like experimental design, measurement, sampling, and more. Our decisions about each of these will determine how precise our estimate is, and whether it is subject to bias. But before we get to those topics, let's start by thinking about *why* we might do an experiment, a topic that will intersect with our key themes of BIAS REDUCTION and GENERALIZABILITY.

[1] Perhaps you're already saying, "that's not what I thought experiments were for! I thought they were for testing hypotheses." Bear with us and we hope we'll convince you that our definition is a bit more general, and that testing a hypothesis is one thing you can do with a measurement.

1.1 Observational studies don't reveal causality

If you're reading this book, there's probably something about psychology you want to understand. How is language learned? How is it that we experience emotions like happiness and sadness? Why do humans sometimes work together and other times destroy one another? When psychologists study these centuries-old questions, they often transform them into questions about **causality**.[2]

1.1.1 Describing causal relationships

Consider the age-old question: does money make people happy? This question is – at its heart – a question about what interventions on the

[2] Defining causality is one of the trickiest and oldest problems in philosophy, and we won't attempt to solve it here! But from a psychological perspective, we're fond of Lewis (1973)'s "counterfactual" analysis of causality. On this view, we can understand the claim that *money causes happiness* by considering a scenario where if people *hadn't* been given more money, they *wouldn't* have experienced an increase in happiness.

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 one 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!

Let's get a bit more precise about our causal hypothesis. To illustrate causal relationships, we can use a tool called **directed acyclic graphs** (DAGs, Pearl 1998). Figure 1.1 shows an example of a DAG for money and happiness: the arrow represents our idea about the potential causal link between two variables: money and happiness.[3] The direction of the arrow tells us which way we hypothesize that the causal relationship goes.

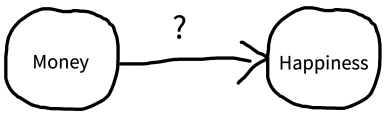


Figure 1.1: The hypothesized causal effect of money on happiness.

[3] 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.

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.

1.1.2 The problems of directionality and confounding

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.

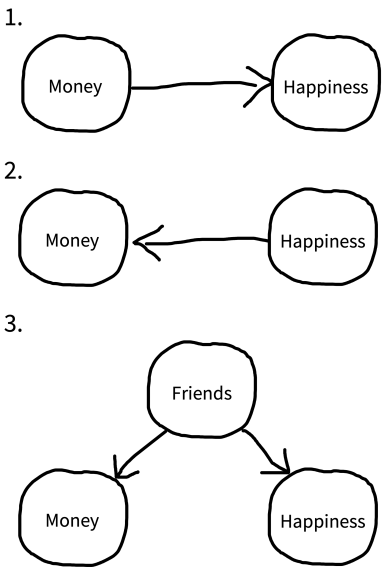


Figure 1.2: Three reasons why money and happiness can be correlated.

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 as well.[4]

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 for reasoning about causal relationships from observational datasets.[5] 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

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 vs. 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** –

[4] 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.

[5] 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.

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 1882).[6] 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 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.

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 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 we can "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 education 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?[7]

[6] 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!

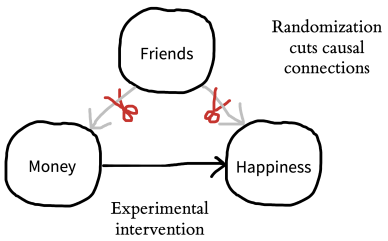


Figure 1.3: In principle, experiments allows us to "snip away" the friend confound by holding it constant (though in practice, it can be tough to figure out how to hold something constant when you are talking about people as your unit of study).

[7] 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, Lawson, and Rhemtulla 2022).

1.2.2 *Randomization saves the day*

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).

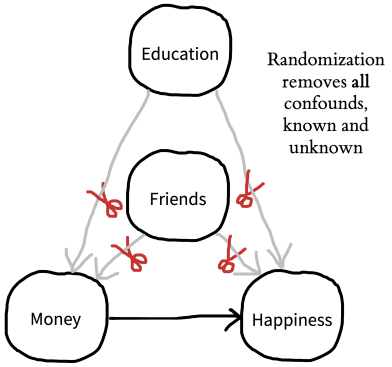


Figure 1.4: If you randomly split a large group of people into groups, the groups will, on average, be equal in every way.

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 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.[8]

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, education attainment, and sleep quality. But just as you can flip a coin and sometimes get heads 9 out of 10 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 stil 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.[9]

[8] 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.

[9] 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.

Depth

Unhappy randomization?

As we've been discussing, random assignment removes confounding by ensuring that – on average – groups are equivalent with respect to all of their characteristics. Equivalence for any *particular* random assignment is more likely the larger your sample is, however. Any individual experiment may be affected by **unhappy randomization**, when a particular confound is unbalanced between groups by chance.

Unhappy randomization is much more common in small experiments than larger ones. To see why, we use a technique called **simulation**. In simulations, we invent data randomly following a set of assumptions: we make up a group of participants and generate their characteristics and their condition assignments. By varying the assumptions we use, we can investigate how particular choices might change the structure of the data.

To look at unhappy randomization, we created many simulated versions of our money-happiness experiment, in which an experimental group receives money and the control group receives none, and then happiness is measured for both groups. We assume that each participant has a set number of friends, and that the more friends they have, the happier they are. So when we randomly assign them to experimental and control groups, we run the risk of unhappy randomization – sometimes one group will have substantially more friends than the other.

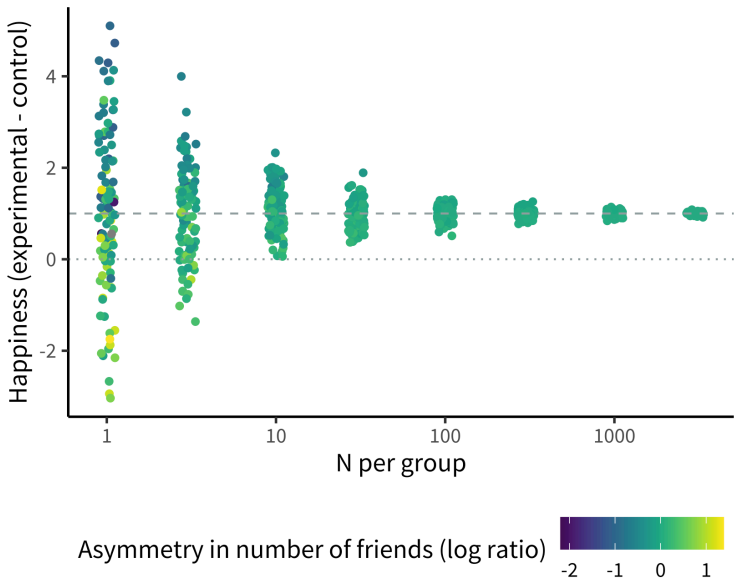


Figure 1.5: Simulated data from our money-happiness experiment. Each dot represents the measured happiness effect (vertical position) for an experiment with a set number of participants in each group (horizontal position). Dot color shows how uneven friendship is between the groups. The dashed line shows the true effect.

Figure 1.5 shows the results of this simulation. Each dot is an experiment, representing one estimate of the happiness effect (how much happiness is gained for the amount of money given to the experimental group). For very small experiments (e.g., with 1 or 3 participants per group), dots are very far from the dashed line showing the true effect – meaning these estimates are extremely noisy! And the reason is unhappy randomization. The upper and lower points are those in which one group had far more friends than the other.

There are three things to notice about this simulation. First, the noise overall goes down as the sample sizes get bigger: larger experiments yield estimates closer to the true effect. Second, the unhappy randomization decreases dramatically as well with larger samples. Although individuals still differ just as much in large experiments, the *group* average number of friends is virtually identical for each condition in the largest groups.

Finally, although the small experiments are individually very noisy, the *average effect* across all of the small experiments is still very close to the true effect. This last point illustrates what we mean when we say that randomized experiments remove confounds. Even though friendship is still an important factor determining happiness in our simulation, the average effect across experiments is correct and each individual estimate is unbiased.

1.3 Generalizability

When we are asking questions about psychology, it's important to think about who we are trying to study. Do we want to know if money increases happiness in *all people*? In people who live in materialistic societies? In people whose basic needs are not being met? We call the group we are trying to study our **population of interest**, and the people who actually participate in our experiment our **sample**. The process of **sampling** is then what we do to recruit people into our experiment.

Sometimes researchers sample from one population, but make a claim about another (usually broader) population. For example, they may run their experiment with a particular sample of U.S. college students, but then generalize to all people (their intended population of interest). The mismatch of sample and population is not always a problem, but quite often causal relationships are different for different populations.

Unfortunately, psychologists pervasively assume that research on U.S. and European samples generalizes to the rest of the world, and it often does not. To highlight this issue, Henrich, Heine, and Norenzayan (2010) coined the acronym WEIRD. This catchy name describes the oddness of making generalizations about all of humanity from experiments on a sample that is quite unusual because it is Western, Educated, Industrialized, Rich, and Democratic. Henrich and colleagues argue that seemingly "fundamental" psychological functions like visual perception, spatial cognition, and social reasoning all differ pervasively across populations – hence, any generalization from an effect estimated with a WEIRD sub-population may be unwarranted.

In the early 2000's, researchers found that gratitude interventions – like writing a brief essay about something nice that somebody did for you – increased happiness in studies conducted in Western countries. Based on these findings, some psychologists believed that gratitude interventions could increase happiness in all people. But it seems they were wrong. A few years later, Layous et al. (2013) ran a gratitude experiment in

two locations: the U.S. and South Korea. Surprisingly, the gratitude intervention decreased happiness in the South Korean sample. The researchers attributed this negative effect to feelings of indebtedness that people in South Korea more prominently experienced when reflecting on gratitude. In this example, we would say that the findings obtained with the U.S. sample may not **generalize** to people in South Korea.

Issues of generalizability extend to all aspects of an experiment, not just its sample. For example, even if our hypothetical cash intervention experiment resulted in gains in happiness, we might not be warranted in generalizing to different ways of providing money. Perhaps there was something special about the amount of money we gave or the way we provided it that led to the effect we observed. Without testing multiple different intervention types, we can't make a broad claim. As we'll see in Chapter 7 and Chapter 9, this issue has consequences for both our statistical analyses and our experimental designs (Yarkoni 2020).

Questions of generalizability are pervasive, but the first step is to simply acknowledge and reason about them. Perhaps all papers should have a Constraints on Generality statement, where researchers discuss whether they expect their findings to generalize across different samples, experimental stimuli, procedures, and historical and temporal features (Simons, Shoda, and Lindsay 2017). This kind of statement would at least remind researchers to be humble: experiments are a powerful tool for understanding how the world works, but there are limits to what any individual experiment can teach us.

1.4 Anatomy of a randomized experiment

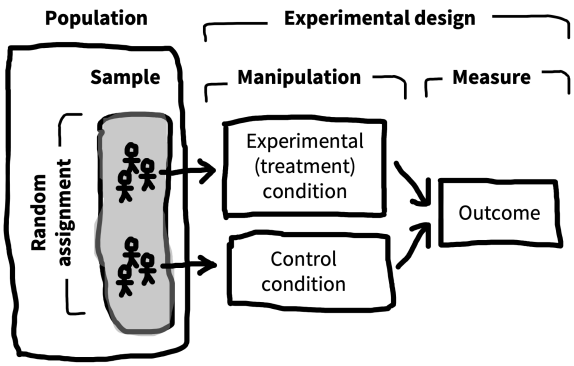


Figure 1.6: Anatomy of a randomized experiment.

Now is a good time for us to go back and consolidate the anatomy of an experiment, since this anatomy is used throughout the book. Figure 1.6 shows a simple two-group experiment like our possible money-happiness intervention. A sample is taken from a larger population, and

then participants in the sample are randomly assigned to one of two conditions (the manipulation) – either the experimental condition, in which money is provided, or the control condition, in which none is given. Then an outcome measure – happiness – is recorded for each participant.

We'll have a lot more to say about all of these components in subsequent chapters. We'll discuss measures in Chapter 8, because good measurement is the foundation of a good experiment. Then in Chapter 9 we'll discuss the different kinds of experimental designs that are possible and their pros and cons. Finally, we'll cover the process of sampling in Chapter 10.

Accident Report

An experiment with very unclear causal inferences

The Stanford Prison Experiment is one of the most famous studies in the history of psychology. Participants were randomly assigned to play the role of "guards" and "prisoners" in a simulation of prison life inside the Stanford Psychology building (Zimbardo 1972). Designed to run for two weeks, the simulation had to be ended after six days due to the cruelty of the participants acting as guards, who apparently engaged in a variety of dehumanizing behaviors towards the simulated prisoners. This result is widely featured in introductory psychology textbooks and is typically interpreted as showing the power of situational factors: in the right context, even undergraduate students at Stanford could quickly be convinced to act out the kind of inhumane behaviors found in the worst prisons in the world (Griggs 2014).

In the years since the study was initially reported, a variety of information has surfaced that makes the causal interpretation of its situational manipulation much less clear (Le Texier 2019). Guards were informed of the objectives of the experiment and given instructions on how to achieve these objectives. The experimenters themselves suggested some harsh punishments whose later use was given as evidence for the emergence of dehumanizing behaviors. Further, both guards and prisoners were coached extensively by the experimenter throughout the study. Some participants have reported that their responses during the study were exaggerated or fabricated (Blum 2018). All of these issues substantially undermine the idea that the assignment of participants' roles (the ostensible experimental manipulation) was the sole cause of the observed behaviors.

The conduct of the study was also unethical. In addition to the question of whether such a study – with all of its risks to the participants – would be ethical at all, a number of features of the study clearly violate the guidelines that we'll learn about in Chapter 4. Participants were prevented from exiting the study voluntarily. The guards were deceived into believing that they were research assistants, rather than participants in the study. And to top it off, the study was reported inaccurately, with reports emphasizing the organic emergence of behaviors, the immersive nature of the simulation, and the extensive documentation of the experiment. In fact, the participants were instructed extensively, the simulation was repeatedly interrupted by mundane details of the research environment, and relatively little of the experiment was captured on video and analyzed.

The Prison Experiment is a fascinating and problematic episode in the history of psychology, but it provides very little causal evidence about the human mind.

1.5 Chapter summary: Experiments

In this chapter, we defined an experiment as a combination of a manipulation and a measure. When combined with randomization, experiments allow us to make strong causal inferences, even when we are studying people (who are hard to hold constant). Nonetheless, there are limits to the power of experiments: there are always constraints on the sample, experimental stimuli, and procedure that limit how broadly we can generalize.

Discussion questions

1. Imagine that you run a survey and find that people who spend more time playing violent video games tend to be more aggressive (i.e., that there is a positive correlation between violent video games and aggression). Following Figure 1.2, list three reasons why these variables may be correlated.
2. Suppose you wanted to run an experiment testing whether playing violent video games causes increases in aggression. What would be your manipulation and what would be your measure? How would you deal with potential confounding by variables like age?
3. Consider an experiment designed to test people's food preferences. The experimenter randomly assigns 30 U.S. preschoolers to be served either asparagus or chicken tenders and then asks them how much they enjoyed their meal. Overall, children enjoyed the meat more; the experimenter writes a paper claiming that humans prefer meat over vegetables. List some constraints on the generalizability of this study. In light of these constraints, is this study (or some modification) worth doing at all?
4. Consider the Milgram study, another classic psychology study (and our case study in Chapter 4). Does this study meet our definition of an experiment?

Readings

– A basic introduction to causal inference from a social science perspective: Huntington-Klein, N. (2022). *The Effect: An Introduction to Research Design and Causality.* Chapman & Hall. Available free online at https://theeffectbook.net.

– A slightly more advanced treatment, focusing primarily on econometrics: Cunningham, S. (2021). *Causal Inference: The Mixtape.* Yale Press. Available free online at https://mixtape.scunning.com.

References

Blum, B. 2018. "The Lifespan of a Lie." Medium. https://medium.com/s/trustissues/the-lifespan-of-a-lie-d869212b1f62.

Griggs, Richard A. 2014. "Coverage of the Stanford Prison Experiment in Introductory Psychology Textbooks." *Teaching of Psychology* 41 (3): 195–203.

Henrich, Joseph, Steven J Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2-3): 61–83.

Layous, Kristin, Hyunjung Lee, Incheol Choi, and Sonja Lyubomirsky. 2013. "Culture Matters When Designing a Successful Happiness-Increasing Activity: A Comparison of the United States and South Korea." *Journal of Cross-Cultural Psychology* 44 (8): 1294–1303.

Le Texier, Thibault. 2019. "Debunking the Stanford Prison Experiment." *American Psychologist* 74 (7): 823.

Lewis, David. 1973. *Counterfactuals*. John Wiley & Sons.

Mill, John Stuart. 1882. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Princilples of Evidence and the Methods of Scientific Investigation*. 8th ed. Harper; Brothers.

Pearl, Judea. 1998. "Graphical Models for Probabilistic and Causal Reasoning." *Quantified Representation of Uncertainty and Imprecision*, 367–89.

Simons, Daniel J, Yuichi Shoda, and D Stephen Lindsay. 2017. "Constraints on Generality (COG): A Proposed Addition to All Empirical Papers." *Perspectives on Psychological Science* 12 (6): 1123–28.

Wysocki, Anna C, Katherine M Lawson, and Mijke Rhemtulla. 2022. "Statistical Control Requires Causal Justification." *Advances in Methods and Practices in Psychological Science* 5 (2): 25152459221095823.

Yarkoni, Tal. 2020. "The Generalizability Crisis." *Behavioral and Brain Sciences* 45: 1–37.

Zimbardo, Philip G. 1972. "Pathology of Imprisonment." *Society* 9 (6): 4–8.

2 Theories

Learning goals

- Define theories and their components
- Contrast different philosophical views on scientific theories
- Analyze features of an experiment that can lead to strong tests of theory
- Discuss the role of formalization in theory development

When you do an experiment, sometimes you just want to see what happens, like a kid knocking down a tower made of blocks. And sometimes you want to know the answer to a specific applied question, like "will giving a midterm vs. weekly quizzes lead students in a class to perform better on the final?" But more often, our goal is to create **theories** that help us explain and predict new observations.

What is a theory? We'll argue here that we should think of psychological theories as sets of proposed relationships among **constructs**, which are variables that we think play causal roles in determining behavior. In this conception of theories, the role of causality is central: theories are guesses about the causal structure of the mind and about the causal relationships between the mind and the world. This definition doesn't include everything that gets called a "theory" in psychology. We describe the continuum between theories and **frameworks** – broad sets of ideas that guide research but don't make specific contact with particular empirical observations.

We begin this chapter by talking about the specific enterprise of constructing psychological theories. We'll then discuss how theories make contact with data, reviewing a bit of the philosophy of science, and give some guidance on how to construct experiments that test theories. We end by discussing the relationship between theories and quantitative models. This material touches on several of our book themes, including GENERALIZABILITY of theories and the need for MEASUREMENT PRECISION to make strong tests of theory.

2.1 What is a psychological theory?

The definition we just gave for a psychological theory is that it is a proposed set of causal relationships among constructs that helps us explain behavior. Let's look at the ingredients of a theory: the constructs and the relationships between them. Then we can ask about how this definition relates to other things that get called "theories" in psychology.

2.1.1 Psychological constructs

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 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, Diener, and Seidlitz 1993).[1]

Even external, apparently non-psychological 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, Foushee, and Srinivasan 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

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.[2]

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.[3] So researchers only get to probe them by measuring them through specific operationalizations.

[1] 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.

[2] 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.

[3] 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 et seq.).

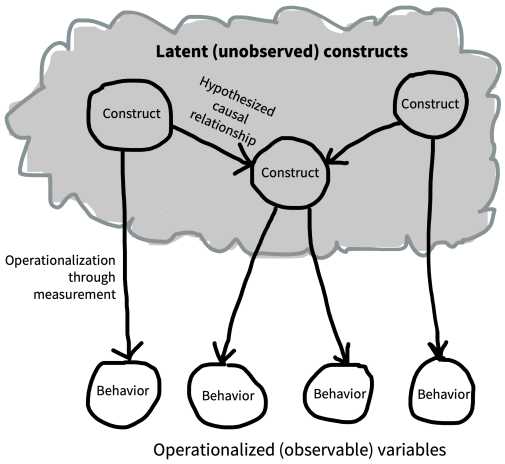


Figure 2.1: A schematic of what a theory might look like.

One poetic way of thinking about this idea is that the theoretical system of constructs "floats… above the plane of observation and is anchored to it by the rules of measurement." (Hempel 1952). 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.[4] 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.

[4] The relationship between money and happiness is actually much more complicated than what we're assuming here. For example, Killingsworth, Kahneman, and Mellers (2023) describes a collaboration between two sets of researchers that 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

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 that 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 (1992)'s 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!

EST 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. EST 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.

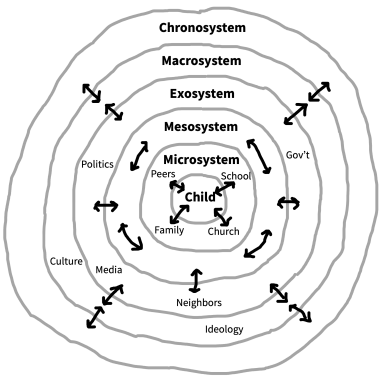


Figure 2.2: The diagram often used to represent Bronfenbrenner's ecological systems theory. Note that circles no longer denote discrete constructs; arrows can be interpreted as causal relationships, but all constructs are assumed to be fully connected.

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 the lens of EST.

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.

Accident report

The cost of a bad theory

Theory development isn't just about knowledge for knowledge's sake – it has implications for the technologies and policies built off the theories.

One case study comes from Edward Clarke's infamous theory regarding the deleterious effects of education for women (Clarke 1884). Clarke posited that (1) cognitive and reproductive processes relied on the same fixed pool of energy, (2) relative to men, women's reproductive processes required more energy, and that (3) expending too much energy on cognitive tasks like education depleted women of the energy needed to maintain a healthy reproductive system. Based on case studies, Clarke suggested that education was causing women to become ill, experience fertility issues, and birth weaker children He thus concluded that "boys must study and work in a boy's way, and girls in a girl's way" (p. 18).

Clarke's work is a chilling example of the implication of a poorly-developed theory. In this scenario, Clarke had neither instruments that allowed him to measure his constructs or experiments to measure the causal connections between them. Instead, he merely highlighted case studies that were consistent with his idea (while simultaneously dismissing cases that were inconsistent). His ideas eventually lost favor – especially as they were subjected to more rigorous tests. But Clarke's arguments were used to attempt to dissuade women from pursuing higher education and hindered educational policy reform.

2.2 How do we test theories?

Our view of psychological theories is that they describe a set of relationships between different constructs. How can we test theories and decide which one is best? We'll first describe **falsificationism**, a historical viewpoint on this issue that has been very influential in the past and that connects to ideas about statistical inference presented in Chapter 6. We'll then turn to a more modern viewpoint, **holism**, that recognizes the interconnections between theory and measurement.

2.2.1 Falsificationism

One historical view that resonates with many scientists is the philosopher Karl Popper's **falsificationism**. In particular, there is a simplistic version of falsificationism that is often repeated by working scientists, even though it's much less nuanced than what Popper actually said! On this view, a scientific theory is a set of hypotheses about the world that instantiate claims like the connection between money and happiness.[5] What makes a statement a *scientific* hypothesis is that it can be disproved (i.e., it is **falsifiable**) by an observation that contradicts it. For example, observing a lottery winner who immediately becomes depressed would falsify the hypothesis that receiving money makes you happier.

For the simplistic falsificationist, theories are never **confirmed**. The hypotheses that form parts of theories are universal statements. You can never prove them right; you can only fail to find falsifying evidence. Seeing hundreds of people get happier when they received money would not prove that the money-happiness hypothesis was universally true. There could always be a counter-example around the corner.

This theory doesn't really describe how scientists work. For example, scientists like to say that their evidence "supports" or "confirms" their theory, and falsificationism rejects this kind of talk. A falsificationist says that confirmation is an illusion; that the theory is simply surviving to be tested another day. This strict falsificationist perspective is unpalatable to many scientists. After all, if we observe that hundreds of people get happier when they receive money, it seems like this should at least slightly increase our confidence that money causes happiness![6]

[5] Earlier we treated the claim that money caused happiness as a theory. It is one! It's just a very simple theory that has only one hypothesized connection in it.

[6] An alternative perspective comes from the Bayesian tradition that we'll learn more about in Chapters 5 and 6. In a nutshell, Bayesians propose that our subjective belief in a particular hypothesis can be captured by a probability, and that our scientific reasoning can then be described by a process of normative probabilistic reasoning (Strevens 2006). The Bayesian scientist distributes probability across a wide range of alternative hypotheses; observations that are more consistent with a hypothesis increase the hypothesis's probability (Sprenger and Hartmann 2019).

2.2.2 A holistic viewpoint on theory testing

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 auxilliary 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 explain), as well as the auxiliary hypotheses – operationalizations – that link it to specific observations.

2.3 *Designing experiments to test theory*

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).[7]

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

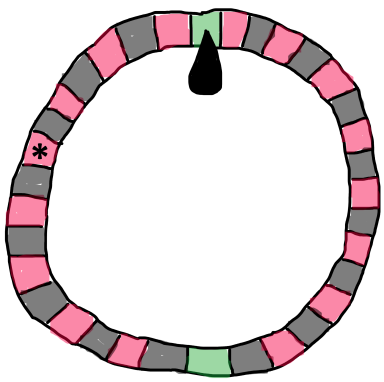


Figure 2.3: A roulette wheel. Betting on red is not that risky, but betting all your chips on a particular value (*) is much riskier.

[7] 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.

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, Kahneman, and Mellers 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![8]

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 very 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.)

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 the process of 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

[8] Theories are often developed iteratively. It's common to start with a theory that is less precise and hence, that has fewer opportunities for risky tests. But 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.

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.[9]

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!

[9] We can use this idea, which comes from Bayesian statistics, to try to figure out what the *right* experiment is by considering which specific experimental conditions derive 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 now called **optimal experiment design** (Myung, Cavagnaro, and Pitt 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.

2.4 Formalizing theories

Say we have a set of constructs we want to theorize about. How do we describe our ideas about the relationships between them so that we can make precise predictions that can be compared with other theories? As one writer noted, mathematics is "unreasonably effective" as a vocabulary for the sciences (Wigner 1990). Indeed, there have been calls for greater formalization of theory in psychology for at least the last 50 years (Harris 1976).

Depth

A universal law of generalization?

How do you take what you know and apply it to a new situation? One answer is that you use the same answer that has worked in similar situations. To do this kind of extrapolation, however, you need a notion of similarity. Early learning theorists tried to measure similarity by creating an association between a stimulus – say a projected circle of light of a particular size – and a reward by repeatedly presenting them together. After this association was learned, they would test generalization by showing circles of different sizes and measuring the strength of the expectation for a reward. These experiments yielded generalization curves: the more similar the stimulus, the more people and other animals would give the same response, signaling generalization.

Shepard (1987) was interested in unifying the results of these different experiments. The first step in this process was establishing a **stimulus space**. He used a procedure called "multidimensional scaling" to infer how close stimuli were to each other on the basis of how strong the generalization between them was. When he plotted the strength of the generalization by the distance between stimuli within this space (their similarity), he found an incredibly consistent pattern: generalization decreased exponentially as similarity decreased.

He argued that this described a "universal law" that governed the relationship between similarity and generalization for almost any stimulus, whether it was the size of circles, the color of patches of light, or the similarity between speech sounds. Later work has even extended this same framework to highly abstract dimensions such as the

relationships between numbers of different types [e.g., being even, being powers of 2, etc.; Tenenbaum (2000)].

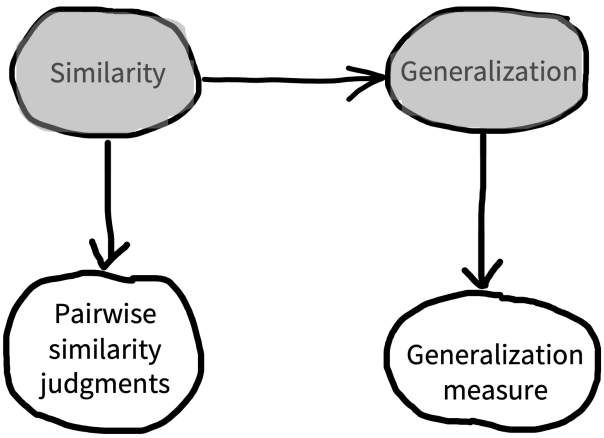


Figure 2.4: The causal theory of similarity and generalization posited by Shepard (1987).

The pattern shown in Shepard's work is an example of **inductive theory building**. In the vocabulary we're developing, Shepard ran (or obtained the data from) randomized experiments in which the manipulation was stimulus dimension (e.g., circle size) and the measure was generalization strength. Then the theory that Shepard proposed was that manipulations of stimulus dimension acted to change the perceived similarity between the stimuli. His theory thus linked two constructs: stimulus similarity and generalization strength (Figure 2.4). Critically the causal relationship he described was not just a qualitative relationship but instead a specific mathematical form.

Shepard wrote in the conclusion of his 1987 paper, "Possibly, behind the diverse behaviors of humans and animals, as behind the various motions of planets and stars, we may discern the operation of universal laws." While Shepard's dream is an ambitious one, it defines an ideal for psychological theorizing.

There is no one approach that will be right for theorizing across all areas of psychology (Oberauer and Lewandowsky 2019; Smaldino 2020). Mathematical theories [such as Shepard (1987); see Depth box] have long been one tool that allows for precise statements of particular relationships.

Computational or formal artifacts are not themselves psychological theories, but they can be used to create psychological theories via the mapping of constructs onto entities in the model and the use of the principles of the formalism to instantiate psychological hypotheses or assumptions (Guest and Martin 2021).[10] Yet stating such clear and general laws feels out of reach in many cases. If we had more Shepard-style theorists or theories, perhaps we'd be in a better place. Or perhaps such "universal laws" are simply out of reach for most of human behavior.

An alternative approach creates statistical models of data that incorporate substantive assumptions about the structure of the data. We use such models all the time for data analysis. The trouble is, we often

don't interpret them as having substantive assumptions about the structure of the data, even when they do (Fried 2020)! But if we examine these assumptions explicitly, even the simplest statistical models can be productive tools for building theories.

For example, if we set up a simple linear regression model to estimate the relationship between money and happiness, we'd be positing a linear relationship between the two variables – that an increase in one would always lead to a proportional increase in the other.[11] If we fit the model to a particular dataset, we could then look at the weights of the model. Our theory might then then be something like "giving people $100 causes 0.2 points of increase in happiness on a self-report scale."

[11] Linear models are ubiquitous in the social sciences because they are convenient to fit, but as theoretical models they are deeply impoverished. There is a lot you can do with a linear regression, but in the end, most interesting processes are not linear combinations of factors!

Obviously, this regression model is not a very good theory of the broader relationship between money and happiness, since it posits that everyone's happiness would be at the maximum on the 10 point scale if you gave them (at most) $4500. It also doesn't tell us how this theory would generalize to other people, other measures of happiness, or other aspects of the psychological representation of money such as income or wealth.

From our viewpoint, these sorts of questions are not distractions – they are the critical work of moving from experiment to theory (Smaldino 2020)! In Chapter 7, we try to draw out this idea further, reconstruing common statistical tests as models that can be repurposed to express contentful scientific hypotheses while recognizing the limitations of their assumptions.

One of the strengths of modern cognitive science is that it provides a very rich set of tools for expressing more complex statistical models and linking them to data. For example, the modern Bayesian cognitive modeling tradition grew out of work like Shepard's; in these models, a system of equations defines a probability distribution that can be used to estimate parameters, predict new data, or make other inferences (Goodman, Tenenbaum, and Contributors 2016). And neural network models – which are now fueling innovations in artificial intelligence – have a long history of being used as substantive models of human psychology (Elman, Bates, and Johnson 1996). One way to think about all these alternatives is as being on a gradient from the general, inspirational frameworks we described above all the way down through computational models and then to statistical models that can be fit to specific datasets (Figure 2.5).

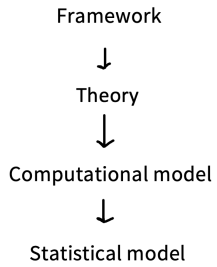


Figure 2.5: A gradient of specificity in theoretical tools. Figure inspired by Guest and Martin (2021).

In our discussion, we've presented theories as static entities that are presented, tested, confirmed, and falsified. That's a simplification that

doesn't take into account the ways that theories – especially when instantiated as formal models – can be flexibly adjusted to accommodate new data (Navarro 2019). Most modern computational theories are more like a combination of core principles, auxiliary assumptions, and supporting empirical assumptions. The best theories are always being enlarged and refined in response to new data.[12]

[12] In the thinking of the philosopher Imre Lakatos, a "productive" research program is one where the core principles are gradually supplemented with a limited set of additional assumptions to explain a growing base of observations. In contrast, a "degenerate" research program is one in which you are constantly making ad-hoc tweaks to the theory to explain each new datapoint (Lakatos 1976).

2.5 Chapter summary: Theories

In this chapter, we characterized psychological theories as a set of causal relationships between latent constructs. The role of experiments is to measure these causal relationships and to adjudicate between theories by identifying cases where different theories make different predictions about particular relationships.

Discussion questions

1. Identify an influential theory in your field or sub-field. Can you draw the "nomological network" for it? What are the key constructs and how are they measured? Are the links between constructs just directional links or is there additional information about what type of relationship exists? Or does our description of a theory in this chapter not fit your example?
2. Can you think of an experiment that falsified a theory in your area of psychology? To what extent is falsification possible for the kinds of theories that you are interested in studying?

Readings

- A fabulous introduction to issues in the philosophy of science can be found in: Godfrey-Smith, P. (2009). Theory and reality. University of Chicago Press.
- Bayesian modeling has been very influential in cognitive science and neuroscience. A good introduction in cognitive science comes from: Lee, M. D. & Wagenmakers, E. J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press. Much of the book is available free online at https://faculty.sites.uci.edu/mdlee/bgm/.
- A recent introduction to Bayesian modeling with a neuroscience focus: Ma, W. J., Kording, K. P., & Goldreich, D. (2022). *Bayesian models of perception and action: An introduction*. MIT Press. Free online at https://www.cns.nyu.edu/malab/bayesianbook.html.

References

Block, Ned J, and Jerry A Fodor. 1972. "What Psychological States Are Not." *The Philosophical Review* 81 (2): 159–81.

Bronfenbrenner, Urie. 1992. *Ecological Systems Theory*. Jessica Kingsley Publishers.

Clarke, Edward H. 1884. *Sex in Education: Or, a Fair Chance for the Girl*. Boston: Houghton, Mifflin; Company.

Cronbach, L J, and P E Meehl. 1955. "Construct Validity in Psychological Tests." *Psychol. Bull.* 52 (4): 281–302.

Ellwood-Lowe, Monica E, Ruthe Foushee, and Mahesh Srinivasan. 2022. "What Causes the Word Gap? Financial Concerns May Systematically Suppress Child-Directed Speech." *Developmental Science* 25 (1): e13151.

Elman, Jeffrey L, Elizabeth A Bates, and Mark H Johnson. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. Vol. 10. MIT press.

Fried, Eiko I. 2020. "Lack of Theory Building and Testing Impedes Progress in the Factor and Network Literature." *Psychological Inquiry* 31 (4): 271–88.

Goodman, Noah D, Joshua B. Tenenbaum, and The ProbMods Contributors. 2016. "Probabilistic Models of Cognition." https://probmods.org/.

Guest, Olivia, and Andrea E Martin. 2021. "How Computational Modeling Can Force Theory Building in Psychological Science." *Perspectives on Psychological Science* 16 (4): 789–802.

Harris, Richard J. 1976. "The Uncertain Connection Between Verbal Theories and Research Hypotheses in Social Psychology." *Journal of Experimental Social Psychology* 12 (2): 210–19.

Hempel, Carl G. 1952. *Fundamentals of Concept Formation in Empirical Science*. University of Chicago Press.

Killingsworth, Matthew A, Daniel Kahneman, and Barbara Mellers. 2023. "Income and Emotional Well-Being: A Conflict Resolved." *Proceedings of the National Academy of Sciences* 120 (10): e2208661120.

Kraus, Michael W, Ivuoma N Onyeador, Natalie M Daumeyer, Julian M Rucker, and Jennifer A Richeson. 2019. "The Misperception of Racial Economic Inequality." *Perspectives on Psychological Science* 14 (6): 899–921.

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Princeton University Press.

Lakatos, Imre. 1976. "Falsification and the Methodology of Scientific Research Programmes." In *Can Theories Be Refuted?*, 205–59. Springer.

Lindley, Dennis V. 1956. "On a Measure of the Information Provided by an Experiment." *The Annals of Mathematical Statistics*, 986–1005.

Meehl, Paul E. 1978. "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *J. Consult. Clin. Psychol.* 46 (4): 806–34.

———. 1990. "Why Summaries of Research on Psychological Theories Are Often Uninterpretable." *Psychological Reports* 66 (1): 195–244.

Myung, Jay I, Daniel R Cavagnaro, and Mark A Pitt. 2013. "A Tutorial on Adaptive Design Optimization." *Journal of Mathematical Psychology* 57 (3-4): 53–67.

Navarro, Danielle J. 2019. "Between the Devil and the Deep Blue Sea: Tensions Between Scientific Judgement and Statistical Model Selection." *Computational Brain & Behavior* 2 (1): 28–34.

Oberauer, Klaus, and Stephan Lewandowsky. 2019. "Addressing the Theory Crisis in Psychology." *Psychonomic Bulletin & Review* 26 (5): 1596–1618.

Roberts, Seth, and Harold Pashler. 2000. "How Persuasive Is a Good Fit? A Comment on Theory Testing." *Psychological Review* 107 (2): 358.

Rohrer, Julia M. 2018. "Thinking Clearly about Correlations and Causation: Graphical Causal Models for Observational Data." *Advances in Methods and Practices in Psychological Science* 1 (1): 27–42.

Sandvik, Ed, Ed Diener, and Larry Seidlitz. 1993. "Subjective Well-Being: The Convergence and Stability of Self-Report and Non-Self-Report Measures." *Journal of Personality* 61 (3): 317–42.

Shepard, Roger N. 1987. "Toward a Universal Law of Generalization for Psychological Science." *Science* 237 (4820): 1317–23.

Smaldino, Paul E. 2020. "How to Translate a Verbal Theory into a Formal Model." *Social Psychology*.

Sprenger, Jan, and Stephan Hartmann. 2019. *Bayesian Philosophy of Science*. Oxford University Press.

Strevens, Michael. 2006. "The Bayesian Approach to the Philosophy of Science." In *Encyclopedia of Philosophy, Second Edition*, edited by D. M. Borchert, 495–502. Macmillan Reference.

Tenenbaum, Joshua B. 2000. "Rules and Similarity in Concept Learning." *Advances in Neural Information Processing Systems* 12: 59–65.

Ungar, Michael. 2002. "A Deeper, More Social Ecological Social Work Practice." *Social Service Review* 76 (3): 480–97.

Wigner, Eugene P. 1990. "The Unreasonable Effectiveness of Mathematics in the Natural Sciences." In *Mathematics and Science*, 291–306. World Scientific.

3 Replication

Learning goals

- Define and distinguish reproducibility and replicability
- Synthesize the meta-scientific literature on replication and the causes of replication failures
- Reason about the relation of replication to theory building

In the previous chapters, we introduced experiments, their connection with causal inference, and their role in building psychological theory. In principle, repeated experimental work combined with theory building should yield strong research programs that explain and predict phenomena with increasing scope.

Yet in recent years there has been an increasing recognition that this idealized view of science might not be a good description of what we actually see when we look at the psychology literature. Many classic findings may be wrong, or at least overstated. Their statistical tests might not be trustworthy. The actual numbers are even wrong in many papers! And even when experimental findings are "real," they may not generalise broadly to different people and different situations.

How do we know about these problems? A burgeoning field called **metascience** is providing the evidence. metascience is research *about research*, for example investigating how often findings in a literature can be successfully built on, or trying to figure out how widespread some negative practice is. metascience allows us to go beyond one-off anecdotes about a particular set of flawed results or rumors about bad practices. Perhaps the most obvious sign that something is wrong is that when independent scientists team up in metascience projects and try to repeat previous studies, they often do not get the same results.

Before we begin reviewing this evidence, let's discuss the different ways in which a scientific finding can be repeated. Figure 3.1 gives us a basic starting point for our definitions. For a particular finding in a paper, if we take the same data, do the same analysis, and get the same result, we call that finding **reproducible** (sometimes, **analytically** or **computationally reproducible**). If we collect *new* data using the same methods, do the same analysis, and get the same result, we call that a **replication** and

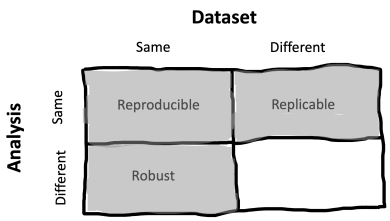


Figure 3.1: A framework for understanding different terms related to the repeatabilty of scientific findings (based on Whitaker 2017).

say that the finding is **replicable**. If we do a different analysis with the same data, we call this a **robustness check** and if we get a similar result, we say that the finding is **robust**.[1] We leave the last quadrant empty because there's no specific term for it in the literature – the eventual goal is to draw **generalizable** conclusions but this term means more than just having a finding that is reproducible and replicable.

In this chapter, we'll primarily discuss reproducibility and replicability (we'll talk about robustness a bit in Chapter 11). We'll start out by reviewing key concepts around reproducibility and replicability as well as some important metascience findings. This literature suggests that when you read an average psychology paper, your default expectation should be that it might not replicate!

We'll then discuss some of the main reasons *why* findings might not replicate – especially **analytic flexibility** and **publication bias**. We end by taking up the issue of how reproducibility and replicability relate to theory building in psychology, and the role of **open science** in this discussion. This discussion focuses on the key role of TRANSPARENCY (one of our major book themes) in supporting theory building.

[1] You might have observed that a lot of work is being done here by the word "same." How do we operationalize same-ness for experimental procedures, statistical analyses, samples, or results? These are difficult questions that we'll touch on below. Keep in mind that there's no single answer and so these terms are always going to helpful guides rather than exact labels.

3.1 Reproducibility

Scientific papers are full of numbers: sample sizes, measurements, statistical results, and visualizations. For those numbers to have meaning, and for other scientists to be able to verify them, we need to know where they came from (their **provenance**). The chain of actions that scientists perform on the raw data, all the way through to reporting numbers in their papers, is sometimes called the *analysis pipeline*. For much of history, scientific papers have only provided a verbal, description of the analysis pipeline, usually with little detail.[2]

Moreover, researchers typically do not share key research objects from this pipeline, such as the analysis scripts or the raw data (Hardwicke, Thibault, et al. 2021).[3] Without code and data, the numbers reported in scientific papers are often not reproducible – an independent scientist cannot repeat all of the steps in the analysis pipeline and get the same results as the original scientists.

Reproducibility is desirable for a number of reasons. Without it:

- Errors in calculation or reporting could lead to disparities between the reported result and the actual result,
- Vague verbal descriptions of analytic computations could keep readers from understanding the computations that were actually performed,

[2] The situation is nicely summed up by a prescient quote from Buckheit and Donoho (1995): "… a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."

[3] For many years, professional societies, like the American Psychological Association, have mandated data sharing (https://www.apa.org/ethics/code), but only for purposes of verification, and only "on request" – in other words, scientists could keep data hidden by default and it was their responsibility to share if another scientist requested access. In practice, this kind of policy does not work; data are rarely made available on request (Wicherts et al. 2006). We believe this situation is untenable. We provide a longer argument justifying data sharing in Chapter 4 and discuss some of the practicalities of sharing in Chapter 13.

- The robustness of data analyses to alternative model specifications cannot be checked, and
- Synthesizing evidence across studies, a key part of building a cumulative body of scientific knowledge, is much more difficult.

From this list, error detection and correction is probably the most pressing issue. But are errors common? There are plenty of individual instances of errors that are corrected in the published literature (e.g., some of us found an error in Cesana-Arlotti et al. 2018), and we ourselves have made significant analytic errors (e.g., Frank et al. 2013). But these kinds of experiences don't tell us about the frequency of errors more generally (or the consequences of error for the conclusions that researchers draw).[4]

Estimating the frequency of errors is a meta-scientific issue that researchers have attempted to answer over the years. If errors are frequent, that would suggest a need for changes in our policies and practices to reduce their frequency! Unfortunately, the lack of data availability creates a problem: it's hard to figure out if calculations are wrong if you can't check them in the first place. Here's one clever approach to this issue. In standard American Psychological Association (APA) reporting format, inferential statistics must be reported with three pieces of information: the test statistic, the degrees of freedom for the test, and the p-value (e.g., $t(18) = -0.74$, $p = 0.47$). Yet these pieces of information are redundant with one another. Thus, reported statistics can be checked for consistency simply by evaluating whether they line up with one another – that is, whether the p-value recomputed from the t and degrees of freedom matches the reported value.

Bakker and Wicherts (2011) performed this kind of statistical consistency analysis on a sample of 281 papers, and found that around 18% of statistical results were incorrectly reported. Even more worrisome, around 15% of articles contained at least one decision error – that is, a case where the error changed the direction of the inference that was made (e.g., from significant to insignificant).[5] Nuijten et al. (2016) used an automated method called "statcheck"[6] to confirm and extend this analysis. They checked p-values for more than 250,000 psychology papers in the period 1985–2013 and found that around half of all papers contained at least one incorrect p-value!

These findings provide a lower bound on the number of errors in the literature and suggest that reproducibility of analyses is likely very important. However, they only address the consistency of statistical reporting. What would happen if we tried to repeat the entire analysis pipeline from start to finish? It's rather difficult to answer this question

[4] There is a very interesting discussion of the pernicious role of scientific error on theory building in Gould (1996)'s "The Mismeasure of Man." Gould examines research on racial differences in intelligence and documents how scientific errors that supported racial differences were often overlooked. Errors are often caught asymmetrically; we are more motivated to double-check a result that contradicts our biases.

[5] Confirming Gould's speculation (see note above), most of the reporting errors that led to decision errors were in line with the researchers' own hypotheses.

[6] Statcheck is now available as a web app (http://statcheck.io) and an R package so that you can check your own manuscripts!

at a large scale: firstly, it takes a long time to run a reproducibility check; and secondly, the lack of access to raw data means that for most scientific papers, checking reproducibility is impossible.

Nevertheless, a few years ago a group of us spotted an opportunity to check reproducibility by examining studies published in two journals that either required or encouraged data sharing. Hardwicke et al. (2018) and Hardwicke, Bohn, et al. (2021) first identified studies that shared data, then narrowed those down to studies that shared *reusable* data (the data were accessible, complete, and comprehensible). For 60 of these articles, we then attempted to reproduce numerical values related to a particular statistical result in the paper. The process was incredibly labor-intensive, with articles typically requiring 5–10 hours of work each. And the results were concerning: the targeted values in only about a third of articles were completely reproducible without help from the original authors! In many cases, after – sometimes extensive – correspondence with the original authors, they provided additional information that was not reported in the original paper. After author contact, the reproducibility success rate improved to 62% (Figure 3.2). The remaining papers appeared to have some values that neither we, nor the original authors, could reproduce. Importantly, we didn't identify any patterns of non-reproducibility that seriously undermined the conclusions drawn in the original articles; however, other reproducibility studies have found a distressingly high number of decision errors (Artner et al. 2020), albeit with a slightly higher success rate overall.

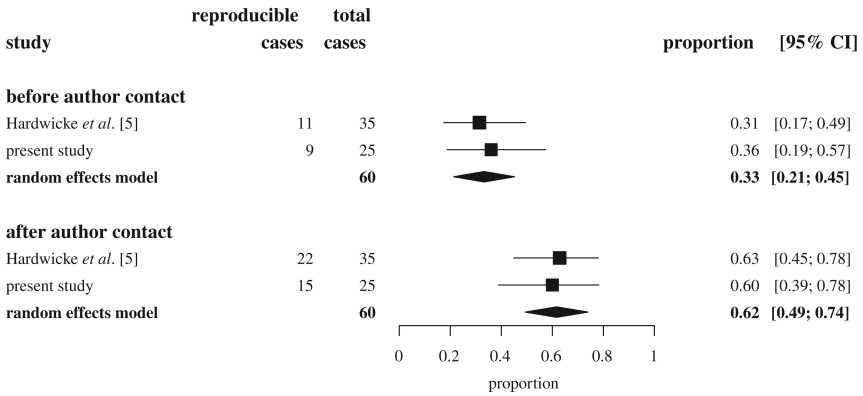


Figure 3.2: Analytic reproducibility of results from open-data articles in *Cognition* and *Psychological Science*. From Hardwicke, Bohn, et al. (2021), Figure 1 (licensed under CC BY 4.0).

In sum: transparency is a critical imperative for decreasing the frequency of errors in the published literature. Reporting and computation errors are frequent in the published literature, and the identification of these errors depends on the findings being reproducible. If data are not available, then errors usually cannot be found.

Case study

The Open Science Collaboration

Around 2011, we were teaching our Experimental Methods course for the first time, based on a course model that we had worked on with Rebecca Saxe (Frank and Saxe 2012). The idea was to introduce students to the nuts and bolts of research by having them run replications. A guy named Brian Nosek was on sabbatical nearby, and over coffee we learned that he was starting up an ambitious project to replicate a large sample of studies published in top psychology journals in 2008.

In the course that year we chose replication projects from the sample that Nosek had told us about. Four of these projects were executed very well and were nominated by the course TAs for inclusion in the broader project. A few years later, when the final group of 100 replication studies was completed, we got a look at the results, shown in Figure 3.3.

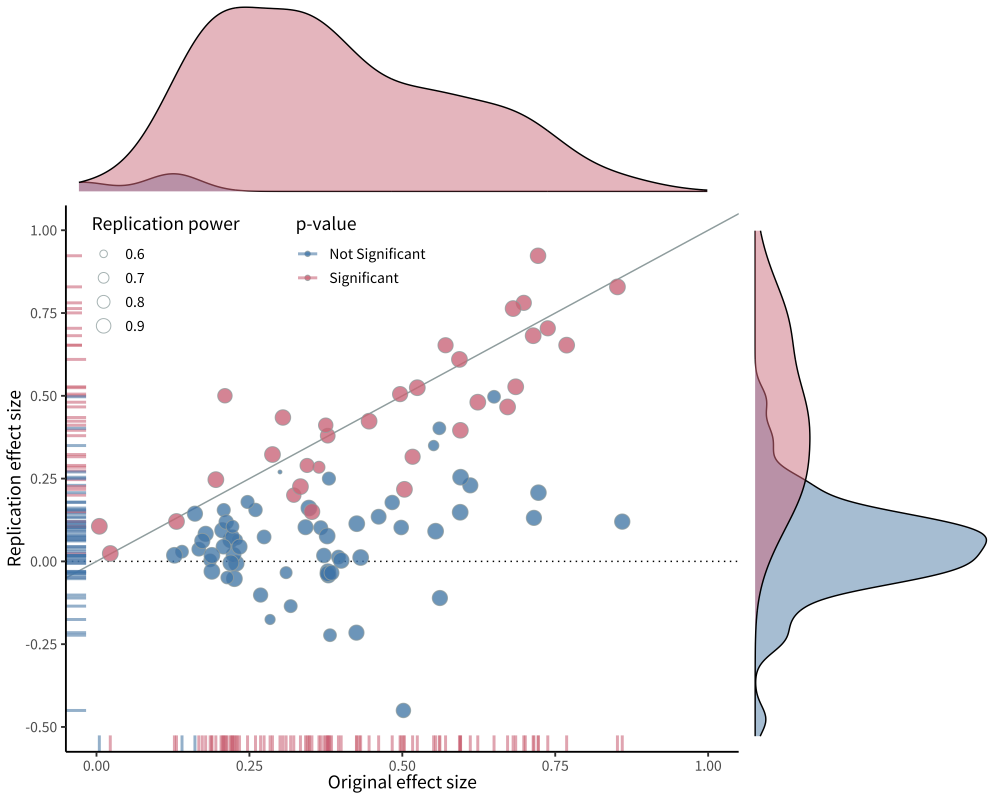


Figure 3.3: Results from the Open Science Collaboration (2015). Each point represents one of the studies in the sample, with the horizontal position giving the original effect size and the vertical position giving the replication effect size. Dot size represents estimated statistical power. The grey line represents a perfect replication.

The resulting metascience paper, which we and others refer to as the "replication project in psychology" (RPP), made a substantial impression on both psychologists and the broader research community, defining both a field of psychology metascience studies and providing a template for many-author collaborative projects (Open Science Collaboration 2015). But the most striking thing was the result: disappointingly, only around a third of the repli-

cations had similar findings to the original studies. The others yielded smaller effects that were not statistically significant in the replication sample (almost all of the original studies were significant). RPP provided the first large-scale evidence that there were systematic issues with replicability in the psychology literature.

RPP's results – and their interpretation – were controversial, however, and much ink was spilled on what these data showed. In particular, critics pointed to different degrees of fidelity between the original studies and the replications; insufficient levels of statistical power and low evidential value in the replications; non-representative sampling of the literature; and difficulties identifying specific statistical outcomes for replication success (Gilbert et al. 2016; Anderson et al. 2016; Etz and Vandekerckhove 2016). In our view, many of these critiques have merit, and you can't simply interpret the results of RPP as an unbiased estimate of the replicability of results in the literature, contra the title.

And yet, RPP's results are still important and compelling, and they undeniably changed the direction of the field of psychology. Many good studies are like this – they have flaws but they inspire follow up studies that can address those problems. For several of us personally, working on this project was also transformative in that it showed us the power of collaborative work. Together we could do a study that no one of us had any hope of completing on our own, and potentially make a difference in our field.

3.2 Replication

Beyond verifying a paper's original analysis pipeline, we are often interested in understanding whether the study can be replicated – if we repeat the study methods and obtain new data, do we get similar results? To quote from Popper (2005), "the scientifically significant… effect may be defined as that which can be regularly [replicated] by anyone who carries out the appropriate experiment in the way prescribed."

Replications can be conducted for many reasons (Schmidt 2009). One goal can be to verify that the results of an existing study can be obtained again if the study is conducted again in exactly the same way, to the best of our abilities. A second goal can be to gain a more precise estimate of the effect of interest by conducting a larger replication study, or combining the results of a replication study with the existing study. A third goal can be to investigate whether an effect will persist when, for example, the experimental manipulation is done in a different, but still theory-consistent, manner. Alternatively, we might want to investigate whether the effect persists in a different population. Such replications are often efforts to "replicate and extend," and are common both when the same research team wants to conduct a sequence of experiments that each build on one another or when a new team wants to build on a result from a paper they have read (Rosenthal 1990).

Much of the metascience literature (and attendant debate and discussion) has focused on the first goal – simple verification. This focus has been so intense that the term "replication" has become associated with skepticism or even attacks on the foundations of the field. This dynamic

is at odds with the role that replication is given in a lot of philosophy of science, where it is assumed to be a typical part of "normal science."

3.2.1 Conceptual frameworks for replication

The key challenge of replication is **invariance** – Popper's stipulation that a replication be conducted "in the way prescribed" in the quote above. That is, what are the features of the world over which a particular observation should be relatively constant, and what are those that are specified as the key ingredients for the effect? Replication is relatively straightforward in the physical and biological sciences, in part because of presupposed theoretical background that allows us to make strong inferences about invariance. If a biologist reports an observation about a particular cell type from an organism, the color of the microscope is presumed not to matter to the observation.

These invariances are far harder to state in psychology, for both the procedure of an experiment and its sample. Procedurally, should the color of the experimental stimulus matter to the measured effect? In some cases yes, in some cases no.[7] Yet the task of postulating how a scientific effect should be invariant to lab procedures pales in comparison to the task of postulating how the effect should be invariant across different human populations![8]

A lot is at stake in this discussion. If Dr. Frog publishes a finding with US undergraduates and Dr. Toad then "replicates" the procedure in Germany, to what extent should we be perturbed if the effect is different in magnitude or absent?[9] Meta-researchers have made a number of replication taxonomies to try and quantify the degree of methodological consistency between two experiments.

Some researchers have tried to distinguish "direct replications"[10] and "conceptual replications". Direct replications are those that attempt to reproduce all of the salient features of the prior study, up to whatever invariances the experimenters believe are present (e.g., color of the paint, gender of the experimenter, etc.). In contrast, conceptual replications are typically paradigms that attempt to test the same hypothesis via different operationalizations of the manipulation and/or the measure. We agree with Zwaan et al. (2018): labeling this second type of experiment as a "replication" is a little misleading. Rather, so-called "conceptual replications" are actually different tests of the same part of your theory. Such tests can be extremely valuable, but they serve a different goal than replication.

[7] A fascinating study by Baribault et al. (2018) proposes a method for empirically understanding psychological invariances. Treating a subliminal priming effect as their model system, they sampled thousands of "micro-experiments" in which small parameters of their experimental procedure were randomly sampled. These parameters allowed for measurement of their effect of interest, averaging across this irrelevant variation. In their case, it turned out that color did not matter.

[8] In some sense, the research program of some branches of the social sciences amounts to an understanding of invariances across human cognition.

[9] Presumably not very much if Dr. Toad gave the original instructions in English instead of in German – that's another one of these pesky invariances that we are always worrying about!

[10] These also get called **exact replications** sometimes. We think this term is misleading because similarity between two different experiments is always going to be on a gradient, and where you cut this continuum is always going to be a theory-laden decision. One person's "exact" is another's "inexact."

Accident report

"Small Telescopes"

We've been discussing the question of invariance with respect to procedure and sample, but we haven't really discussed invariance with respect to the studies' statistical results. To what extent can we consider two statistical results to be "the same"? Several obvious metrics, including those used by RPP, have important limitations (Simonsohn 2015). For example, if one finding is statistically significant and the other isn't, they still could have effect sizes that are actually quite close to one another, in part because one might have a larger sample size than the other. Or you could have two significant findings that nevertheless have very different effect sizes.

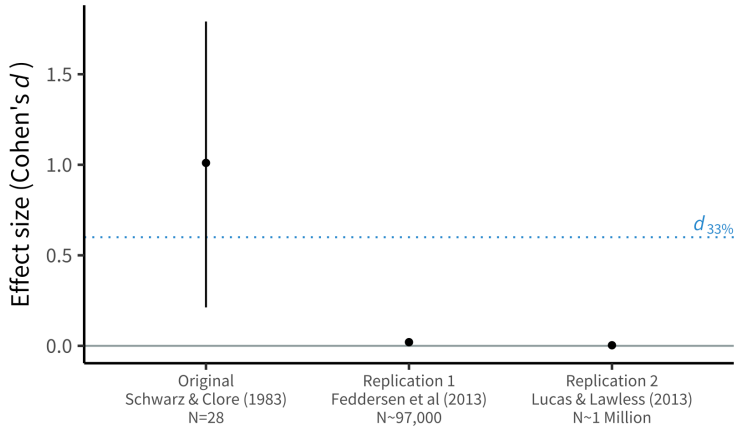


Figure 3.4: The original finding by Schwarz and Clore (1983) and two replications with much larger samples. All three estimates include a 95% confidence interval, but the confidence intervals are very small for the two replication studies. The blue dotted line shows the smallest effect that the origial study could reasonably have detected. Based on Simonsohn (2015).

In a classic study, Schwarz and Clore (1983) reported that participants (N=28) rated their life satisfaction as higher on sunny days than rainy days, suggesting that they mis-attributed temporary happiness about the weather to longer-term life satisfaction. However, when two more recent studies examined very large samples of survey responses, they yielded estimates of the effect that were much smaller. (All of these effects have been standardized so they are on the same scale using a metric called Cohen's d that we will introduce more formally in Chapter 5). In one survey, the effect was statistically significant but extremely small; in the other it was essentially zero (Figure 3.4). Using statistical significance as the metric of replication success, you might be tempted to say that the first of these studies was a successful replication and the second was a failed replication.

Simonsohn points out that this interpretation doesn't make sense, using the analogy of a study's sample size as a telescope. Following this analogy, Schwarz and Clore had a very small telescope (i.e., a small sample size), and they pointed it in a particular direction and claimed to have observed a planet (i.e., a nonzero effect). Now it might turn out that there *was* a planet at that location when you look with a much larger telescope (first replication), and it might turn out that there *wasn't* (second replication). Regardless, however, the original small telescope was simply not powerful enough to have seen whatever was there. Both studies fail to replicate the original observation, regardless of whether their observed effect was in the same direction.

Following Simonsohn's example, numerous metrics for replication success have been proposed (Mathur and VanderWeele 2020). The best of these move away from the idea that there is a binary test of whether an individual

replication was successful and towards a comparison of the two effects and whether they appear consistent with the same theory. Gelman (2018) suggests the "time reversal" heuristic – rather than thinking of a replication as a success or a failure, consider the alternative world in which the replication study had been performed first and the original study followed it.

If we leave behind the idea that the original study has precedence, it makes much more sense to consider the sum total of the evidence across multiple experiments. Using this approach, it seems pretty clear that the weather misattribution effect is, at best, a tiny factor in people's overall judgments of their life satisfaction, even if a small study once found a larger effect.

3.2.1 The metascience of replication

In RPP, replication teams reported subjectively that 39% of replications were successful, with 36% reporting a significant effect in the same direction as the original. How generalizable is this estimate – and how replicable *is* psychological research more broadly? Based on the discussion above, we hope we've made you skeptical that this is a well-posed question, at least without additional qualifiers. Any answer is going to have to provide details about the scope of this claim, the definition of replication being used, and the metric for replication success. On the other hand, *versions* of this question have led to a number of empirical studies that help us better understand the scope of replication issues.

Many subsequent empirical studies of replication have focused on particular subfields or journals, with the goal of informing particular field-specific practices or questions. For example, Camerer et al. (2016) replicated all of the between-subject laboratory articles published in two top economics journals in the period 2011–2014. They found a replication rate of 61% of significant effects in the same direction of the original, higher than the rate in RPP but lower than the naive expectation based on their level of statistical power. Another study attempted to replicate all 21 behavioral experiments published in the journals *Science* and *Nature* from 2010–2015, finding a replication rate of 62% significant effects (Camerer et al. 2018). This study was notable because they followed a two-step procedure – after an initial round of replications, they followed up on the failures by consulting with the original authors and pursuing extremely large sample sizes. The resulting estimate thus is less subject to many of the critiques of the original RPP paper. While these types of studies do not answer all the questions that were raised about RPP, they suggest that replication rates for top experiments are not as high as we'd like them to be, even when care is taken with the sampling and individual study protocols.

Other scientists working in the same field can often predict when an experiment will fail to replicate. Dreber et al. (2015) showed that predic-

tion markets (where participants bet small sums of real money on replication outcomes) made fairly accurate estimates of replication success in the aggregate. This result has itself now been replicated several times (e.g., in the Camerer et al., 2018 study described earlier). Maybe even more surprisingly, there's some evidence that machine learning models trained on the text of papers can predict replication success (Yang, Youyou, and Uzzi 2020; Youyou, Yang, and Uzzi 2023), though more work still needs to be done to validate these models and understand the features they use. More generally, these two lines of research suggest the possibility of isolating consistent factors that lead to replication success or failure. (In the next section we consider what these factors are in more depth.)

Although more work still needs to be done to get generalizable estimates of replicability, taken together, the metascience literature does provide some clarity on what we should expect. Altogether, the chance of a significant finding in a (well-powered) replication study of a generic experiment in social and cognitive psychology is likely somewhere around 56%. Furthermore, the replication effect will likely be on average 53% as large (Nosek et al. 2021).

On the other hand, these large-scale replication studies have substantial limitations as well. With relatively few exceptions, the studies chosen for replication used short, computerized tasks that mostly would fall into the categories of social and cognitive psychology. Further, and perhaps most troubling from the perspective of theory development, they tell us only whether a particular experimental effect can be replicated. They tell us much less about whether the construct that the effect was meant to operationalize is in fact real! We'll return to the difficult issue of how replication and theory construction relate to one another in the final section of this chapter.

Some have called the narrative that emerges from the sum of these metascience studies the "replication crisis." We think of it as a major tempering of expectations with respect to the published literature. Your naive expectation might reasonably be that you could read a typical journal article, select an experiment from it, and replicate that experiment in your own research. The upshot of this literature is, unfortunately, if you try selecting and replicating an exeriment, you might well be disappointed by the result.

Accident report

Consequences for the study, consequences for the person

"Power posing" is the idea that adopting a more open and expansive physical posture might also change your confidence. Carney, Cuddy, and Yap (2010) told 42 participants that they were taking part in a study of physiological recording. They then held two poses, each for a minute. In one condition, the poses were expansive (e.g., legs out, hands on head); in another condition, the poses were contractive (e.g., arms and legs crossed). Participants in the expansive pose condition showed increases in testosterone and decreases in salivary cortisol (a stress marker), they took a greater number of risk in a gambling task, and they reported that they were more "in charge" in a survey. This result suggested that a two-minute manipulation could lead to striking physiological and psychological changes – in turn leading to power posing becoming firmly enshrined as part of the set of recommended strategies in business and elsewhere. The original publication contributed to the rise of the researchers' careers, including becoming a principal piece of evidence in a hugely-popular TED talk by Amy Cuddy, one of the authors.

Followup work has questioned these findings, however. A replication study with a larger number of participants (N=200) failed to find evidence for physiological effects of power-posing, even as it did find some effects on participants' own beliefs (Ranehill et al. 2015). And a review of the published literature suggested that many findings appeared to be the result of some sort of publication bias, as far too many of them had *p*-values very close to the .05 threshold (Simmons and Simonsohn 2017). In light of this evidence, the first author of the replication study bravely made a public statement that she does not believe that "power pose" effects are real (Carney 2016).

From the scientific perspective, it's very tempting to take this example as a case in which the scientific ecosystem corrects itself. Although many people continue to cite the original power posing work, we suspect the issues are well-known throughout the social psychology community, and overall interest from the lay public has gone down. But this narrative masks the very real human impacts of the self-correction process, which can raise ethical questions about the best way to address issues in the scientific record.

The process of debate and discussion around individual findings can be bruising and complicated. In the case of power posing, Cuddy herself was tightly associated with the findings and many critiques of the findings became critiques of the individual. Several commentators used Cuddy's name as a stand-in for low-quality psychological results, likely because of her prominence and perhaps because of her gender and age as well. These comments were harmful to Cuddy personally and her career more generally.

Scientists should critique, reproduce, and replicate results – these are all parts of the progress of normal science. But it's important to do this in a way that's sensitive to the people involved. Here are a few guidelines for courteous and ethical conduct:

- Always communicate about the work, never the person. Try to use language that is specific to the analysis or design being critiqued, rather than the person who did the analysis or thought up the design.
- Avoid using language that assumes negative intentions, e.g. "the authors misleadingly state that …"
- Ask someone to read your paper, email, blogpost, or tweet before you hit send. It can be very difficult to predict how someone else will experience the tone of your writing; a reader can help you make this judgement.
- Consider communicating personally before communicating publicly. As Joe Simmons, one critic in the power-posing debate said, "I wish I'd had the presence of mind to pick up the phone and call [before publishing my critique]" (Dominus 2017). Personal communication isn't always necessary (and can be difficult due to asymmetries of power or status), but it can be helpful.

As we will argue in the next chapter, we have an ethical duty as scientists to promote good science and critique low quality science. But we also have a duty to our colleagues and communities to be good to one another.

3.3 Causes of replication failure

Depth

Context, moderators, and expertise

There are many explanations for failed replications. The wonderful thing about metascience is that these explanations can be tested empirically!

Let's start with the idea that specific experimental operationalizations of a theory might be "context sensitive," especially in subfields, like social psychology, whose theories inherently refer to environmental context (Van Bavel et al. 2016). Critics brought this issue up for RPP, where there were several studies in which the original experimental materials were tailored to one cultural context but then were deployed in another context, potentially leading to failure due to mismatch (Gilbert et al. 2016).

Context sensitivity seems like a great explanation because in some sense, it *must* be right. If the context of an experiment includes the vast network of learned associations, practices, and beliefs that we all hold, then there's no question that an experiment's materials tap into this context to one degree or another. For example, if your experiment relies on the association between *doctor* and *nurse* concepts, you would expect this experiment to work differently in the past when *nurse* meant something more like *nanny* (Ramscar 2016).

On the other hand, as an explanation of specific replication failures, context sensitivity has not fared very well. The "Many Labs" projects were a series of replication projects in which *multiple* labs independently attempted to replicate several original studies. (In contrast, in RPP and similar studies, a single replication was conducted for each original study.) Some of the Many Labs projects assessed variation in replication success across different labs. In ManyLabs 2, Klein et al. (2018) replicated 28 findings, distributed across 125 different samples and more than 15,000 participants. ManyLabs 2 found almost no support for the context sensitivity hypothesis as an explanation of replication failure. In general, when effects failed to replicate, they did so when conducted in person as well as when conducted online, and these failures were consistent across many cultures and labs.

On the other hand, a review of several Many Labs-style replication projects indicated, on re-analysis, that population effects differed across replication labs even when the replication protocols were very similar to one another (Olsson-Collentine, Wicherts, and Assen 2020; Errington et al. 2021). So context sensitivity is almost certainly present – and we'll return to the broader issues of generalizability, context, and invariance in the next section – but so far we have not identified specific forms of context sensitivity that reliably affect replication success.

These observations – that 1) direct replications vary in how successful they are, but 2) we cannot identify specific contextual moderators – together suggest the possible presence of "hidden moderators." That is, when faced with a successful original study and a failed replication, there may be some unknown factor(s) that moderates the effect.

We've personally had several experiences that corroborate the idea that there are hidden moderators. For example, in Lewis and Frank (2016), we were unsuccessful in replicating a simple categorization experiment. We then made a series of iterative changes to the stimuli and instructions, for example changing the color and pattern of the stimuli (Figure 3.5), eventually resulting in a larger (and statistically significant) effect – though still much smaller than the original. Critically, however, each alteration that we made to the procedure yielded a very small change in the effect, and it would have taken us many thousands of participants to figure exactly which alteration made the difference. (If you're keeping score, here's a case where stimulus color *did* matter to the outcome of the experiment!).

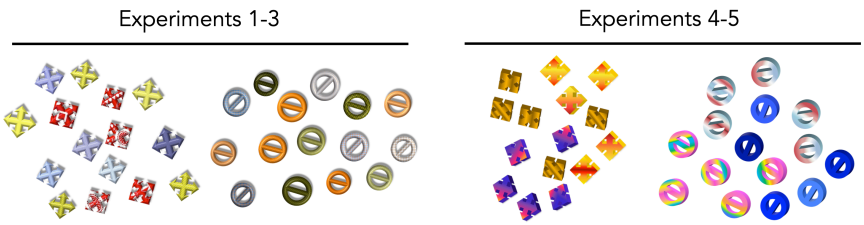


Figure 3.5: Stimuli from Lewis and Frank (2016) (https://github.com/mllewis/xtSamp).

Another explanation for replication failure that is often cited is experimenter expertise (e.g., Schwarz and Strack 2014). On this hypothesis, replications fail because the researchers performing the replication do not have sufficient expertise to execute the study. Like context sensitivity, this explanation is almost certainly true for some replications. In our own work, we have repeatedly performed experiments that failed due to our own incompetence!

Yet as an explanation of the pattern of metascience findings, the expertise hypothesis hasn't been supported empirically. First, team expertise was not a predictor of replication success in RPP (cf. Bench et al. 2017). More convincingly, Many Labs 5 selected ten findings from RPP with unsuccessful replications and systematically evaluated whether formal expert peer review of the protocols, including by the authors of the original study, would lead to a larger effect sizes. Despite a massive sample size and extremely thorough review process, there was little to no change in the effects for the vetted protocols relative to the original protocol used in RPP (Ebersole et al. 2020).

Context, moderators, and expertise seem like reasonable explanations for individual replication failures. Certainly, we should expect them to be explanatory! But for these hypotheses to be operationalized in such a way that they carry weight in our evaluation of the meta-scientific evidence, they must be evaluated empirically rather than accepted uncritically. When such evaluations have been carried out, they have failed to support a large role for these factors.

The general argument of this chapter is that everything is not all right in experimental psychology, and hence that we need to change our methodological practices to avoid negative outcomes like irreproducible papers and unreplicable results. Towards that goal, we have been presenting meta-scientific evidence on reproducibility and replicability. But this evidence has been controversial, to say the least! Do large-scale replication studies like RPP – or for that matter, smaller-scale individual replications of effects like "power posing" – really lead to the conclusion that our methods require changes? Or are there reasons why a lower replication rate is actually consistent with a cumulative, positive vision of psychology?

One line of argument addresses this question through the dynamics of scientific change. There are many versions, but one is given by Wilson, Harris, and Wixted (2020). The idea is that progress in psychology consists of a two-step process by which candidate ideas are "screened" by virtue of small, noisy experiments that reveal promising but tentative ideas that can then be "confirmed" by large-scale replications. On this kind of view, it's business as usual to find that many randomly-selected findings don't hold up in large-scale replications and so we shouldn't be

distressed by results like those of RPP. The key to progress is to finding a small set that *do* hold up, which will lead to new areas of inquiry. We're not sure this is view is either a good description of current practice or a good normative goal for scientific progress, but we won't focus on that critique of Wilson et al.'s argument here. Instead, since book is written for experimenters-in-training, we assume that *you* do not want your experiment to be a false positive from a noisy screening procedure, regardless of your feelings about the rest of the literature!

In RPP and subsequent metascience studies, original studies with lower p-values, larger effect sizes, and larger sample sizes were more likely to replicate successfully (Yang, Youyou, and Uzzi 2020). From a theoretical perspective, this result is to be expected, because the p-value literally captures the probability of the data (or any "more extreme") under the null hypothesis of no effect. So a lower p-value should indicate a lower probability of a spurious result.[11] In some sense, the fundamental question about the replication metascience literature is why the p-values *aren't* better predictors of replicability! For example, Camerer et al. (2018) computes an expected number of successful replications on the basis of the effects and sample sizes – and their proportion of successful replications is substantially lower than that number.[12]

One explanation is that the statistical evidence presented in papers often dramatically overstates the true evidence from a study. That's because of two pervasive and critical issues: **analytic flexibility** (also known as **p-hacking** or **questionable research practices**) and **publication bias.**[13]

Publication bias refers to the relative preference (of scientists and other stakeholders, like journals) for experiments that "work" than those that do not, where "work" is typically defined as yielding a significant result at $p < .05$. Because of this preference, it is typically easier to publish positive (statistically significant) results. The relative absence of negative results leads to biases in the literature. Intuitively, this bias will lead to a literature filled with papers where $p < .05$. Negative findings will then remain unpublished, living in the proverbial "file drawer" (Rosenthal 1979).[14] In a literature with a high degree of publication bias, many findings will be spurious because experimenters got lucky and published the study that "worked" even if that success was due to chance variation. In this situation, these spurious findings will not be replicable and so the overall rate of replicability in the literature will be lowered.[15]

It's our view that publication bias and its even more pervasive cousin, analytic flexibility, are likely to be key drivers of lower replicability. We admit that the meta-scientific evidence for this hypothesis isn't unambiguous, but that's because there's no sure-fire way to diagnose analytic flexibility in a particular paper – since we can almost never reconstruct

[11] In Chapter 6 we will have a lot more to say about $p < .05$ but for now we'll mostly just treat it as a particular research outcome.

[12] This calculation, as with most other metrics of replication success, assumes that the underlying population effect is exactly the same for the replication and the original. This is a limitation because there could be unmeasured moderators that could produce genuine substantive differences between the two estimates.

[13] These terms basically mean the same thing and are not used very precisely in the literature. p-hacking is an informal term that sounds like you know you are doing something bad; sometimes people do, and sometimes they don't. Questionable research practices is a more formal-sounding term that is in principle vague enough to encompass many ethical failings but in practice gets used to talk about p-hacking. Unless p-hacking intent is crystal clear, we favor two clunkier terms: "data-dependent decision-making" and "undisclosed analytic flexibility" describe the actual practices more precisely: trying many different things after looking at data, typically without reporting all of them.

[14] One estimate is that 96% of (non-preregistered) papers report positive findings (Scheel, Schijen, and Lakens 2021)! We'll have a lot more to say about analytic flexibility and publication bias in Chapters 11 and 16, respectively.

[15] The mathematics of the publication bias scenario strikes some observers as implausible: most psychologists don't run dozens of studies and report only one out of each group (Nelson, Simmons, and Simonsohn 2018). Instead, a more common scenario is to conduct many different analyses and then report the most successful, creating some of the same effects as publication bias – a promotion of spurious variation – without a file drawer full of failed studies.

the precise choices that were made in the data collection and analysis process! On the other hand, it is possible to analyze indicators of publication bias in specific literatures and there are several cases where publication bias diagnostics appear to go hand in hand with replication failure. For example, in the "power posing" example described above, Simmons and Simonsohn (2017) noted strong evidence of analytic flexibility throughout the literature, leading them to conclude that there was no evidential value in the literature. And in the case of "money priming" (incidental exposures to images or text about money that were hypothesized to lead to changes in political attitudes), strong evidence of publication bias (Vadillo, Hardwicke, and Shanks 2016) was accompanied by a string of failed replications (Rohrer, Pashler, and Harris 2015).

Accident report

Analytic flexibility reveals a fountain of eternal youth

The way they tell it, Joseph Simmons, Leif Nelson, and Uri Simonsohn wrote their paper on "false positive psychology" (Simmons, Nelson, and Simonsohn 2011) as an attempt at catharsis (Simmons, Nelson, and Simonsohn 2018). They were fed up with work that they felt exploited flexibility in data analysis to produce findings blessed with $p < .05$ but likely did not reflect replicable effects. They called this practice **p-hacking**: trying different things to get your p-value to be below .05.

Their paper reported on a simple experiment: they played participants either the Beatles song, "when I'm 64," or a control song and then asked them to report their date of birth (Simmons, Nelson, and Simonsohn 2011). This manipulation resulted in a significant one and a half year rejuvenation effect. Listening to the Beatles seemed to have made their participants younger!

This result is impossible, of course. But the authors produced a statistically significant difference between the groups that, by definition, was a **false positive** – a case where the statistical test indicated that there was a difference between groups despite no difference existing. In essence, they did so by trying many possible analyses and "cherry-picking" the one that produced a positive result. This practice of course invalidates the inference that the statistical test is supposed to help you make. Several of the practices they followed included:

- Selectively reporting dependent measures (e.g., collecting several measures and reporting only one)
- Selectively dropping manipulation conditions
- Conducting their statistical test and then testing extra participants if they did not see a significant finding
- Adjusting for gender as a covariate in their analysis if doing so resulted in a significant effect

Many of the practices that the authors followed in their rejuvenation study were (and maybe still are!) commonplace in the research literature. John, Loewenstein, and Prelec (2012) surveyed research psychologists on the prevalence of what they called **questionable research practices**. Most participants admitted to following some of these practices – including exactly the same practices followed by the rejuvenation study.

For many in the field, "false positive psychology" was a galvanizing moment, leading them to recognize how common practices could lead to completely spurious (or even impossible) conclusions. As Simmons, Nelson, and Simonsohn wrote in their 2018 article, "Everyone knew [p-hacking] was wrong, but they thought it was wrong the way it is wrong to jaywalk. We decided to write 'False-Positive Psychology' when simulations revealed that it was wrong the way it is wrong to rob a bank."

3.4 Replication, theory building, and open science

Empirical measures of reproducibility and replicability in the experimental psychology literature are low – lower than we may have naively suspected and lower than we want. How do we address these issues? And how do these issues interact with the goal of building theories? In this section, we discuss the relationship between replication and theory – and the role that open and transparent research practices can play.

3.4.1 Reciprocity between replication and theory

Analytic reproducibility is a prerequisite for theory building because if the twin goals of theories are to explain and to predict experimental measurements, then an error-ridden literature undermines this goal. If some proportion of all numerical values reported in the literature were simple, unintentional typos, this situation would create an extra level of noise – irrelevant random variation – impeding our goal of getting precise enough measurements to distinguish between theories. But the situation is likely worse: errors are much more often in the direction that favors authors' own hypotheses. Thus, irreproducibility not only decreases our precision, it also increases the bias in the literature, creating obstacles to the fair evaluation of theories with respect to data.

Replicability is also foundational to theory building. Across a range of different conceptions of how science works, scientific theories are evaluated with respect to their relationship to the world. They must be supported, or at least fail to be falsified, by specific observations. It may be that some observations are by their nature un-repeatable (e.g., a particular astrophysical event might be observed once a human lifetime). But for laboratory sciences – and experimental psychology can be counted among these, to a certain extent at least – the independent and skeptical evaluation of theories requires repeatability of measurements.

Some authors have argued (following the philosopher Heraclitus), "you can't step in the same river twice" (McShane and Böckenholt 2014) – meaning, the circumstances and context of psychological experiments are constantly changing and no observation will be identical to another. This is of course technically true from a philosophical perspective. But that's where theory comes in! As we discussed above, our theories postulate the invariances that allow us to group together similar observations and generalize across them.

In this sense, replication is critical to theory, but theory is also critical to replication. Without a theory of "what matters" to a particular outcome, we really are stepping into an ever-changing river. But a good theory can concentrate our expectations on a much smaller set of causal

relationships, allowing us to make strong predictions about what factors should and shouldn't matter to experimental outcomes. To return to an example we discussed earlier, should stimulus color matter to the outcome of an experiment? Our theory could tell us that it shouldn't matter for a priming experiment (Baribault et al. 2018) but that it should for a generalization experiment (Lewis and Frank 2016).

3.4.2 Deciding when to replicate to maximize epistemic value

As a scientific community, how much emphasis should we place on replication? In the words of Newell (1973), "you can't play 20 questions with nature and win". A series of well-replicated measurements does not itself constitute a theory. Theory construction is its own important activity. We've tried to make the case here that a reproducible and replicable literature is a critical foundation for theory building. That doesn't necessarily mean you have to do replications all the time.

More generally, any scientific community needs to trade off between exploring new phenomena and confirming previously reported effects. In a thought-provoking analysis, Oberauer and Lewandowsky (2019) suggest that perhaps replications also aren't the best test of theoretical hypotheses. In their analysis, if you don't have a theory then it makes sense to try and discover new phenomena and then to replicate them. If you *do* have a theory, you should expend your energy in testing new predictions rather than repeating the same test across multiple replications. Analyses such as Oberauer and Lewandowsky (2019) can provide a guide to our allocation of scientific effort.

Our goal in this book is somewhat different than the general goal of metascientists considering how science should be conducted. Once *you* as a researcher decide to do a particular experiment, we think you will want to maximize its scientific value and so you will want it to be replicable. But we aren't suggesting that you should necessarily do a replication study. There are many concerns that go into whether to replicate – including not only whether you are trying to gather evidence about a particular phenomenon, but also whether you are trying to master techniques and paradigms related to it. As we said at the beginning of this chapter, not all replication is for the purpose of verification, and you as a researcher can make an informed decision about what experimental strategy is best for you.

3.4.3 Open science

The **open science movement** is, in part, a response – really a set of responses – to the challenges of reproducibility and replicability. The

open science (and now the broader **open scholarship**) movement is a broad umbrella (Figure 3.6), but in this book we take open science to be a set of beliefs, research practices, results, and policies that are organized around the central roles of transparency and verifiability in scientific practice.[16] The core of this movement is the idea of "nullius in verba" (the motto of the British Royal Society, which roughly means "take no one's word for it."[17]

Transparency initiatives are critical for ensuring reproducibility. As we discussed above, you cannot even evaluate reproducibility in the absence of data sharing. Code sharing can go even further towards helping reproducibility, as code makes the exact computations involved in data analysis much more explicit than the verbal descriptions that are the norm in papers (Hardwicke et al. 2018). Further, as we will discuss in Chapter 13, the set of practices involved in preparing materials for sharing can themselves encourage reproducibility by leading to better organizational practices for research data, materials, and code.

Transparency also plays a major role in advancing replicability. This point may not seem obvious at first – why would sharing things openly lead to more replicable experiments? – but it is one of the major theses of this book, so we'll unpack it a bit. Here are a couple of routes by which transparent practices lead to greater replication rates.

1. Sharing of experimental materials enables replications that closely follow the original study's methods. One critique of many replications has been that they differ in key respects from the originals. Sometimes those deviations were purposeful, but in other cases they were simply because the replicators could not use the original experimental materials. Sharing materials solves this problem.

2. Sharing sampling and analysis plans allows replication of key aspects of design and analysis that may not be clear in verbal descriptions, for example exclusion criteria or details of data preprocessing.

3. Sharing of analytic decision-making via preregistration can lead to a decrease in p-hacking and other practices that can introduce bias. The strength of statistical evidence in the original study is a predictor of replicability in subsequent studies. If original studies are preregistered, they are more likely to report effects that are not subject to inflation via questionable research practices.

4. Preregistration can also clarify the distinction between confirmatory and exploratory findings, helping subsequent experimenters to make a more informed judgment about which effects are likely to be good targets for replication.

[16] Another part of the open science umbrella involves a democratization of the scientific process through efforts to open access to science. This process involves both removal of barriers to access the scientific literature but also efforts to remove barriers to scientific training – especially to groups historically underrepresented in the sciences. The hope is that these processes increase both the set of people and the range of perspectives contributing to science. We view these changes as no less critical than the transparency aspects of the open science movement, though more indirectly related to the current discussion of reproducibility and replicability.

[17] At least that's a reasonable paraphrase; there's some interesting discussion about what this quote from Horace really means in a letter by Gould (1991).

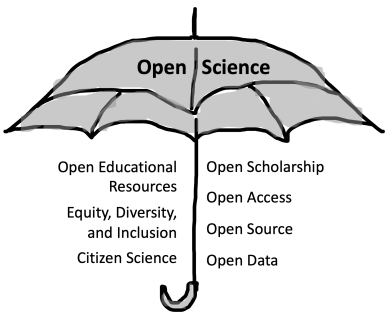


Figure 3.6: The broad umbrella of open science (adapted from an image created for the Stanford Lane Library Blog).

For all of these reasons, we believe that open science practices can play a critical role in increasing reproducibility and replicability.

3.4.4 A crisis?

So, is there a "replication crisis"? The common meaning of "crisis" is "a difficult time." The data we reviewed in this chapter suggest that there are real problems in the reproducibility and replicability of the psychology literature. But there's no evidence that things have gotten worse. If anything, we are optimistic about the changes in practices that have happened in the last ten years. So in that sense, we are not sure that a crisis narrative is warranted.

On the other hand, for Kuhn (1962), the term "crisis" had a special meaning: it is a period of intense uncertainty in a scientific field brought on by the failure of a particular paradigm (Chapter 2). A crisis typically heralds a shift in paradigm, in which new approaches and phenomena come to the fore.

In this sense, the replication crisis narrative isn't mutually exclusive with other crisis narratives, including the "generalizability crisis" (Yarkoni 2020) and the "theory crisis" (Oberauer and Lewandowsky 2019). All of these are symptoms of discontent with the status quo. We share this discontent! We are writing this book to encourage further changes in experimental methods and practices to improve reproducibility and replicability outcomes – many of them driven by the broader set of ideas referred to as "open science." These changes may not lead to a paradigm shift in the Kuhnian sense, but we hope that they lead to eventual improvements. In that sense, we think agree with those who say that the "replication crisis" has led to a "credibility revolution" (Vazire 2018).

3.5 Chapter summary: Replication

In this chapter we introduce the notions of reproducibility – getting the same numbers from the same analysis – and replicability – getting the same conclusions from a new dataset. Both of these are critical prerequisites of a cumulative scientific literature, yet the metascience literature has suggested that the rate of both reproducibility and replicability in the published literature is quite a bit lower than we would hope. A strong candidate explanation for low reproducibility is simply that code and data are rarely shared alongside published research. Lowered replicability is more difficult to explain, but our best guess is that analytic flexibility ("p-hacking") is at least partially to blame. On our account, replication is a meta-scientific tool for understanding the status of the

scientific literature rather than an end in itself. Instead, we see the open science movement, a movement focused on the role of transparency in the scientific process, as a promising response to issues of reproducibility and replicability.

Discussion questions

1. How would you design a measure of the context sensitivity of an experiment? Think of a measure you could apply *post hoc* to a description of an experiment (e.g., from reading a paper) so that you could take a group of experiments and annotate how context-sensitive they are on some scale.
2. Take the measure you designed above. How would you test that this measure really captured context sensitivity in a way that was not circular? What would be an "objective measure" of context sensitivity?
3. What proportion of reproducibility failures do you think are due to questionable practices by experimenters vs. just plain errors? How would you test your hypothesis?

Readings

- Still a very readable and entertaining introduction to the idea of p-hacking: Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science, 22(11), 1359–1366. https://doi.org/10.1177/0956797611417632.
- A recent review of issues of replication in psychology: Nosek, B. et al. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. Annual Review of Psychology, 73, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157.

References

Anderson, CJ, S Bahnik, M Barnett-Cowan, FA Bosco, J Chandler, CR Chartier, and otherss. 2016. "Response to Comment on 'Estimating the Reproducibility of Psychological Science'." *Science* 351 (6277): 1037–37.

Artner, Richard, Thomas Verliefde, Sara Steegen, Sara Gomes, Frits Traets, Francis Tuerlinckx, and Wolf Vanpaemel. 2020. "The Reproducibility of Statistical Results in Psychological Research: An Investigation Using Unpublished Raw Data." *Psychological Methods*.

Bakker, Marjan, and Jelte M Wicherts. 2011. "The (Mis) Reporting of Statistical Results in Psychology Journals." *Behavior Research Methods* 43 (3): 666–78.

Baribault, Beth, Chris Donkin, Daniel R Little, Jennifer S Trueblood, Zita Oravecz, Don Van Ravenzwaaij, Corey N White, Paul De Boeck, and Joachim Vandekerckhove. 2018. "Metastudies for Robust Tests of Theory." *Proceedings of the National Academy of Sciences* 115 (11): 2607–12.

Bench, Shane W, Grace N Rivera, Rebecca J Schlegel, Joshua A Hicks, and Heather C Lench. 2017. "Does Expertise Matter in Replication? An Examination of the Reproducibility Project: Psychology." *Journal of Experimental Social Psychology* 68: 181–84.

Buckheit, Jonathan B, and David L Donoho. 1995. "Wavelab and Reproducible Research." In *Wavelets and Statistics*, 55–81. Springer.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36.

Camerer, Colin F, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015." *Nature Human Behaviour* 2 (9): 637–44.

Carney, Dana R. 2016. "My Position on Power Poses." *Unpublished Manuscript. Haas School of Business, University of California*. https://faculty.haas.berkeley.edu/dana_carney/pdf_my%20position%20on%20power%20poses.pdf.

Carney, Dana R, Amy JC Cuddy, and Andy J Yap. 2010. "Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance." *Psychological Science* 21 (10): 1363–68.

Cesana-Arlotti, N, A Martı□n, E Téglás, L Vorobyova, R Cetnarski, and L L Bonatti. 2018. "Erratum for the Report 'Precursors of Logical Reasoning in Preverbal Human Infants'." *Science* 361 (6408).

Dominus, Susan. 2017. "When the Revolution Came for Amy Cuddy." *When the Revolution Came for Amy Cuddy*. https://www.nytimes.com/2017/10/18/magazine/when-the-revolution-came-for-amy-cuddy.html.

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A Nosek, and Magnus Johannesson. 2015. "Using Prediction Markets to Estimate the Reproducibility of Scientific Research." *Proceedings of the National Academy of Sciences* 112 (50): 15343–47.

Ebersole, Charles R, Maya B Mathur, Erica Baranski, Diane-Jo Bart-Plange, Nicholas R Buttrick, Christopher R Chartier, Katherine S Corker, et al. 2020. "Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability." *Advances in Methods and Practices in Psychological Science* 3 (3): 309–31.

Errington, Timothy M, Maya Mathur, Courtney K Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. 2021. "Investigating the Replicability of Preclinical Cancer Biology." *Elife* 10: e71601.

Etz, Alexander, and Joachim Vandekerckhove. 2016. "A Bayesian Perspective on the Reproducibility Project: Psychology." *PloS One* 11 (2): e0149794.

Frank, Michael C, and Rebecca Saxe. 2012. "Teaching Replication." *Perspectives on Psychological Science* 7: 595–99.

Frank, Michael C, Jonathan A Slemmer, Gary F Marcus, and Scott P Johnson. 2013. ""Information from Multiple Modalities Helps 5-Month-Olds Learn Abstract Rules": Erratum." *Developmental Science* 16 (2): 324. https://doi.org/doi.org/10.1111/desc.12060.

Gelman, Andrew. 2018. "Don't Characterize Replications as Successes or Failures." *Behavioral and Brain Sciences* 41.

Gilbert, Daniel T, Gary King, Stephen Pettigrew, and Timothy D Wilson. 2016. "Comment on 'Estimating the Reproducibility of Psychological Science'." *Science* 351 (6277): 1037–37.

Gould, Stephen Jay. 1991. "Royal Shorthand." *Science* 251 (4990): 142–42.

———. 1996. *The Mismeasure of Man*. WW Norton & company.

Hardwicke, Tom E, Manuel Bohn, Kyle MacDonald, Emily Hembacher, Michèle B Nuijten, Benjamin N Peloquin, Benjamin E deMayo, Bria Long, Erica J Yoon, and Michael C Frank. 2021. "Analytic Reproducibility in Articles Receiving Open Data Badges at the Journal Psychological Science : An Observational Study." *Royal Society Open Science*.

Hardwicke, Tom E, Maya B Mathur, Kyle Earl MacDonald, Gustav Nilsonne, George Christopher Banks, Mallory Kidwell, Alicia Hofelich Mohr, et al. 2018. "Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the Journal Cognition."

Hardwicke, Tom E, Robert T. Thibault, Jessica Kosie, Joshua D. Wallach, Mallory C. Kidwell, and John Ioannidis. 2021. "Estimating the Prevalence of Transparency and Reproducibility-Related Research Practices in Psychology (2014-2017)." *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620979806.

John, Leslie K, George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling." *Psychological Science* 23 (5): 524–32.

Klein, Richard A, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, et al. 2018. "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1 (4): 443–90.

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Princeton University Press.

Lewis, Molly L, and Michael C Frank. 2016. "Understanding the Effect of Social Context on Learning: A Replication of Xu and Tenenbaum (2007b)." *Journal of Experimental Psychology: General* 145 (9): e72.

Mathur, Maya B, and Tyler J VanderWeele. 2020. "New Statistical Metrics for Multisite Replication Projects." *J. R. Stat. Soc. Ser. A Stat. Soc.* 183 (3): 1145–66.

McShane, Blakeley B, and Ulf Böckenholt. 2014. "You Cannot Step into the Same River Twice: When Power Analyses Are Optimistic." *Perspectives on Psychological Science* 9 (6): 612–25.

Nelson, Leif D, Joseph Simmons, and Uri Simonsohn. 2018. "Psychology's Renaissance." *Annual Review of Psychology* 69: 511–34.

Newell, Allen. 1973. "You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of This Symposium." In *Visual Information Processing*, edited by W G Chase. Academic Press.

Nosek, Brian A, Tom E Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S Corker, Anna Dreber Almenberg, Fiona Fidler, et al. 2021. "Replicability, Robustness, and Reproducibility in Psychological Science." *Annual Review of Psychology*.

Nuijten, Michèle B, Chris H J Hartgerink, Marcel A L M van Assen, Sacha Epskamp, and Jelte M Wicherts. 2016. "The Prevalence of Statistical Reporting Errors in Psychology (1985–2013)." *Behavior Research Methods* 48 (4): 1205–26.

Oberauer, Klaus, and Stephan Lewandowsky. 2019. "Addressing the Theory Crisis in Psychology." *Psychonomic Bulletin & Review* 26 (5): 1596–1618.

Olsson-Collentine, Anton, Jelte M Wicherts, and Marcel ALM van Assen. 2020. "Heterogeneity in Direct Replications in Psychology and Its Association with Effect Size." *Psychological Bulletin* 146 (10): 922.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251).

Popper, Karl. 2005. *The Logic of Scientific Discovery*. Routledge.

Ramscar, Michael. 2016. "Learning and the Replicability of Priming Effects." *Current Opinion in Psychology* 12: 80–84.

Ranehill, Eva, Anna Dreber, Magnus Johannesson, Susanne Leiberg, Sunhae Sul, and Roberto A Weber. 2015. "Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women." *Psychological Science* 26 (5): 653–56.

Rohrer, Doug, Harold Pashler, and Christine R Harris. 2015. "Do Subtle Reminders of Money Change People's Political Views?" *Journal of Experimental Psychology: General* 144 (4): e73.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86 (3): 638.

———. 1990. "Replication in Behavioral Research." *Journal of Social Behavior and Personality* 5 (4): 1.

Scheel, Anne M, Mitchell RMJ Schijen, and Daniël Lakens. 2021. "An Excess of Positive Results: Comparing the Standard Psychology Literature with Registered Reports." *Advances in Methods and Practices in Psychological Science* 4 (2): 25152459211007467.

Schmidt, Stefan. 2009. "Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences." *Review of General Psychology* 13: 90–100.

Schwarz, Norbert, and Gerald L Clore. 1983. "Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States." *Journal of Personality and Social Psychology* 45 (3): 513.

Schwarz, Norbert, and Fritz Strack. 2014. "Does Merely Going Through the Same Moves Make for a 'Direct' Replication? Concepts, Contexts, and Operationalizations." *Social Psychology* 45 (4): 305–6.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66.

———. 2018. "False-Positive Citations." *Perspectives on Psychological Science* 13 (2): 255–59.

Simmons, Joseph P, and Uri Simonsohn. 2017. "Power Posing: P-Curving the Evidence." *Psychological Science*.

Simonsohn, Uri. 2015. "Small Telescopes: Detectability and the Evaluation of Replication Results." *Psychol. Sci.* 26 (5): 559–69.

Vadillo, Miguel A, Tom E Hardwicke, and David R Shanks. 2016. "Selection Bias, Vote Counting, and Money-Priming Effects: A Comment on Rohrer, Pashler, and Harris (2015) and Vohs (2015)." *Journal of Experimental Psychology: General*.

Van Bavel, Jay J, Peter Mende-Siedlecki, William J Brady, and Diego A Reinero. 2016. "Contextual Sensitivity in Scientific Reproducibility." *Proceedings of the National Academy of Sciences* 113 (23): 6454–59.

Vazire, Simine. 2018. "Implications of the Credibility Revolution for Productivity, Creativity, and Progress." *Perspectives on Psychological Science* 13 (4): 411–17.

Whitaker, K. 2017. "Publishing a Reproducible Paper." Presentation. https://doi.org/https://doi.org/10.6084/m9.figshare.5440621.v2.

Wicherts, Jelte M, Denny Borsboom, Judith Kats, and Dylan Molenaar. 2006. "The Poor Availability of Psychological Research Data for Reanalysis." *American Psychologist* 61 (7): 726.

Wilson, Brent M, Christine R Harris, and John T Wixted. 2020. "Science Is Not a Signal Detection Problem." *Proceedings of the National Academy of Sciences* 117 (11): 5559–67.

Yang, Yang, Wu Youyou, and Brian Uzzi. 2020. "Estimating the Deep Replicability of Scientific Findings Using Human and Artificial Intelligence." *Proceedings of the National Academy of Sciences* 117 (20): 10762–68.

Yarkoni, Tal. 2020. "The Generalizability Crisis." *Behavioral and Brain Sciences* 45: 1–37.

Youyou, Wu, Yang Yang, and Brian Uzzi. 2023. "A Discipline-Wide Investigation of the Replicability of Psychology Papers over the Past Two Decades." *Proceedings of the National Academy of Sciences* 120 (6): e2208863120.

Zwaan, Rolf Antonius, Alexander Etz, Richard E Lucas, and Brent Donnellan. 2018. "Making Replication Mainstream."

4 Ethics

Learning goals

- Distinguish between consequentialist, deontological, and virtue ethics frameworks
- Identify key ethical issues in performing experimental research
- Discuss ethical responsibilities in analysis and reporting of research
- Describe ethical arguments for open science practices

The fundamental thesis of this book is that experiments are the way to estimate causal effects, which are the foundations of theory. And as we discussed in Chapter 1, the reason why experiments allow for strong causal inferences is because of two ingredients: a manipulation – in which the experimenter changes the world in some way – and randomization. Put a different way, experimenters learn about the world by randomly deciding to do things to their participants! Is that even allowed?

Experimental research raises a host of ethical issues that deserve consideration. What can and can't we do to participants in an experiment, and what considerations do we owe to them by virtue of their decision to participate? To facilitate our discussion of these issues, we start by briefly introducing the standard philosophical frameworks for ethical analysis. We then use those to discuss problems of experimental ethics, first from the perspective of participants and then second from the perspective of the scientific ecosystem more broadly. We end with an ethical argument for TRANSPARENCY.

We have placed this chapter in the Foundations section of the book because we think it's critical to start the conversation about your ethical responsibilities as an experimentalist and researcher even before you start planning a study. We'll come back to the ethical frameworks we describe here in Chapter 12, which deals specifically with participant recruitment and the informed consent process.

Case study

Shock treatment

A decade after surviving prisoners were liberated from the last concentration camp, Adolf Eichmann, one of the Holocaust's primary masterminds, was tried for his role in the mass genocide (Baade 1961). While reflecting on his rationale for forcibly removing, torturing, and eventually murdering millions of Jews, an unrepentant Eichmann claimed that he was "merely a cog in the machinery that carried out the directives of the German Reich" and therefore was not directly responsible (Kilham and Mann 1974). This startling admission gave a young researcher an interesting idea: "Could it be that Eichmann and his million accomplices in the Holocaust were just following

orders? Could we call them all accomplices?" (Milgram 1974).

Stanley Milgram aimed to make a direct test of whether people would comply under the direction of an authority figure no matter how uncomfortable or harmful the outcome. He invited participants into the laboratory to serve as a teacher for an activity (Milgram 1963). Participants were told that they were to administer electric shocks of increasing voltage to another participant, the student, in a nearby room whenever the student provided an incorrect response. In reality, there were no shocks, and the student was an actor who was in on the experiment and only pretended to be in pain when the 'shocks' were administered. Participants were encouraged to continue administering shocks despite clearly audible pleas from the student to stop. In one of Milgram's studies, nearly 65% of participants administered the maximum voltage to the student.

This deeply unsettling result has become, as Ross and Nisbett (2011) say, "part of our society's shared intellectual legacy," informing our scientific and popular conversation in myriad different ways. At the same time, modern re-analyses of archival materials from the study have called into question whether the deception in the study was effective, casting doubt on its central findings (Perry et al. 2020).

Regardless of its scientific value, Milgram's study blatantly violates modern ethical norms around the conduct of research. Among other violations, the procedure involved **coercion** that undermined participants' right to withdraw from the experiment. This coercion appeared to have negative consequences: Milgram noted that a number of his participants displayed anxiety symptoms and nervousness. This observation was distressing and led to calls for this sort of research to be declared unethical (e.g., Baumrind 1964). The ethical issues surrounding Milgram's study are complex, and some are relatively specific to the particulars of his study and moment (Miller 2009). But the controversy around the study was an important part of convincing the scientific community to adopt stricter policies that protect study participants from unnecessary harm.

4.1 Ethical frameworks

Was Milgram's experiment (see Case Study) really ethically wrong – in the sense that it should not have been performed? You might have the intuition that is was unethical, due to the harms that the participants experienced or the way they were (sometimes) deceived by the experimenter. Others might consider arguments in defense of the experiment, perhaps that what we learned from it was sufficiently valuable to justify its being conducted. Beyond simply arguing back and forth, how could we approach this issue more systematically?

Ethical frameworks offer tools for analyzing such situations. In this section, we'll introduce three of the most commonly used frameworks and discuss how each of these could be applied to Milgram's paradigm.

4.1.1 Consequentialist theories

Ethical theories provide principles for what constitute good actions. The simplest theory of good actions is the **consequentialist** theory: good actions lead to good results. The most famous consequentialist position is the **utilitarian position**, originally defined by the philosopher John

Stuart Mill (Flinders 1992). This view emphasizes decision-making based on the "greatest happiness principle", or the idea that an action should be considered morally good based on the degree of happiness or pleasure people experience because of it, and likewise that an action should be considered morally bad based on the degree of unhappiness or pain people experience by the same action (Mill 1859).

A consequentialist analysis of Milgram's study considers the study's negative and positive effects and weighs these against one another. Did the study cause harm to its participants? On the other hand, did the study lead to knowledge that prevented harm or caused positive benefits?

Consequentialist analysis can be a straightforward way to justify the risks and benefits of a particular action, but in the research setting it is unsatisfying. Many horrifying experiments would be licensed by a consequentialist analysis and yet feel untenable to us. Imagine a researcher forced you to undergo a risky and undesired medical intervention because the resulting knowledge might benefit thousands of others. This experiment seems like precisely the kind of thing our ethical framework should rule out!

4.1.2 Deontological approaches

Harmful research performed against participants' will or without their knowledge is repugnant; we consider the Tuskegee Syphilis Experiment, a horrifying example of such research (Case Study, below). Considering such cases, a few rules seem obvious, for example: "researchers must ask participants' permission before conducting research on them." Principles like this one are now formalized in all ethical codes for research. They exemplify an approach called **deontological** (or duty-based) ethics.

Deontology emphasizes the importance of taking ethically permissible actions, regardless of their outcome (Biagetti, Gedutis, and Ma 2020). In general, university ethics boards take a deontological approach to ethics (Boser 2007). In the context of research, there are four primary principles being applied:

(1) **Respect for autonomy**. This principle requires that people participating in research studies can make their own decisions about their participation, and that those with diminished autonomy (children, neuro-divergent people, etc.) should receive equal protections (Beauchamp, Childress, et al. 2001). Respecting someone's autonomy also means providing them with all the information they need to make an informed decision about

whether to participate in a research study (giving **consent**) and giving them further context about the study they have participated in after it is done (**debriefing**).

(2) **Beneficence**. This principle means that researchers are obligated to protect the well-being of participants for the duration of the study. Beneficence has two parts. The first is to do no harm. Researchers must take steps to minimize the risks to participants and to disclose any known risks at the onset. If risks are discovered during participation, researchers must notify participants of their discovery and make reasonable efforts to mitigate these risks, even if that means stopping the study altogether. The second is to maximize potential benefits to participants.[1]

(3) **Nonmaleficence**. This principle is similar to beneficence (in fact, beneficence and nonmaleficence were a single principle when they were first introduced in the Belmont Report, which we'll discuss later) but differs in its emphasis on doing/causing no harm. In general, harm is bad – but deontology is about intent, not impact, so harm is sometimes warranted when the intent is morally good. For example, administering a vaccine may cause some discomfort and pain, but the intent is to protect the patient from developing a deadly virus in the future. The harm is justifiable under this framework.

(4) **Justice**. This principle means that both the benefits and risks of a study should be equally distributed among all participants. For example, participants should not be systematically assigned to one condition over another based on features of their identity such as socioeconomic status, race and ethnicity, or gender.

Analyzed from the perspective of these principles, Milgram's study raises several red flags. First, Milgram's study reduced participants' autonomy by making it difficult for them to voluntarily end their involvement (participants were told up to four times to continue administering shocks even after they expressed clear opposition). Second, the paradigm was designed in a way that it was likely to cause harm to its participants by putting them in a very stressful situation. Further, Milgram's study may have induced *unnecessary* harm on certain participants by failing to screen participants for existing mental health issues before beginning the session.

[1] In practice, this doesn't mean compensating participants with exorbitant amounts of money or gifts, which might cause other issues, like exerting an undue influence on low-income participants to participate. Instead "maximizing benefits" is interpreted as identifying all possible benefits of participation in the research and making them available where possible.

Depth

Was Milgram justified?

Was the harm done in Milgram's experiment justifiable given that it informed our understanding of obedience and conformity? We can't say for sure. What we can say is that in the 10 years following the publication of Milgram's study, the number of papers on (any kind of) obedience increased and the nature of these papers expanded from a focus on religious conformity to a broader interest in social conformity, suggesting that Milgram changed the direction of this research area. Additionally, in a followup that Milgram conducted, he reported that 84% of participants in the original study said they were happy to have been involved (Milgram 1974). On the other hand, given concerns about validity in the original study, perhaps its influence on the field was not warranted (Perry et al. 2020).

Many researchers believe there was no ethical way to conduct Milgram's experiment while also protecting the integrity of the research goals, but some have tried. One study recreated a portion of the original experiment, with some critical changes (Burger 2007). Before enrolling in the study, participants completed both a phone screening for mental health concerns, addiction, or extreme trauma, and a formal interview with a licensed clinical psychologist, who identified signs of depression or anxiety. Those who passed these assessments were invited into the lab for a Milgram-type learning study. Experimenters clearly explained that participation was voluntary and the decision to participate could be reversed at any point, either by the participant themselves or by a trained clinical psychologist who was present for the duration of the session. Additionally, shock administration never exceeded 150 volts (compared to 450 volts in the original study), and experimenters debriefed participants extensively following the end of the session. This modified replication study found similar patterns of obedience as Milgram's; further, one year later, no participants expressed any indication of stress or trauma associated with their involvement in the study.

4.1.1 Virtue-based Approaches

A final way that we can approach ethical dilemmas is through a virtue framework. A **virtue** is a trait, disposition, or quality that is thought to be a moral foundation (Annas 2006). Virtue ethics suggests that people can learn to be virtuous by observing those actions in others they admire (Morris and Morris 2016). Proponents of virtue ethics say this works for two reasons: (1) people are generally good at recognizing morally good traits in others and (2) people receive some fulfillment from living virtuously. Virtue ethics differs from deontology and utilitarianism because it focuses on a person's character rather than on the nature of a rule or the consequences of an action.

From a research perspective, virtue ethics tells us that in order to behave virtuously, we must make decisions that consider the context surrounding the experiment (Dillern 2021). In other words, researchers should evaluate how their studies might influence a participant's behaviors, especially when those behaviors deviate from typical expectations. This process is also meant to be adaptive, meaning that researchers must be vigilant about both the changing mental states of their participants dur-

ing the experimental session and whether the planned procedure is no longer acceptable.

How can we apply this ethical framework to Milgram's experiment? Many virtue ethicists would probably conclude that Milgram's approach was neither appropriate (for participants) nor adaptive. Upon noticing increasing levels of participant distress, an experimenter following the virtue ethics framework should have chosen to end the session early or – even better – to have minimized participant distress from the beginning.

4.2 Ethical responsibilities to research participants

Milgram's shock experiment was just one of dozens of unethical human subjects studies that garnered the attention and anger of the public in the United States. In 1978, the US National Commission for the Protection of Human Services of Biomedical and Behavioral Research released the **Belmont Report**, which described protections for the rights of human subjects participating in research studies (Adashi, Walters, and Menikoff 2018). Perhaps the most important message found in the report was the notion that "investigators should not have sole responsibility for determining whether research involving human subjects fulfills ethical standards. Others, who are independent of the research, must share the responsibility." In other words, ethical research requires both transparency and external oversight.

4.2.1 Institutional review boards

The creation of **institutional review boards** (IRBs) in the United States was an important result of the Belmont Report. While regulatory frameworks and standards vary across national boundaries, ethical review of research is ubiquitous across countries. In what follows, we focus on the US regulatory framework as it has been a model for other ethical review systems but we use the clearer label "ethics review boards" for IRBs.

An ethics board is a committee of people who review, evaluate, and monitor human subjects research to make sure that participants' rights are protected when they participate in research (Oakes 2002). Ethics boards are local; every organization that conducts human subjects or animal research is required to have its own ethics board or to contract with an external one. If you are based at a university, yours likely has its own, and its members are probably a mix of scientists, doctors, professors, and community residents.[2]

[2] The local control of ethics boards can lead to very different practices in ethical review across institutions, which is obviously inconsistent consistent with the idea that ethical standards should be uniform! In addition, critics have wondered about the structural issue that institutional ethics boards have an incentive to decrease liability for the institution, while private boards have an incentive to provide approvals to the researchwers who pay them (Lemmens and Freedman 2000).

When a group of researchers have a research question they are interested in pursuing with human subjects, they must receive approval from their local ethics board before beginning any data collection. The ethics board reviews each study to make sure:

1. A study poses no more than **minimal risk** to participants. This means the anticipated harm or discomfort to the participant is not greater than what would be experienced in everyday life. It is possible to perform a study that poses **greater than minimal risk**, but it requires additional monitoring to detect any adverse events that may occur.
2. Researchers obtain **informed consent** from participants before collecting any data. This requirement means experimenters must disclose all potential risks and benefits so that participants can make an informed decision about whether or not to participate in the study. Importantly, informed consent does not stop after participants sign a consent form. If researchers discover any new potential risks or benefits along the way, they must disclose these discoveries to all participants (see Chapter 12).
3. Sensitive information remains **confidential**. Although regulatory frameworks vary, researchers typically have an obligation to their participants to protect all identifying information recorded during the study (see Chapter 13).
4. Participants are recruited **equitably** and without **coercion**. Before ethics boards became standard, researchers often coercively recruited marginalized and vulnerable populations to test their research questions, rather than making participation in research studies voluntary and providing equitable access to the opportunity to participate.

Case study

The Tuskegee Syphilis Study

In 1929, The United States Public Health Service (USPHS) was perplexed by the effects of syphilis in Macon County, Alabama, an area with an overwhelmingly Black population (Brandt 1978). Syphilis is a sexually transmitted bacterial infection that can either be in a visible and active stage or in a latent stage. At the time of the study's inception, roughly 36% of Tuskegee's adult population had developed some form of syphilis, one of the highest infection rates in America (White 2006).

The USPHS recruited 400 Black males from 25–60 years of age with latent syphilis and 200 Black males without the infection to serve as a control group to participate (Brandt 1978). The USPHS sought the help of the Macon County Board of Health to recruit participants with the promise that they would provide treatment for community

members with syphilis. The researchers sought poor, illiterate Black people and, instead of telling them that they were being recruited for a research study, merely informed them that they would be treated for "bad blood".

Because the study was interested in tracking the natural course of latent syphilis without any medical intervention, the USPHS had no intention of providing any care to its participants. To assuage participants, the USPHS distributed an ointment that had not been shown to be effective in the treatment of syphilis, and only small doses of a medication actually used to treat the infection. In addition, participants underwent a spinal tap which was presented to them as another form of therapy and their "last chance for free treatment."

By 1955, just over 30% of the original participants had died from syphilis complications. It took until the 1970s before the final report was released and (the lack of) treatment ended. In total, 128 participants died of syphilis or complications from the infection, 40 wives became infected, and 19 children were born with the infection (Katz and Warren 2011). The damage rippled through two generations, and many never actually learned what had been done to them.

The Tuskegee experiment violates nearly every single guideline for research described above – indeed in its many horrifying violations of research participants' agency, it provides a blueprint for future regulation to prevent any aspect of it from being repeated: Investigators did not obtain informed consent. Participants were not made aware of all known risks and benefits involved with their participation. Instead, they were deceived by researchers who led them to believe that diagnostic and invasive exams were directly related to their treatment.

Perhaps most shocking, participants were denied appropriate treatment following the discovery that penicillin was effective at treating syphilis (Mahoney, Arnold, and Harris 1943). The USPHS requested that medical professionals overseeing their care outside of the research study not offer treatment to participants so as to preserve the study's methodological integrity. This intervention violated participants' rights to equal access to care, which should have taken precedence over the results of the study.

Finally, recruitment was both imbalanced and coercive. Not only were participants selected from the poorest of neighborhoods in the hopes of finding vulnerable populations with little agency, but they were also bribed with empty promises of treatment and a monetary incentive (payment for burial fees, a financial obstacle for many sharecroppers and tenant farmers at the time).

4.2.1 Risks and benefits

Imagine that you were approached about participating in a research study at your local university. You were only told you would be paid $25 in exchange for completing an hour of cognitive tasks on a computer. Now imagine that halfway through the session, the experimenter revealed they would also need to collect a blood sample, "which should only take a couple of minutes and which will really help the research study." Would you agree to the sample? Would you feel uncomfortable in any way?

Participants need to understand the risks and benefits of participation in an experiment before they give consent. To do otherwise compromises their autonomy (a key deontological principle). In the case of this hypothetical experiment, a new and unexpected invasive component of an experiment is coercive: participants would have to choose to forfeit

their expected compensation to opt out. They also might feel that they have been deceived by the experimenter.

In human subjects research, **deception** is a specific technical term that refers to cases when (1) experimenters withhold any information about its goals or intentions, (2) experimenters hide their true identity (such as when using actors), (3) some aspects of the research are under- or overstated to conceal information, or (4) participants receive any false or misleading information. The use of deception requires special consideration from a human subjects perspective (Kelman 2017; Baumrind 1985).

Even assuming they are disclosed properly without coercion or deception, the risks and benefits of a study must be assessed from the perspective of the *participant*, not the experimenter. By doing so, we allow participants to make an informed choice. In the case of the blood sample, the risks to the participant were not disclosed, and the benefits were stated in terms of the research project (and the experimenter).

The benefits of participation in research can either be direct or indirect, and it is important to specify which type participants may receive. While some clinical studies and interventions may offer some direct benefit due to participation, many of the benefits of basic science research are indirect. Both have their place in science, but participants must ultimately determine the degree to which each type of benefit motivates their own involvement in a study (Shatz 1986).

4.3 Ethical responsibilities in analysis and reporting of research

Accident report

What data?

Dutch social psychologist Diederick Stapel contributed to more than 200 articles on social comparison, stereotype threat, and discrimination, many published in the most prestigious journals. Stapel reported that affirming positive personal qualities buffered against dangerous social comparison, that product advertisements related to a person's attractiveness changed their sense of self, and that exposure to intelligent in-group members boosted a person's performance on future tasks (Stapel and Linde 2012; Trampe, Stapel, and Siero 2011; Gordijn and Stapel 2012). These findings were fresh and noteworthy at the time of publication, and Stapel's papers were cited thousands of times. The only problem? Stapel's data were made up.

Stapel has admitted that when he first began fabricating data, he would make small tweaks to a few data points (Stapel 2012). Changing a single number here and there would turn a flat study into an impressive one. Having achieved comfortable success (and having aroused little suspicion from journal editors and others in the scientific

community), Stapel eventually began creating entire data sets and passing them off as his own. Several colleagues began to grow skeptical of his overwhelming success, however, and brought their concerns to the Psychology Department at Tilburg University. By the time the investigation of his work concluded, 58 of Stapel's papers were **retracted**, meaning that the publishing journal withdrew the paper after discovering that its contents were invalid.

Everyone agrees that Stapel's behavior was deeply unethical. But should we consider cases of falsification and fraud to be different in kind from other ethical violations in research? Or is fraud merely the endpoint in a continuum that might include other practices like p-hacking? Lawyers and philosophers grapple with the precise boundary between sloppiness and neglect, and it can be difficult to know which one is at play when a typo or coding mistake changes the conclusion of a scientific paper. Similarly, if a researcher engages in so-called "questionable research practices," at what point should they be considered to have made an ethical violation as opposed to simply performing their research poorly?

The ethical frameworks above provide a framework for thinking about this topic. For the consequentialist, sloppy science can lead to good outcomes for the scientist (quicker publication) but bad outcomes for the rest of the scientific community who have to waste time and effort on papers that may not be correct. For the deontologist, the scientist's intention plays a key role: it is not a generally acceptable principle to knowingly use sub-standard practices. And for the virtue ethicist, sloppiness is not a morally good trait. On all analyses, researchers have a duty to pursue their work carefully.

As scientists, we not only have a responsibility to participants, we are also responsible for what we do with our data and for the kinds of conclusions we draw. Cases like Stapel's (see Accident Report) seem stunning, but they are part of a continuum. Codes of professional ethics for organizations like the American Psychological Association encourage researchers to take care in the management and analysis of their data so as to avoid errors and misstatements (Association 2022).

Researchers also have an obligation not to suppress findings based on their own beliefs about the right answer. One unfortunate way that this suppression can happen is when researchers selectively report their research, leading to **publication bias**, as you learned in Chapter 3. Researchers' own biases can be another (invalid) rationale for not publishing: it's also an ethical violation to suppress findings that contradict your theoretical commitments.

Importantly, researchers don't have an obligation to publish *everything* they do. Publishing in the peer-reviewed literature is difficult and time-consuming. There are plenty of reasons not to publish an experimental finding! For example, there's no reason to publish a result if you believe it is truly uninformative because of a confound in the experimental design. You also aren't typically committing an ethical violation if you decide to quit your job in research and so you don't publish a study from your dissertation.[3] The primary ethical issue arises when you use the *result* of a study – and how it relates to your own beliefs or to a threshold like $p < .05$ – to decide whether to publish it or not.

As we'll discuss again and again in this book, the preparation of research reports must also be done with care and attention to detail (see Chapter 14). Sloppiness in writing up results can lead to imprecise or overbroad claims; and if that sloppiness extends to the reporting of data, and analysis, it may lead to irreproducibility as well.

Further, professional ethics dictate that published contributions to the literature be original. In general, the text of a paper must not be **plagiarized** (copied) from the text of other reports whether by you or by another author without attribution. Copying from others outside of a direct, attributed quotation is obviously an ethical violation because it leads to credit for text being given to you rather than the true author. But self-plagiarism is also not acceptable – it is a violation to receive credit multiple times for the same product.[4]

[4] Standards on this issue differ from field to field. Our sense is that the rule on self-plagiarism applies primarily to duplication of content between journal papers. So, for example, barring any specific policy of the funder or journal, it is acceptable to use text from one of your own grant proposals in a journal paper. It is also typically acceptable to reuse text from a conference abstract or preregistration (that you wrote, of course) when prepare a journal paper.

4.4 Ethical responsibilities to the broader scientific community

The open science principles that we will describe throughout this book are not only important correctives to issues of reproducibility and replicability, they are also ethical duties.

The sociologist Robert Merton described a set of norms that science is assumed to follow: communism – that scientific knowledge belongs to the community; universalism – that the validity of scientific results is independent of the identity of the scientists; disinterestedness – that scientists and scientific institutions act for the benefit of the overall enterprise; and organized skepticism – that scientific findings must be critically evaluated (Merton 1979).

If the products of science aren't open, it is very hard to be a scientist by Merton's definition. To contribute to the communal good, papers need to be openly available. And to be subject to skeptical inquiry, experimental materials, research data, analytic code, and software must be all available so that analytic calculations can be verified and experiments can be reproduced. Otherwise, you have to accept arguments on authority rather than by virtue of the materials and data.

Openness is not only definitionally part of the scientific enterprise, it's also good for science and individual scientists (Gorgolewski and Poldrack 2016). Open access publications are cited more (Eysenbach 2006; Gargouri et al. 2010). Open data also increases the potential for citation and reuse, and maximizes the chances that errors are found and corrected.

But these benefits mean that researchers have a responsibility to their funders to pursue open practices so as to seek the maximal return on funders' investments. And by the same logic, if research participants contribute their time to scientific projects, the researchers also owe it to these participants to maximize the impact of their contributions (Brakewood and Poldrack 2013). For all of these reasons, individual scientists have a duty to be open – and scientific institutions have a duty to promote transparency in the science they support and publish.

How should these duties be balanced against researchers' other responsibilities? For example, how should we balance the benefit of data sharing against the commitment to preserve participant privacy? And, since transparency policies also carry costs in terms of time and effort, how should researchers consider those costs against other obligations?

First, open practices should be a default in cases where risks and costs are limited. For example, the vast majority of journals allow authors to post accepted manuscripts in their un-typeset form to an open repository. This route to "green" open access is easy, cost free, and – because it comes only after articles are accepted for publication – confers essentially no risks of scooping. As a second example, the vast majority of analytic code can be posted as an explicit record of exactly how analyses were conducted, even if posting data is sometimes more complicated due to privacy restrictions. These kinds of "incentive compatible" actions towards openness can bring researchers much of the way to a fully transparent workflow, and there is no excuse not to take them.

Second, researchers should plan for sharing and build a workflow that decreases the costs of openness. As we discuss in Chapter 13, while it can be costly and difficult to share data after the fact if they were not explicitly prepared for sharing, good project management practices can make this process far simpler (and in many cases completely trivial).

Finally, given the ethical imperative towards openness, institutions like funders, journals, and societies need to use their role to promote open practices and to mitigate potential negatives (Nosek et al. 2015). Scholarly societies have an important role to play in educating scientists about the benefits of openness and providing resources to steer their members towards best practices for sharing their publication and other research products. Similarly, journals can set good defaults, for example by requiring data and code sharing except in cases where a strong justification is given. Funders of research can – and increasingly, do – signal their interest in openness through data sharing mandates.

4.5 Chapter summary: Ethics

In this chapter, we discussed three ethical frameworks and evaluated how they can be applied to our own research through the lens of Milgram's famous obedience experiment. Studies like Milgram's prompted serious conversations about how best to reconcile experimenter goals with participant well-being. The publication of the Belmont Report and later creation of ethics boards in the United States standardized the way scientists approach human subjects research, and created much-needed accountability. We also addressed our ethical responsibilities to the scientific community, both in how we report our data and how we distribute it. We hope that we have convinced you that careful, open science is an ethical imperative for researchers!

Discussion questions

1. The COVID-19 pandemic led to an immense amount of "rapid response" research in psychology that aimed to discover – and influence – the way people reasoned about contagion, vaccines, masking, and other aspects of the public health situation. What are the specific ethical concerns that researchers should be aware of for this type of research? Are there reasons for more caution in this kind of research than in other "run of the mill" research?
2. Think of an argument against open science practices – for example, that following open science practices is especially burdensome for researchers with more limited resources (you can make up another if you want!). Given our argument that researchers have an ethical duty to openness, how would you analyze this argument under the three different ethical frameworks we discussed?

Readings

- The Belmont Report has shaped US research ethics policy from its publication to the present day. It's also short and quite readable: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html.
- A rich reference with several case studies on science misconduct and with strong arguments for open science: Ritchie, S. (2020). Science fictions: How fraud, bias, negligence, and hype undermine the search for truth. Metropolitan Books.

Part II

Statistics

References

Adashi, Eli Y, LeRoy B Walters, and Jerry A Menikoff. 2018. "The Belmont Report at 40: Reckoning with Time." *American Journal of Public Health* 108 (10): 1345–48.

Annas, Julia. 2006. "Virtue Ethics." *The Oxford Handbook of Ethical Theory*, 515–36.

Association, American Psychological. 2022. "Ethical Principles of Psychologists and Code of Conduct." 2022. https://www.apa.org/ethics/code.

Baade, Hans W. 1961. "The Eichmann Trial: Some Legal Aspects." *Duke LJ*, 400.

Baumrind, Diana. 1964. "Some Thoughts on Ethics of Research: After Reading Milgram's" Behavioral Study of Obedience."." *American Psychologist* 19 (6): 421.

———. 1985. "Research Using Intentional Deception: Ethical Issues Revisited." *American Psychologist* 40 (2): 165.

Beauchamp, Tom L, James F Childress, et al. 2001. *Principles of Biomedical Ethics*. Oxford University Press, USA.

Biagetti, Maria Teresa, Aldis Gedutis, and Lai Ma. 2020. "Ethical Theories in Research Evaluation: An Exploratory Approach." *Scholarly Assessment Reports*.

Boser, Susan. 2007. "Power, Ethics, and the IRB: Dissonance over Human Participant Review of Participatory Research." *Qualitative Inquiry* 13 (8): 1060–74.

Brakewood, Beth, and Russell A Poldrack. 2013. "The Ethics of Secondary Data Analysis: Considering the Application of Belmont Principles to the Sharing of Neuroimaging Data." *Neuroimage* 82: 671–76.

Brandt, Allan M. 1978. "Racism and Research: The Case of the Tuskegee Syphilis Study." *Hastings Center Report*, 21–29.

Burger, Jerry. 2007. "Replicating Milgram." *APS Observer* 20 (11).

Dillern, Thomas. 2021. "The Scientific Judgment-Making Process from a Virtue Ethics Perspective." *Journal of Academic Ethics* 19 (4): 501–16.

Eysenbach, Gunther. 2006. "Citation Advantage of Open Access Articles." *PLoS Biology* 4 (5): e157.

Flinders, David J. 1992. "In Search of Ethical Guidance: Constructing a Basis for Dialogue." *International Journal of Qualitative Studies in Education* 5 (2): 101–15.

Gargouri, Yassine, Chawki Hajjem, Vincent Larivière, Yves Gingras, Les Carr, Tim Brody, and Stevan Harnad. 2010. "Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research." *PloS One* 5 (10): e13636.

Gordijn, Ernestine H, and Diederik A Stapel. 2012. "Behavioural Effects of Automatic Interpersonal Versus Intergroup Social Comparison (Retraction of Vol 45, Pg 717, 2006)." *British Journal of Social Psychology* 51 (3): 498–98.

Gorgolewski, Krzysztof J., and Russell A. Poldrack. 2016. "A Practical Guide for Improving Transparency and Reproducibility in Neuroimaging Research." *PLOS Biology* 14 (7): e1002506. https://doi.org/10.1371/journal.pbio.1002506.

Katz, Ralph V, and Rueben C Warren. 2011. *The Search for the Legacy of the USPHS Syphilis Study at Tuskegee*. Lexington Books.

Kelman, Herbert C. 2017. "Human Use of Human Subjects: The Problem of Deception in Social Psychological Experiments." In *Research Design*, 189–204. Routledge.

Kilham, Wesley, and Leon Mann. 1974. "Level of Destructive Obedience as a Function of Transmitter and Executant Roles in the Milgram Obedience Paradigm." *Journal of Personality and Social Psychology* 29 (5): 696.

Lemmens, Trudo, and Benjamin Freedman. 2000. "Ethics Review for Sale? Conflict of Interest and Commercial Research Review Boards." *The Milbank Quarterly* 78 (4): 547–84.

Mahoney, John F, RC Arnold, and AD Harris. 1943. "Penicillin Treatment of Early Syphilis—a Preliminary Report." *American Journal of Public Health and the Nations Health* 33 (12): 1387–91.

Merton, Robert K. 1979. "The Normative Structure of Science." *The Sociology of Science: Theoretical and Empirical Investigations*, 267–78.

Milgram, Stanley. 1963. "Behavioral Study of Obedience." *The Journal of Abnormal and Social Psychology* 67 (4): 371.

———. 1974. *Obedience to Authority: An Experimental View*. Harper & Row.

Mill, John Stuart. 1859. "Utilitarianism (1863)." *Utilitarianism, Liberty, Representative Government*, 7–9.

Miller, Arthur G. 2009. "Reflections on "Replicating Milgram" (Burger, 2009)." *American Psychologist* 64 (1): 20–27. https://doi.org/https://doi.org/10.1037/a0014407.

Morris, Marilyn C, and Jason Z Morris. 2016. "The Importance of Virtue Ethics in the IRB." *Research Ethics* 12 (4): 201–16.

Nosek, Brian A, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. "Promoting an Open Research Culture." *Science* 348 (6242): 1422–25. https://doi.org/10.1126/science.aab2374.

Oakes, J Michael. 2002. "Risks and Wrongs in Social Science Research: An Evaluator's Guide to the IRB." *Evaluation Review* 26 (5): 443–79.

Perry, Gina, Augustine Brannigan, Richard A Wanner, and Henderikus Stam. 2020. "Credibility and Incredulity in Milgram's Obedience Experiments: A Reanalysis of an Unpublished Test." *Social Psychology Quarterly* 83 (1): 88–106.

Ross, Lee, and Richard E Nisbett. 2011. *The Person and the Situation: Perspectives of Social Psychology*. Pinter & Martin Publishers.

Shatz, David. 1986. "Autonomy, Beneficence, and Informed Consent: Rethinking the Connections." *Cancer Investigation* 4 (3): 257–69.

Stapel, Diederik A. 2012. *Ontsporing*. Prometheus Amsterdam.

Stapel, Diederik A, and Lonneke AJG van der Linde. 2012. ""What Drives Self-Affirmation Effects? On the Importance of Differentiating Value Affirmation and Attribute Affirmation": Retraction of Stapel and van Der Linde (2011)." *Journal of Personality and Social Psychology* 103 (3): 505. https://doi.org/https://doi.org/10.1037/a0029745.

Trampe, Debra, Diederik A Stapel, and Frans W Siero. 2011. "Retracted: The Self-Activation Effect of Advertisements: Ads Can Affect Whether and How Consumers Think about the Self." *Journal of Consumer Research* 37 (6): 1030–45.

White, Robert M. 2006. "Effects of Untreated Syphilis in the Negro Male, 1932 to 1972: A Closure Comes to the Tuskegee Study, 2004." *Urology* 67 (3): 654.

5 Estimation

Learning goals

- Estimate the causal effect of a manipulation
- Discuss differences between frequentist and Bayesian estimation
- Reason about standardized effect sizes and their strengths and weaknesses
- Quantify the relationship between variables

> "In every quantitative paper we read, every quantitative talk we attend, and every quantitative article we write, we should all ask one question: *what is the estimand*? The estimand is the object of inquiry – it is the precise quantity about which we marshal data to draw an inference. Yet, too often social scientists skip the step of defining the estimand. Instead, they leap straight to describing the data they analyze and the statistical procedures they apply. Without a statement of the estimand, it becomes impossible for the reader to know whether those procedures were appropriate." (Lundberg, Johnson, and Stewart 2021)

In the first section of this book, our goal was to set up some of the theoretical ideas that motivate our approach to experimental design and planning. We introduced our key thesis, namely that experiments are about measuring causal effects. We also began to discuss our key themes, TRANSPARENCY, MEASUREMENT PRECISION, BIAS REDUCTION, and GENERALIZABILITY.

In this next section of the book – treating statistical topics – we will integrate these ideas with an analytic toolkit for **estimating** effects and quantifying their size (this chapter), making **inferences** about how these estimates relate to a population (Chapter 6), and building **models** for estimation and inference in more complex settings (Chapter 7). Although this book does not provide an extensive treatment of statistics, we hope that these chapters provide a foundations – and an opinionated perspective – for beginning the statistical analysis of your experimental data, with a focus on MEASUREMENT PRECISION.

Case study

The Lady Tasting Tea

The birth of modern statistical inference arose from the age old conundrum of how to best make a cup of tea. The statistician Ronald Fisher was apparently at an afternoon tea party when a lady declared that she could tell the difference when tea was added to milk vs. milk to tea. Rather than taking her at her word, Fisher devised an experimental and data analysis procedure to test her claim.

The lady would have to judge a set of six new cups of tea and sort them into milk-first vs. tea-first sets. Her data would then be analyzed to determine whether her level of correct choice exceeded that expected by chance. While this process now sounds like a quotidian experiment that might be done on a cooking reality show, it seems unremarkable in hindsight only because it set the standard for the way science was done going forward.

The important and unusual element of the experiment was its treatment of potential design confounds such which cup of tea was prepared first, which cup of tea was presented first, or the material that the cups were made out of. Prior experimental practice would have been to try to equate all of the cups as closely as possible, decreasing the influence of confounds. Fisher recognized that this strategy was insufficient because of the presence of unobserved confounders. Only by randomizing all other aspects of the experiment could he make strong causal inferences about the treatment (milk then tea vs. tea then milk). We discussed the causal power of random assignment in Chapter 1 – the Lady Tasting Tea experiment is a key touchstone in the popularization of randomized experiments!

5.1 Estimating a quantity

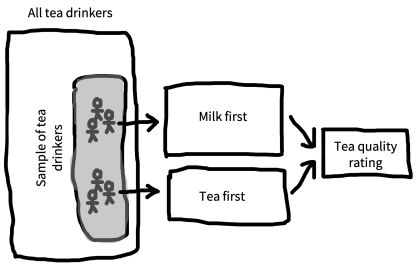


Figure 5.1: The structure of our tea tasting experiment.

If experiments are about estimating effects, how do we actually use our experimental data to make these estimates? For our example we'll design a slightly more modern version of Fisher's experiment, shown in Figure 5.1.

Our causal theory is that the tea quality is affected by milk-tea ordering, so we'll test that by rating tea quality both milk-first and tea-first, represented by a DAG like the one in Figure 5.2. Our intended population to generalize to is the set of all tea drinkers, and towards that goal we sample a set of tea-drinkers. In practice, we might do a field trial in a cafe in which we approach patrons and ask them to participate in our experiment in exchange for a free cup of tea. Although this sample size

An important piece of context for the work of Ronald Fisher, Karl Pearson, and other early pioneers of statistical inference is that they were all strong proponents of eugenics. Fisher was the founding Chairman of the Cambridge Eugenics Society. Pearson was perhaps even worse, an avowed Social Darwinist who believed fervently in Eugenic legislation. These views are repugnant and provide important context for their statistical contributions.

is almost certainly too small to get precise estimates, for the purpose of this example, we'll sample 18 tea drinkers – nine in each condition.

As our manipulation, we follow Fisher in randomly assigning participants (who of course should give consent to participate) into to one of our two conditions: milk-first and tea-first.[1] This design is a between-participants design, so each participant gets only one cup of tea. They receive their cup of tea and taste it. Then as our measure, we ask for a rating of the tea on a continuous scale from 1 (terrible) to 7 (delicious).[2]

An example dataset from our experiment is shown in Figure 5.3. Eventually, we'll want to estimate the effect of milk-first preparation on quality ratings (our effect of interest). But for now, our goal will be to estimate the quality of the tea when it is milk-first [which some data suggest is actually the better way, at least for British tea drinkers; Kennedy (2003)]. More formally, we want to use our sample of 9 milk-first tea judgments to estimate a number that we can't directly observe, namely the true perceived quality of all possible milk-first cups. We'll call this number a **population parameter** for reasons that will become clear in a moment.

We'll try to go easy on notation but some amount will hopefully make things clearer. We will use θ_{M} ("theta") to denote the parameter we want to estimate (the **population parameter**) and $\hat{\theta}_{\mathrm{M}}$, its **sample estimate**.[3]

5.1.1 Maximum likelihood estimation

OK, you are probably saying, if we want our estimate of milk-first quality, shouldn't we just take the average rating across the 9 cups of milk-first tea? The answer is yes. But let's unpack that choice: taking the sample mean as our estimate $\hat{\theta}_{\mathrm{M}}$ is an example of an estimation approach called **maximum likelihood estimation**. In general terms, maximum likelihood estimation is a two-step process.

First, we assume a **model** for how the data were generated.[4] This model is specified in terms of certain population parameters. In our example, the model is as simple as they come: we just assume there is some average level of tea quality and that the measurements vary around it. Let's take a look at the data from the milk-first condition, shown in Figure 5.4. Our observations are clustered around the mean, but they also show some variation. Some are higher and some are lower. Variation

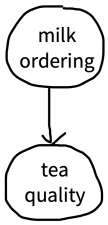


Figure 5.2: A directed acyclic graph representing our causal theory of tea quality.

[1] Technically, randomized experiments were not invented by Fisher. Perhaps the earliest example of a (somewhat) randomized experiment was a trial of scurvy treatments in the 1700s (Dunn 1997). Peirce and Jastrow (1884) also report a strikingly modern use of randomized stimulus presentation (via shuffling cards). Nevertheless, Fisher's statistical work popularized randomized experiments throughout the sciences, in part by integrating them with a set of analytic methods.

[2] Right now we're going to assume that our ratings are just simple numerical values and not worry about the fact that they come from a rating scale that is bounded (e.g., can't go above 7). If you're curious about **Likert scales** (the name for discrete numerical rating scales; pronounced LICK-ERT), we'll talk a bit more about them in Chapter 8.

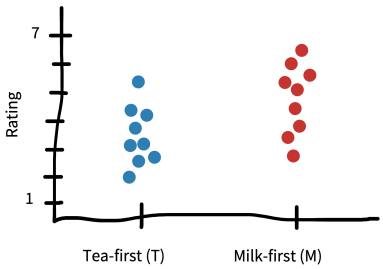


Figure 5.3: Schematic data from the tea tasting experiment.

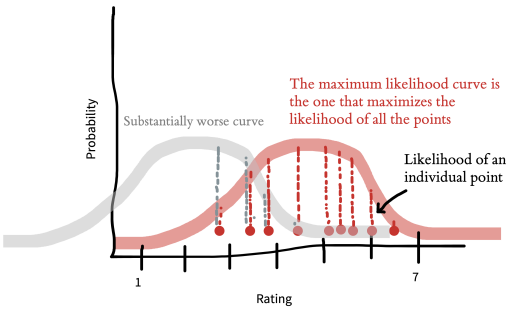
[3] Statisticians use "hats" like this to denote estimates from a specific sample. One way to remember this is that the "person in the hat" is wearing a hat to dress up as the actual quantity.

of this type is a feature of every data set. This variation can be summarized via a **probability distribution**, a mathematical entity that describes the properties of possible datasets.

The only probability distribution we'll discuss here is the ubiquitous **normal distribution** (also sometimes called a "Gaussian distribution"). A normal distribution has two **parameters** (numbers that define its shape), a **mean** and a **standard deviation**. These two parameters define the shape of the curve. The mean (θ_M) describes where its center goes, and the standard deviation describes how wide it is. Technically, the mean is the **expected value** for new samples from the distribution. Our best guess about the value of these new samples is that they are at the mean. We can write this more formally by introducing $E[M]$ to denote the expectation of the variable M.

The standard deviation σ_M is then a way of describing the expected *variation* in these samples. A bigger standard deviation means that we expect samples to be on average further from the mean. We can write this formally as $\sigma_M = \sqrt{E[(M - \theta_M)^2]}$: the standard deviation is the expected absolute distance between individual samples and the mean, with the square and square root being necessary to compute distance.

Using a probability distribution to describe our dataset gives us a way of summarizing our observations through the parameters of the distribution and encoding an assumption about what future observations might look like. How do we fit a normal distribution to our data? We try to find the values of the population parameters that make our observed data as likely as possible. Let's start with the mean.



For example, if our sample mean is $\hat{\theta}_{\mathrm{M}} = 4.5$, what underlying value of θ_{M} would make these data most likely to occur? Well, suppose the underlying parameter were $\theta_{\mathrm{M}} = 2.5$. Then it would be pretty unlikely that our sample mean would be so much bigger. So $\hat{\theta}_{\mathrm{M}} = 2.5$ is a poor estimate of the population parameter based on these data (Figure 5.5). Conversely, if the parameter were $\theta_{\mathrm{M}} = 6.5$, it would be a bit unlikely that our sample mean would be so much *smaller*. The value of $\hat{\theta}_{\mathrm{M}}$ that

[4] This sense of "model" is actually a formal instantiation of the type of causal model we discussed in Chapter 1. As you get deeper into causal modeling, typically what you do is define a causal "story" for the statistical process that generated a dataset, using both DAGs and the kinds of probability distributions we define below.

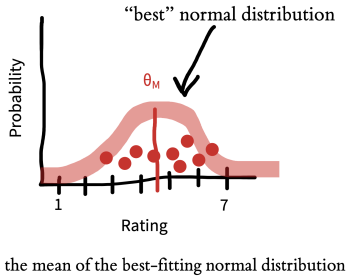


Figure 5.4: The best-fitting normal distribution for data from the milk-first condition.

Figure 5.5: Comparison of the best-fitting normal distribution and a substantially worse curve.

makes these data most likely is just 4.5 itself: the sample mean! That is why the sample mean in this case is the maximum likelihood estimate.

5.1.2 *Bayesian estimation*

The maximum likelihood estimation example above describes a common approach to estimating parameters, where the researcher completely puts aside their prior expectations about what these values might be. This approach is an example of a **frequentist** statistical approach, an approach that focuses on the long-run performance of estimation procedures.

Often this approach makes sense, especially when we have no prior expectations about the values we are estimating. But sometimes we *do* have relevant beliefs about the value. For example, before we perform our tea experiment, we don't know exactly what θ_{M} will be, but it seems a bit unlikely that tea would be consistently rated as either horrible (1) or perfect (7). We have what you might call *weak prior expectations* about the kinds of ratings we'll receive.

These kind of expectations are most useful when we have a very small amount of data. Remember that our goal is to estimate a population parameter using the sample data, and small data sets can be rather noisy. Taking into account our prior expectations can help to temper the influence of noise. For example, if our very first participant in the experiment rated their tea as terrible, we wouldn't want to jump to the conclusion that the tea was actually bad. Instead, we might speculate that the participant was having a bad day or just brushed their teeth. On the other hand, if all of our participants gave bad ratings to their tea, the data would be more persuasive; in that case, we might want to tell the cafe that they are serving substandard tea. The extent to which our prior expectations should moderate our conclusions should vary with the amount of sample data; with only a little data, our prior expectations should have more influence, but as we gather more, we should put greater weight on the data.

How do we quantify this tradeoff between our prior expectations and our current observations? We can do this via **Bayesian estimation** of $\hat{\theta}_{\text{M}}$. Bayesian estimation provides a principled framework for integrating prior beliefs and data. These estimation techniques can be very helpful in cases where data are sparse or prior beliefs are strong.

In Bayesian estimation, we observe some data d, consisting of the set of responses in the experiment. Now we can use **Bayes' rule**, a tool from

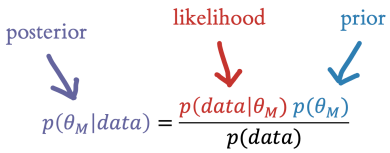


Figure 5.6: Bayes rule, annotated.

basic probability theory, to estimate this number (Figure 5.6). Each part of this equation has a name, and it's worth becoming familiar with them. The thing we want to compute, $p(\theta_M|\text{data})$, is called the **posterior probability** – it tell us what we should believe about the population parameter on tea quality, given the data we observed.[5]

The first part of the numerator is $p(\text{data}|\theta_M)$, the probability of the data we observed given our hypothesis about the participant's ability. This part is called the **likelihood**.[6] This term tells us about the relationship between our hypothesis and the data we observed – so if we think the tea is of high quality (say $\theta_M = 6.5$) then the probability of observing a bunch of low quality ratings will be fairly low.

The second term in the numerator, $p(\theta_M)$, is called the **prior**. This term encodes our beliefs about the likely distribution of tea quality. Intuitively, if we think that the tea is likely of high quality, we should require more evidence to convince us that it's bad. In contrast, if we think it's probably bad, a few examples of low ratings might serve to convince us.

Figure 5.7 gives an example of the combination of prior and data. In this example, we look at what difference the prior makes after observing 9 ratings. If we go in assuming that the tea is likely to be bad, the posterior mean (purple line) will be pushed downward relative to the maximum likelihood estimate (red line). This prior is operating only over on ratings – estimates of tea quality. Later on when we talk about comparing milk-first and tea-first ratings to get an estimate of the experimental effect, we could consider putting a prior on tea *discrimination* (e.g., the experimental effect).

Priors aren't usually as strong as the one shown above. Figure 5.8 shows how the picture shifts when we have a weaker prior reflecting a flatter, more widely spread belief about the distribution of ratings. Now the posterior mean (purple) is closer to the maximum likelihood mean (red). This situation is more common – the prior encodes a weak assumption that ratings won't cluster around the ends of the scale.

The effect of the prior is also decreased when you have more data. Take a look at Figure 5.9. The prior is the same as in Figure 5.7, but we have more data. As a result, the posterior distribution is much more peaked and also much closer to the data – the prior makes much less difference.

Bayesian estimation is most important when you have strong beliefs and not a lot of data. That can be a case where you have just a few participants in your experiment, but it's also good – and perhaps more common – to use Bayesian methods when you have a lot of data, but maybe

[5] We're making the posterior purple to indicate the combination of likelihood (red) and prior (blue).

[6] Speaking informally, "likelihood" is just a synonym for probability, but in Bayesian estimation, "likelihood" is a technical term specifically referring to probability of the data given our hypothesis. This ambiguity can get a bit confusing.

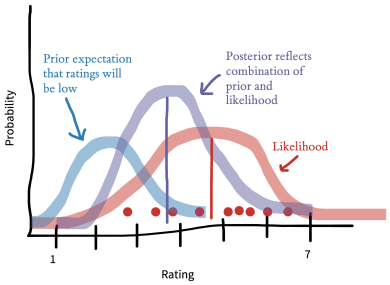


Figure 5.7: Bayesian inference about tea ratings with a strong prior on low values.

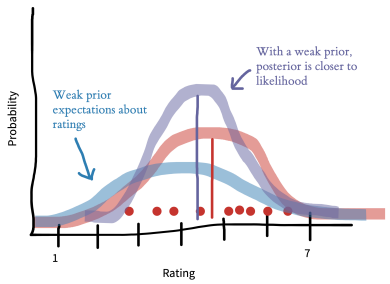


Figure 5.8: Bayesian inference about tea ratings with a weak prior on low values.

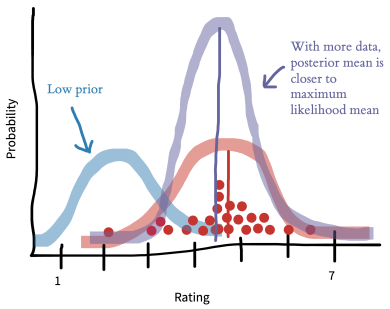


Figure 5.9: Bayesian inference about tea ratings with a a strong prior on low values and more data.

not that much data about particular units that you care about. For example, you might have a large dataset about the effects of an educational intervention but not that much data about how it affects a particular subgroup. Bayesian estimates and maximum likelihood estimates will exactly coincide either under a flat prior (a prior that makes any value equally likely) or as the amount of data goes to infinity.

5.2 Estimating and comparing effects

We've now covered estimating a single parameter (the mean for people who had milk-first tea) using both frequentist and Bayesian methods. But recall that what we really wanted to do was to estimate the *causal effect* we were interested in, namely the milk-first vs. tea-first effect. In this section, we'll discuss how to estimate the effect, and then how to use **effect size** measures to compare effects across experiments (as well as some of the pros and cons of doing so).[7]

5.2.1 Estimating the treatment effect

Let's refer to the causal effect we care about as our **treatment effect**.[8] In practice, estimating β (a parameter describing the treatment effect) is going to be a pretty straightforward extension to what we did before.

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.[9] 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 (i.e., 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'd probably expect, this estimate $\hat{\beta}$ turns out to be simply the difference in sample means, $\hat{\theta}_M - \hat{\theta}_T$. You can see this difference pictured in Figure 5.10.

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

[7] This method doesn't have to be used only with a causal effect, it can be any between-group difference. In the current example, we can say with certainty that this effect is a causal because our experiment uses random assignment.

[8] 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.

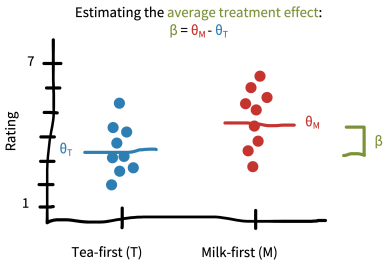


Figure 5.10: Estimating the average treatment effect from the tea-tasting data.

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

effect. If we have no prior beliefs at all, then we could specify a flat prior, $p(\beta) \propto 1$.[10]. If we believe the treatment effect is likely to favor milk-first pouring ($\beta > 0$), we could specify 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.[11]

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.

5.2.2 *Measures of effect size*

Once we have measured something, we need to make a decision about how to describe this effect to others. Sometimes we are working with fairly intuitive relationships that are easy to describe. A researcher might say, for example, that people who received milk-first tea drank the tea, on average, 5 minutes quicker than people who received tea-first tea (i.e., that $\hat{\beta} = 5$ minutes). Time is measured in units like minutes and seconds and so we all have a shared understanding of what 5 minutes means.

But what about our participants' ratings of tea quality, which were provided on an arbitrary 7-point rating scale that we devised? What does it mean to that participants who drank milk-first tea rated it 1 point higher than participants who drank tea-first tea (i.e., that $\hat{\beta} = 1$ point)? And how is this difference comparable to, for instance, a 1-point change on a scale that has similar anchors ("terrible" and "delicious") but uses a 100-point rating system?

To provide a common language for describing these relationships, some researchers use **standardized effect sizes**. A common standardized effect size is Cohen's d, which provides a standardized estimate of the difference between two means. There are many different ways to calculate Cohen's d (Lakens 2013), but all approaches are usually some variant of the following formula:

$$d = \frac{\theta_{\mathrm{M}} - \theta_{\mathrm{T}}}{\sigma_{\text{pooled}}}$$

[10] 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.

[11] 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.

where the difference between means (θ_T and θ_M) is divided by the pooled standard deviation σ_{pooled}. Intuitively, what you're doing is taking the study effect (β) and dividing it – scaling it – by the variation we saw between individuals in the study.

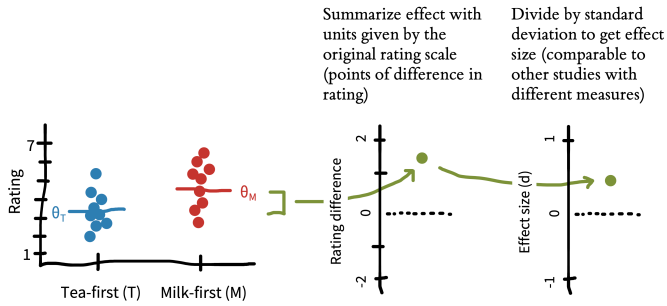


Figure 5.11: Schematic effect size computation.

Let's compute this measure for our tea-drinking study. We can just plug in the estimates we see in Figure 5.10 and compute the standard deviation of our observed data:

$$\hat{d} = \frac{\hat{\theta}_M - \hat{\theta}_T}{\hat{\sigma}_{pooled}} = \frac{4.5 - 3.5}{1.25} = \frac{1}{1.25} = 0.80$$

In other words, the effect size of the difference between the two conditions is .8 standard deviations. This process is shown graphically in Figure 5.11.[12]

We previously said that people who drank milk-first tea had quality ratings that were, on average, 1 point higher on a 7-point scale ($\beta = 1$ point). Cohen's d translates the arbitrary units of our rating scale into a **unit-less** effect size that is measured in terms of the variation in the data. You may find yourself wondering: "why would I ever describe things in terms of standard deviations?" The key benefit is that it allows us to compare the size of the effect between studies that use different measures.

Let's say that we ran a replication of our tea study with two changes: (1) we studied patrons in a US cafe instead of a UK cafe, and (2) we used a 100-point quality rating scale instead of a 7-point scale. Imagine that, just as we found that participants in the UK rated the milk-first tea 1-point higher on a 7-point quality scale, US participants rated the milk-first tea 1-point higher on a *100-point* quality scale. It seems clear that these effects are different because of the difference in scale. But how different?

It might at first seem reasonable just to normalize by the length of the scale. So maybe the UK experimental participants showed a 1/7 rating

[12] Cohen's d, also referred to as a **standardized mean difference** (SMD), can be tricky to apply to more complex experimental designs, such as when you have within-participant designs and multiple measurements of each participant. For some guidance on this topic, see Lakens (2013).

effect and the US participants showed a 1/100 rating effect. The trouble with this move is that it presupposes that participants from two different populations are using two different scales in exactly the same way! For example, maybe US participants made very clumpy judgments that were mostly centered around 50 (perhaps because of a lack of milk tea experience). Standardized effect sizes get around this kind of issue by scaling according to the variability of the data.

Let's compute the effect size for the cross-cultural replication. We'll imagine that participants who drank milk-first tea gave an average rating of 50/100 and participants who drank tea-first tea rated it 49 on average. But if their variability was also relatively lower, perhaps the standard deviation of their ratings was only 5. Using the formula above, we find

$$\hat{d}_{US} = \frac{\hat{\theta}_{\text{M}} - \hat{\theta}_{\text{T}}}{\hat{\sigma}_{\text{pooled}}} = \frac{50 - 49}{5} = \frac{1}{5} = 0.2$$

A Cohen's d of .2 means that US cafe patrons rated their tea .2 standard deviations higher when it was milk-first, much smaller than the .8 standard deviation difference in the UK patrons.

There are no hard and fast rules for interpreting what makes a big effect or a small effect, but people often refer back to a standard suggested by Cohen (1992). On those standards, $d = 0.8$ is a "large effect", and $d = 0.2$ is a "small effect." But these effect size interpretation norms are somewhat arbitrary. The key point here was that US and UK patrons had the same raw score change in quality ratings ($\hat{\beta} = 1$) and standardizing the differences allowed us to communicate that the difference was larger among the UK patrons.

Cohen's d is one of many standardized effect sizes that researchers can use. Just as Cohen's d standardizes differences in group means, there are also generalizations that allow for continuous treatment variables or covariate adjustment (e.g., Pearson's r, as we discuss below; r^2; or η^2). And there is a whole other set of effect-size measures for relationships between binary variables (e.g., odds ratio). We'll be using effect sizes throughout the book, but we'll be using Cohen's d as our example.[13]

[13] If you'd like to learn more about other varieties of effect size, take a look at Fritz, Morris, and Richler (2012) and Lakens (2013).

5.2.3 *Pros and cons of standardizing effect sizes*

Standardizing effect size helps communicate that a 1-point change on a 7-point scale is not the same as a 1-point change on a 100-point scale.

But is it any better to say that the first change represents a 0.80 standard deviation difference and the second a 0.08 standard deviation difference?

Effect sizes allow us to compare results across studies more easily. Across studies, researchers use different measures, different study designs, and different populations. Standardization gives us a "common language" to describe estimated relationships in these varied contexts. This language is helpful when we want to aggregate and compare effects across studies via meta-analysis. And it is also helpful when planning new studies. When trying to figure out how many participants to run in a study, almost all techniques for sample size planning use standardized effect sizes to determine how much data would be needed to reliably detect an effect.

Standardizing effect sizes has limitations, though. For example, if two interventions produce the same absolute change in the same outcome measure, but are studied in different populations in which the variability on the outcome differs substantially, the interventions would produce different standardized mean differences (Baguley 2009) (see the Depth box "Reliability paradoxes!" in Chapter 8).

Imagine we conducted our tea experiment again, but this time with (decaf) tea, and focusing on children. Maybe milk-first tea tastes the same amount better than tea-first tea for kids and for adults. But kids are, as a rule, more variable in their responding than adults. This higher level of variability would lead us to observe a smaller effect size in kids vs. adults. Recall that our UK adult SD was 1.25, and our effect size was $d = .8$. Imagine that children's SD is 2.5. In this scenario, even if tea led to the same 1-point absolute change in ratings among adults and children, the standardized effect size for kids would look half as big:

$$\hat{d}_{kids} = \frac{\hat{\theta}_{\text{M}} - \hat{\theta}_{\text{T}}}{\hat{\sigma}_{\text{pooled}}} = \frac{5-4}{2.5} = \frac{1}{2.5} = .4$$

This example highlights some of the challenges with standardization. If we focused on the fact that both adults and children show a 1-point change in ratings levels ($\hat{\beta} = 1$), we would conclude that milk-first tea ordering is as much better for adults as kids. If we focused on the standardized effect sizes, however, we would conclude that the milk ordering effect is twice as big for adults.

So which is better: describing raw measures or standardized effect sizes? In general our response is "Why not both?" But if you wanted to pick

one or the other, we recommend considering what type of measurement you are using. With measures that yield common measurement units that are likely to be reported in many studies already, use raw scores (Baguley 2009). For example, if your study uses physical units such as milliseconds (e.g., for reaction times) or counts (e.g., for a study tracking an outcome like number of words), these measurements can be quite useful to compare across studies. Reporting raw measurements also can allow you to check whether your measurements make sense – for example, a reaction time of 70 milliseconds is inhumanly fast, while a reaction time of 10 seconds might be extremely slow (at least, for many speeded tasks).

In contrast, we recommend using standardized effect sizes for cases where the measurement is relatively unlikely to be comparable with other studies in its original form, or unlikely to be meaningful on its own. For example, reporting the effect of an intervention on raw math test scores is only meaningful if the reader knows how many items are on the test, how difficult it is, and so forth. In such a case where there it is hard for a reader to be "calibrated" to the specific measurement units you are using, standardized effect sizes may be the best way to report your finding (Kelley and Preacher 2012).

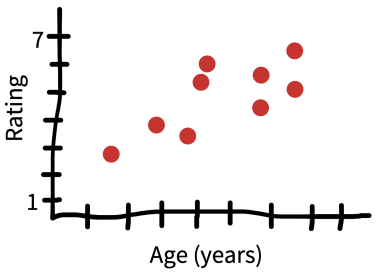


Figure 5.12: The relationship between age and milk-first tea rating.

5.3 Estimating the relationship between variables

Our focus up until now has been in estimating individual effects, but sometimes we also want to estimate the relationship between two different variables. Extending our example, Figure 5.12 shows the relationship between the age of the tea taster and their rating of milk-first tea. It seems that younger people overall like tea less than older people.[14] How could we quantify this result?

The first concept we need is **covariance**. Covariance captures the degree to which we expect two variables to deviate from their means in the same direction. We're looking at milk-first tea ratings M and participant ages A. We can write the covariance between these two variables as

$$Cov(M, A) = E[(M - \theta_M)(A - \theta_A)]$$

This formula expresses the expected product of how much each observation differs from its expectation (mean) along each variable. Figure 5.13 shows these differences, which are multiplied together for each point to get the covariation.[15]

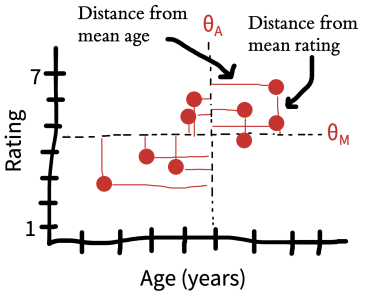


Figure 5.13: Estimating the covariation between age and milk-first tea rating.

[14] Remember, this is a correlational relationship, and there's no causal inference possible here.

[15] This looks a little tricky, but it's actually very related to the basic concepts we've already seen. Remember when we introduced the standard deviation, we described it as the expected distance between new samples from a distribution and the mean of that distribution. The covariance is very related: the standard deviation is just $\sqrt{Cov(X, X)}$, in other words, the square root of the covariance of a variable with itself.

This covariance number gives us an estimate of how much age and ratings covary, but its units are a bit funny: it's hard to know what to make of an expected deviation of 1 point-year. We can do a simple trick to standardize its units and make it into a wonderful form of effect size called the **correlation coefficient** (denoted r). Remember that to create effect sizes above, we divided by the standard deviation of the variable. Here all we have to do is divide by the standard deviation of both variables.

$$r_{M,A} = \frac{Cov(M, A)}{\sigma_M \sigma_A}$$

In other words, the correlation between two variables is the standardized covariation.

The correlation coefficient is the most ubiquitous measure of association between variables. It ranges between -1, where two variables covary in exactly the opposite direction, to 1, when two variables covary perfectly. A correlation means that there is no association between two variables. A correlation of -1 or 1 doesn't mean that these two variables have the same scale, however: it just means that they "move together."

Critically, a correlation is an effect size. Correlations can be compared across different measures and different studies (including both experimental and observational studies), making it a very valuable scale-free comparison tool.

This section has described one way of looking at a correlation coefficient: as standardized covariation. For a great discussion of all the different ways of thinking about correlations, see Lee Rodgers and Nicewander (1988).

5.4 Chapter summary: Estimation

In this chapter, we introduced the idea of estimating both individual measurements and treatment effects from observed data. These ideas are simple but they lay the foundations for hypothesis testing and modeling (our next two chapters). Further, we set up the distinction between Bayesian and frequentist approaches, which we will expand in the next chapter since these traditions provide different inferential tools.

Discussion questions

1. In this chapter you learned about estimation, and in this book more generally, we have argued that the goal of an experiment is to provide a maximally precise estimate of a causal effect. Psychology as a field has often been criticized for focusing too much on inference and too little on estimation. Find an article in the journal Psychological Science that reports on an experiment or series of experiments and read the abstract. Does it mention an estimate of any particular quantity? What might be the benefits of reporting estimates in the study

abstract?

2. Try the same exercise with a paper in the *New England Journal of Medicine* or *Journal of the American Medical Association*. Find a paper and check if there is a mention of any specific quantity being estimated. (We suspect there will be!) Consider this contrast between the medical article and the psychology article. What do you make of this difference between fields?

Readings

- A great narrative introduction to the history and practice of statistics: Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Macmillan.
- An open source statistics textbook that follows a similar approach as Chapters 5 – 7: Poldrack, R. (2022). *Statistical thinking for the 21st century*. Available free online at https://statsthinking21.org.

References

Baguley, Thom. 2009. "Standardized or Simple Effect Size: What Should Be Reported?" *British Journal of Psychology* 100 (3): 603–17.

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112 (1): 155.

Dunn, Peter M. 1997. "James Lind (1716-94) of Edinburgh and the Treatment of Scurvy." *Archives of Disease in Childhood-Fetal and Neonatal Edition* 76 (1): F64–65.

Fritz, Catherine O, Peter E Morris, and Jennifer J Richler. 2012. "Effect Size Estimates: Current Use, Calculations, and Interpretation." *Journal of Experimental Psychology: General* 141 (1): 2.

Kelley, Ken, and Kristopher J Preacher. 2012. "On Effect Size." *Psychological Methods* 17 (2): 137.

Kennedy, Maeve. 2003. "How to Make a Perfect Cuppa: Put Milk in First." *How to Make a Perfect Cuppa: Put Milk in First*. https://www.theguardian.com/uk/2003/jun/25/science.highereducation.

Lakens, Daniël. 2013. "Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs." *Frontiers in Psychology* 4: 863.

Lee Rodgers, Joseph, and W Alan Nicewander. 1988. "Thirteen Ways to Look at the Correlation Coefficient." *The American Statistician* 42 (1): 59–66.

Lundberg, Ian, Rebecca Johnson, and Brandon M Stewart. 2021. "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86 (3): 532–65.

Peirce, Charles Sanders, and Joseph Jastrow. 1884. "On Small Differences in Sensation." *Memoirs of the National Academy of Sciences* 3.

6 Inference

Learning goals

- Discuss the purpose of statistical inference
- Define *p*-values and Bayes Factors
- Consider common fallacies about inference (especially for *p*-values)
- Reason about sampling variability
- Define and reason about confidence intervals

We've been arguing that experiments are about measuring effects. The effects we are interested in are causal effects for a group of people, but that group is almost always bigger than the participants in an experiment. **Statistical inference** is the process of going beyond the specific characteristics of the sample that you measured to make generalizations about the broader **population**.

Chapter 5 already showed us how to make one simple inference: estimating population parameters using both frequentist and Bayesian techniques. Estimating population parameters is an important first step. But often we want to make more sophisticated inferences so that we can answer questions such as:

1. How likely is it that this pattern of measurements was produced by chance variation?
2. Do these data provide more support for one hypothesis or another?
3. How precise is our estimate of an effect?
4. What portion of the variation in the data is due to a particular manipulation (as opposed to variation between participants, stimulus items, or other manipulations)?

Question (1) is associated with one particular type of statistical inference method – **null hypothesis significance testing** (NHST) in the **frequentist** statistical tradition. NHST has become synonymous with data analysis, such that in the vast majority of research papers (and research methods courses), all of the reported analyses are tests of this type. Yet this equivalence is quite problematic.

The move to "go test for significance" before visualizing your data and trying to understand sources of variation (participants, items, manipulations, etc.) is one of the most unhelpful strategies for an experimenter. Whether $p < .05$ or not, a test of this sort gives you literally *one bit* of information about your data.[1] Considering effect sizes and their variation more holistically, including using the kinds of visualizations we advocate in Chapter 15, gives you a much richer sense of what happened in your experiment!

In this chapter, we will describe NHST, the conventional method that many students still learn (and many scientists still use) as their primary method for engaging with data. All practicing experimentalists need to understand NHST, both to read the literature and also to apply this method in appropriate situations. For example, NHST may be a reasonable tool for testing whether an intervention leads to a difference between a treatment condition and an appropriate control. But we will also try to contextualize NHST as a very special case of a broader set of statistical inference strategies. Further, we will continue to flesh out our account of how some of the pathologies of NHST have been a driver of the replication crisis.

If NHST approaches have so many issues, what should replace them? Figure 6.1 shows one way of organizing different inferential approaches. There has been a recent move towards the use of Bayes Factors to quantify the evidence in support of different candidate hypotheses. Bayes Factors can help answer questions like (2). We introduce these tools, and believe that they have broader applicability than the NHST framework and should be known by students. On the other hand, Bayes Factors are not a panacea. They have many of the same problems as NHST when they are applied dichotomously.

Instead of dichotomous frequentist or Bayesian hypothesis testing, we follow our thematic emphasis on MEASUREMENT PRECISION and advocate for a **measurement** strategy, which is more suited towards questions (3) and (4) (Cumming 2014; Kruschke and Liddell 2018). The goal of these strategies is to yield an accurate and precise estimate of the relationships underlying observed variation in the data.

This isn't a statistics book and we won't attempt to teach the full array of important statistical concepts that will allow students to build good models of a broad array of datasets. (Sorry!).[2] But we do want you to be able to reason about inference and modeling. In this chapter, we'll start by making some inferences about our tea-tasting example from the last chapter, using this example to build up intuitions about hypothesis testing and inference. Then in Chapter 7, we'll start to look at more sophisticated models and how they can be fit to real datasets.

[1] In the information theoretic sense, as well as the common sense!

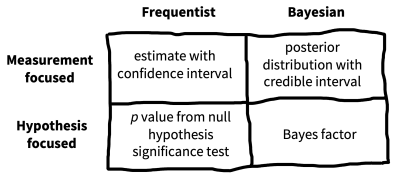


Figure 6.1: Clarifying the distinctions between Bayesian and Frequentist paradigms and the tools they offer for measurement and hypothesis testing. For many settings, we think the measurement mindset is more useful. Adapted from Kruschke and Liddell (2018).

[2] If you're interested in going deeper, here are two books that have been really influential for us. The first is Gelman and Hill (2006) and its successor Gelman, Hill, and Vehtari (2020), which teach regression and multi-level modeling from the perspective of data description. The second is McElreath (2018), a course on building Bayesian models of the causal structure of your data. Honestly, neither is an easy book to sit down and read (unless you are the kind of person who reads statistics books on the subway for fun) but both really reward detailed study. We encourage you to get together a reading group and go through the exercises in one of these together. It'll be well worth while in its impact on your statistical and scientific thinking.

6.1 Sampling variation

In Chapter 5, we introduced Fisher's tea-tasting experiment and discussed how to estimate means and differences in means from our observed data. These so-called "point estimates" represent our best guesses about the population parameters given the data – and possibly also given our prior beliefs. We can also report how much statistical uncertainty is involved in these point estimates.[3] Quantifying and reasoning about this uncertainty is an important goal: in our original study we only had 9 participants in each group, which will only provide a low precision (i.e., highly uncertain) estimate of the population. By contrast, if we repeated the experiment with 200 participants in each group, the data would be far less noisy, and we would have much less uncertainty, even if the point estimates happened to be identical.

[3] As in the previous chapter, we're only capturing *statistical* uncertainty. A holistic view of a particular estimate's credibility also include everything else you know about the study design.

6.1.1 Standard errors

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.

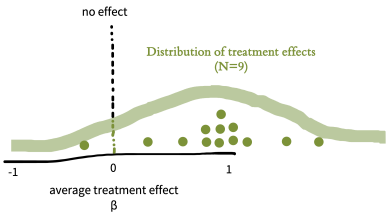


Figure 6.2: Sampling distribution for the treatment effect in the tea-tasting experiment, given many different repetitions of the same experiment, each with N=9 per group. Circles represent average treatment effects from different individual experiments, while the thick line represents the form of the underlying distribution.

Code

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)
```

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.

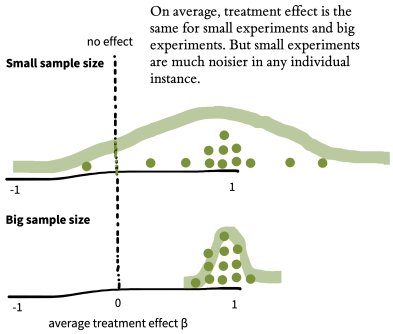


Figure 6.3: Comparing sampling distributions for the treatment effect with smaller and larger size samples.

Code

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 1000 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 `summarise()` workflow from Appendix D to get the estimated treatment effect for each of these simulations.

```
tea_summary <- samps |>
  group_by(sim, condition) |>
  summarise(mean_rating = mean(rating)) |>
  group_by(sim) |>
  summarise(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.

6.1.2 The central limit theorem

We talked in the last chapter about the normal distribution, a convenient and ubiquitous tool for quantifying the distribution of measurements. A shocking thing about sampling distributions for many kinds of estimates – and for *all* maximum likelihood estimates – is that they become normally distributed as the sample size gets larger and larger. This result holds even for estimates that are not even remotely normally distributed in small samples!

For example, say we are flipping a coin and we want to estimate the probability that it lands heads (p_H). If we draw samples each consisting of only $n = 2$ coin flips, Figure 6.4 is the sampling distribution of the estimates ($\hat{p}_H$). This sampling distribution doesn't look normally distributed at all – it doesn't have the characteristic "bell curve" shape! In a sample of only two coin flips, $\hat{p}_H$ can only take on the values 0, 0.5, or 1.

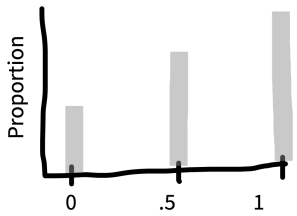


Figure 6.4: Samping distribution of samples from a biased coin (N=2 flips per sample). Bar height is the proportion of flips resulting in a particular mean.

But look what happens as we draw increasingly larger samples in Figure 6.5: We get a normal distribution! This tendency of sampling distributions to become normal as n becomes very large reflects a deep and elegant mathematical law called the **Central Limit Theorem**.

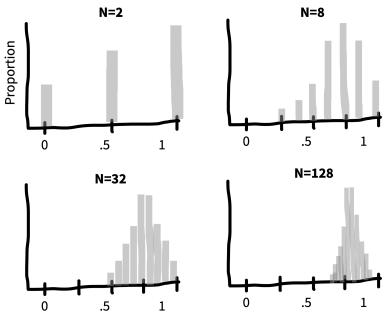


Figure 6.5: Sampling distribution for 2, 8, 32, and 128 flips.

The practical upshot is that the Central Limit Theorem directly helps us characterize the uncertainty of sample estimates. For example, when the sample size is reasonably large (approximately $n > 30$ in the case of sample means) the standard error (i.e., the standard deviation of the sampling distribution) of a sample mean is approximately $\widehat{SE} = \sigma/\sqrt{n}$. The sampling distribution becomes narrower as the sample size increases because we are dividing by the square root of the number of observations.

Code

Even though our figures are hand-drawn, they're based on real simulations. For our central limit theorem simulations, we again use the `map()` function. We set up a tibble with the different values we want to to try (which we call `n_flips`). Then we make use of the `map()` function to run `rbinom()` (random binomial samples) for each value of `n_flips`.

One trick we make use of here is that `rbinom()` takes an extra argument that says how many of these random values you want to generate. Here we generate `nsamps = 1000` samples, giving us 1000 independent replicates at each `n`. But returning an array of 1000 values for a single value of `n_flips` results in something odd: the value for each element of `flips` is an array. To deal with that, we use the `unnest()` function, which expands the array back into a normal tibble.

```
n_samps <- 1000
n_flips_list <- c(2, 8, 32, 128)

sample_p <- tibble(n_flips = n_flips_list) |>
  mutate(flips = map(n_flips, \(f) rbinom(n = n_samps, size = f, prob = .7))) |>
  unnest(cols = flips) |>
  mutate(p = flips / n_flips)
```

6.2 From variation to inference

Let's go back to Fisher's tea-tasting experiment. The first innovation of that experiment was the use of randomization to recover an estimate of the causal effect of milk ordering. But there was more to Fisher's analysis than we described.

The second innovation of the tea-tasting experiment was the idea of creating a model of what might happen during the experiment. Specifically, Fisher described a hypothetical **null model** that would arise if the lady had chosen cups by chance rather than because of some tea sensitivity. In our tea-rating experiment, the null model describes what happens when there is no difference in ratings between tea-first and milk-first cups. Under the null model, the true treatment effect (β) is zero.

Even with an actual treatment effect of zero, across repeated sampling, we should see some variation in $\hat{\beta}$, our *estimate* of the treatment effect. Sometimes we'll get a small positive effect, sometimes a small negative one. Occasionally just by chance we'll get a big effect. This is just sampling variation as we described above.

Fisher's innovation was to quantify the probability of observing various values of $\hat{\beta}$, given the null model. Then, if the observed data that were very low probability under the null model, we could declare that the null was rejected. How unlikely must the observed data be, in order to reject the null? Fisher declared that it is "usual and convenient for experimenters to take 5 percent as a standard level of convenience," establishing the .05 cutoff that has become gospel throughout the sciences.[4]

Let's take a look at what the null model might look like. We already tried out repeating our tea-tasting experiment thousands of times in our discussion of sampling above. Now in Figure 6.6, we do the same thing but we assume that the **null hypothesis** of no treatment effect is true. The plot shows the distribution of treatment effects $\hat{\beta}$ we observe: some a little negative, some a little positive, and a few substantially positive or negative, but mostly zero.

Let's apply the $p < .05$ standard. If our observation has less than a 5% probability under the null model, then the null model is likely wrong. The red dashed lines on Figure 6.6 show the point below which only 2.5% of the data are found and the point above which only 2.5% of the data are found. These are called the **tails** of the distribution. Because we'd be equally willing to accept milk-first tea or tea-first tea being better, we consider both positive and negative observations as possible.[5]

[4] Actually, right after establishing .05 as a cutoff, Fisher then writes that "in the statistical sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon… in order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result." In other words, Fisher was all for replication!

[5] Because we're looking at both tails of the distribution, this is called a "two-tailed" test.

Code

To simulate our null model, we can do the same kind of thing we did before, just specifying to our `make_tea_data()` function that the true difference in effects is zero!

```
n_sims <- 1000
null_model <- tibble(sim = 1:n_sims, n = 18) |>
  mutate(data = map(sim, \(i) make_tea_data(n_total = n, delta = 0))) |>
  unnest(cols = data)
```

Again we use `group_by()` and `summarise()` to get the distribution of treatment effects under the null hypothesis.

```
null_model_summary <- null_model |>
  group_by(sim, condition) |>
  summarise(mean_rating = mean(rating)) |>
  group_by(sim) |>
  summarise(delta = mean_rating[condition == "milk first"] -
              mean_rating[condition == "tea first"])
```

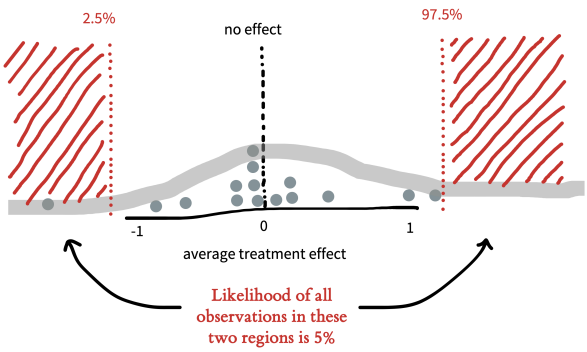


Figure 6.6: One example of the distribution of treatment effects under the null model (with N=9 per group). The red regions indicate the part of the distribution in which less than 5% of observations should fall.

Figure 6.6 captures the logic of NHST: if the observed data fall in the region that has a probability of less than .05 under the null model, then we reject the null. So then when we observe some particular treatment effect $\hat{\beta}$ in a single (real) instance of our experiment, we can compute the probability of these data or any data more extreme than ours under the null model.[6] This probability is our p-value, and if it is small, it gives us license to conclude that the null is false.

As we saw before, the larger the sample size, the smaller the standard error. That's true for the null model too! Figure 6.7 shows the expected null distribution for a bigger experiment.

[6] The "more extreme" part deserves a little explanation. Any individual outcome is relatively unlikely by itself, just because it's surprising that the estimate is that exact value (we're simplifying here, it gets a bit trickier when you are talking about real numbers). What we care about instead is a *group* of values. The ones that are in the middle of the distribution are, considered as a group, quite likely; the ones on the tails are, as a group, less likely. We want to know if the probability of the group of datapoints that includes our observation and anything even further out on the tails is collectively less than .05.

The more participants in the experiment, the tighter the null distribution becomes, and hence the smaller the region in which we should expect a null treatment effect to fall. Because our expectation based on the null becomes more precise, we will be able to reject the null based on smaller treatment effects. In this type of hypothesis testing, as with estimation, our goals matter. If we're merely testing a hypothesis out of curiosity, perhaps we don't want to measure too many cups of tea. But if we were designing the tea strategy for a major cafe chain, the stakes would be higher; in that case, maybe we'd want to do a more extensive experiment!

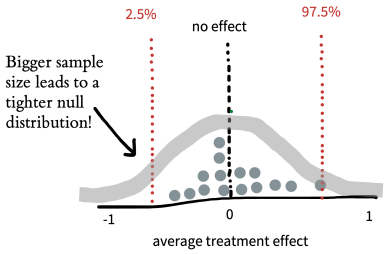


Figure 6.7: Example distribution of treatment effects under the null model for a larger experiment.

Code

We can do a more systematic simulation of the null regions for different sample sizes by simply adding a parameter to our simulation.

```
n_sims <- 10000
null_model_multi_n <- expand_grid(sim = 1:n_sims, n = c(12, 24, 48, 96)) |>
  mutate(sim_data = map(n, \(n_i) make_tea_data(n_total = n_i, delta = 0))) |>
  unnest(cols = sim_data)

null_model_summary_multi_n <- null_model_multi_n |>
  group_by(n, sim, condition) |>
  summarise(mean_rating = mean(rating)) |>
  group_by(n, sim) |>
  summarise(delta = mean_rating[condition == "milk first"] -
              mean_rating[condition == "tea first"])

null_model_quantiles_multi_n <- null_model_summary_multi_n |>
  group_by(n) |>
  summarise(q_025 = quantile(delta, .025),
            q_975 = quantile(delta, .975))
```

Here is the plotting code to produce a comparable figure to our illustration:

```
ggplot(null_model_summary_multi_n, aes(x = delta)) +
  facet_wrap(vars(n), nrow = 1, labeller = label_both) +
  geom_histogram(binwidth = .25) +
  geom_vline(xintercept = 0, color = pal$grey, linetype = "dotted") +
  geom_vline(data = null_model_quantiles_multi_n,
             aes(xintercept = q_025), color = pal$red, linetype = "dotted") +
  geom_vline(data = null_model_quantiles_multi_n,
             aes(xintercept = q_975), color = pal$red, linetype = "dotted") +
  xlim(-2.5, 2.5) +
  labs(x = "Difference in rating", y = "Frequency")
```

One last note: You might notice an interesting parallel between the NHST paradigm and Popper's falsificationist philosophy (introduced in Chapter 2). In both cases, you never get to *accept* the actual hypothesis of interest. The only thing you can do is observe evidence that is inconsistent with the null hypothesis. The added limitation of NHST is that the only hypothesis you can falsify is the null![7]

6.3 Making inferences

In the tea-tasting example we were just considering, we were trying to make an inference from our sample to the broader population. In particular, we were trying to test whether milk-first tea was rated as better than tea-first tea. Our inferential goal was a clear, binary answer: is milk-first tea better?

By defining a p-value, we got one procedure for giving this answer. If $p < .05$, we reject the null. Then we can look at the direction of the difference and, if it's positive, declare that milk-first tea is "significantly" better. Let's compare this procedure to a different process that builds on the Bayesian estimation ideas we described in the previous chapter. We can then come back to examine NHST in light of that framework.

6.3.1 Bayes Factors

Bayes Factors are a method for quantifying the support for one hypothesis over another, based on an observed dataset. They don't tell you the probability that a particular hypothesis is right, but they let you compare two different ones.

Informally, we've now discussed two different distinct hypotheses about the tea situation: our participants could have *no* tea discrimination ability – leading to chance performance. We call this H_0. Or they could have some non-zero ability – leading to greater than chance performance. We call this H_1. The Bayes Factor is simply the likelihood of the data (in the technical sense used above) under H_1 vs. under H_0 (Figure 6.8). The Bayes Factor is a ratio, so if it is greater than 1, the data are more likely under H_1 than they are under H_0 – and vice versa for values between 1 and 0. A BF of 3 means there is three times as much evidence for H_1 than H_0, or equivalently 1/3 as much evidence for H_0 as H_1.[8]

[7] A historical note: what we describe here as NHST is not what either Fisher's method *or* the Neyman-Pearson method that we introduce below. It's what Gigerenzer (1989) called "the silent hybrid solution," in which the more continuous approach to p-values that Fisher advocated for got rolled into the hypothesis testing approach of Neyman and Pearson. This hybrid – which neither Fisher nor Neyman and Pearson would have liked – is what we now mostly take for granted as the received NHST approach.

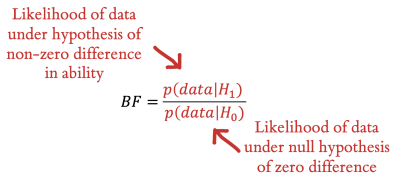


Figure 6.8: The Bayes Factor (BF).

[8] Sometimes people refer to the BF in favor of H_1 as the BF_{10} and the BF in favor of H_0 as the BF_{01}. This notation is a bit confusing because the first of these looks like the number 10.

Code

Bayes Factors are delightfully easy to compute using the `BayesFactor` R package. All we do is feed in the two sets of ratings to the `ttestBF()` function!

```
library(BayesFactor)

tea_bf <- ttestBF(x = filter(tea_data, condition == "milk first")$rating,
                  y = filter(tea_data, condition == "tea first")$rating,
                  paired = FALSE)
```

There are a couple of things to notice about the Bayes Factor. The first is that, like a p-value, it is inherently a continuous measure. You can artificially dichotomize decisions based on the Bayes Factor by declaring a cutoff (say, BF > 3 or BF > 10), but there is no intrinsic threshold at which you would say the evidence is "significant." Some guidelines for interpretation (from S. N. Goodman 1999) are shown in Table 6.1.[9] On the other hand, cutoffs like BF > 5 or $p < .05$ are not very informative. So although we provide this table to guide interpretation, we caution that you should always report and interpret the actual Bayes Factor, not whether it is above or below some cutoff.

[9] Some like the guidelines provided by Jeffreys (1961), which include categories such as "barely worth mentioning" (1 > BF > 3).

Table 6.1: S. N. Goodman (1999) interpretation guidelines for Bayes Factors.

BF range	Interpretation
< 1	Negative (supports H_0)
1–5	Weak
5–10	Moderate
10–20	Moderate to strong
20–100	Strong to very strong

The second thing to notice about the Bayes Factor is that it doesn't depend on our prior probability of H_1 vs. H_0. We might think of H_1 as very implausible. But the BF is independent of that prior belief. So that means it's a measure of how much the evidence should shift our beliefs away from our prior. One nice way to think about this is that the Bayes Factor computes how much our beliefs – whatever they are – should be changed by the data (Morey and Rouder 2011).

In practice, the thing that is both tricky and good about Bayes Factors is that you need to define an actual model of what H_0 and H_1 are. That process involves making some assumptions explicit. We won't go into how to make these models here – this is a big topic that is covered extensively in books on Bayesian data analysis.[10] The goal here is just to give a general sense of what Bayes Factors are.

[10] Two good ones beyond the McElreath book mentioned above are Gelman et al. (1995), which is a bit more statistical, and Kruschke (2014), which is a bit more focused on psychological data analysis. An in-prep web-book by Nicenboim et al. (https://vasishth.github.io/bayescogsci/book/) also looks great.

6.3.2 p-values

Now let's turn back to NHST and the p-value. We already have a working definition of what a p-value is from our discussion above: it's the **probability of the data (or any data that would be more extreme) under the null hypothesis**. How is this quantity related to either our Bayesian

estimate or the BF? Well, the first thing to notice is that the p-value is very close (but not identical) to the likelihood itself.[11]

Next we can use a simple statistical test, a t-test, to compute p-values for our experiment. In case you haven't encountered one, a t-test is a procedure for computing a p-value by comparing the distribution of two variables using the null hypothesis that there is no difference between them.[12] The t-test uses the data to compute a **test statistic** whose distribution under the null hypothesis is known. Then the value of this statistic can be converted to p-values for making an inference.

[11] The likelihood – for both Bayesians and frequentists – is the probability of the data, just like the p-value. But unlike the p-value, it doesn't include the probability of more extreme data as well.

[12] t-tests can also be used in cases where one sample is being compared to some baseline.

Code

The standard `t.test()` function is built into R via the default `stats` package. Here we simply make sure to specify the variety of test we want by using the flags `paired = FALSE` and `var.equal = TRUE` (denoting the assumption of equal variances).

```
tea_t <- t.test(x = filter(tea_data, condition == "milk first")$rating,
                y = filter(tea_data, condition == "tea first")$rating,
                paired = FALSE, var.equal = TRUE)
```

Imagine we conduct a tea-tasting experiment with $N = 48$ and perform a t-test on our experimental results. In this case, we see that the difference between the two groups is significant at $p < .05$: $t(46) = 2.86$, $p = .006$.

The expression $t(46) = 2.86$, $p = .006$ is the standard way to report of a t-test according to the American Psychological Association. The first part of this report gives the t value, qualified by the **degrees of freedom** for the test in parentheses. We won't focus much on the idea of degrees of freedom here, but for now it's enough to know that this number quantifies the amount of information given by the data, in this case 48 datapoints minus the two means (one for each of the samples).

Let's compare p values and Bayes Factors (computed using the default setup in the `BayesFactor` R package). In Table 6.2), the rows represent simulated experiments with varying total numbers of participants (N and varying average treatment effects. Both p and BF go up with more participants and larger effects. In general, BFs tend to be a bit more conservative than p-values, such that $p < .05$ can sometimes translate to a BF of less than 3 (Benjamin et al. 2018). For example, take a look at the row with 48 participants and an effect size of 1: the p value is less than .05, but the Bayes Factor is only 2.0.

Table 6.2: Comparison of p-value and BF for several different (randomly-generated) tea-tasting scenarios.

N	Effect size	p-value	BF
12	0.5	> .999	0.5
12	1.0	.076	1.4
12	1.5	.002	18.7
24	0.5	.858	0.4
24	1.0	.061	1.5
24	1.5	.009	5.6
48	0.5	.002	17.7
48	1.0	.033	2.0
48	1.5	< .001	133.6
96	0.5	.038	1.5
96	1.0	< .001	12218.2
96	1.5	< .001	3081.4

The critical thing about p-values, though, is not just that they are a kind

of data likelihoods. It is that they are used in a *specific inferential procedure*. The logic of NHST is that we make a binary decision about the presence of an effect. If $p < .05$, the null hypothesis is rejected; otherwise not. As Fisher (1949) wrote,

> It should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. (p. 19)

The main problem with p-values from a scientific perspective is that researchers are usually interested in not just rejecting the null hypothesis but also in the evidence for the alternative (the one we are interested in). The Bayes Factor is one approach to quantifying positive evidence *for* the alternative hypothesis in a Bayesian framework. This issue with the Fisher approach to p-values has been known for a long time, though, and so there is an alternative frequentist approach as well.

6.3.3 The Neyman-Pearson approach

One way to "patch" NHST is to introduce a decision-theoretic view, shown in Figure 6.9.[13] On this view, called the Neyman-Pearson view, there is a real H_1, albeit one that is not specified. Then the true state of the world could be that H_0 is true or H_1 is true. The $p < .05$ criterion is the threshold at which we are willing to reject the null, and so this constitutes our **false positive** rate α. But we also need to define a **false negative** rate, which is conventionally called β.[14]

Setting these rates is a decision problem: If you are too conservative in your criteria for the intervention having an effect, then you risk a false negative, where you incorrectly conclude that it doesn't work. And if you're too liberal in your assessment of the evidence, then you risk a false positive.[15] In practice, however, people usually leave α at .05 and try to control the false negative rate by increasing their sample size.

As we saw in Figure 6.6, the larger the sample, the better your chance of rejecting the null for any given non-null effect. But these chances will depend also on the effect size you are estimating. This formulation gives rise to the idea of classical power analysis, which we cover in Chapter 10. Most folks who defend binary inference are interested in using the Neyman-Pearson approach. In our view, this approach has its place (it's especially useful for power analysis) but it still suffers from the substantial issues that plague all binary inference techniques.

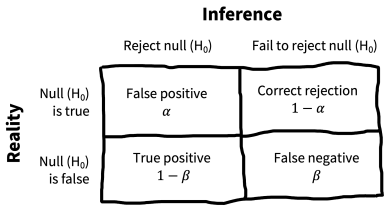


Figure 6.9: Standard decision matrix for the Neyman-Pearson approach to statistical inference.

[13] A little bit of useful history here is given in Cohen (1990), and we also recommend Gigerenzer (1989) for a broader perspective.

[14] Unfortunately, β is very commonly used for regression coefficients – and for that reason we've used it as our symbol for causal effects. We'll be using these βs in the next chapter as well. Those βs are not to be confused with false negative rates. Sorry, this is just a place where statisticians have used the same Greek letter for two different things.

[15] To make really rational decisions, you could couple this chart to some kind of utility function that assessed the costs of different outcomes. For example, you might think it's worse to proceed with an intervention that doesn't work than to stay with business as usual. In that case, you'd assign a higher cost to a false positive and accordingly try to adopt a more conservative criterion. We won't cover this kind of decision analysis here, but Pratt et al. (1995) is a classic textbook on statistical decision theory if you're interested.

Depth

Nonparametric resampling under the null

Hypothesis testing requires knowing the null distribution. In the examples above, it was easy to use statistical theory to work out the null distribution using knowledge of the binomial or normal distribution. But sometimes we don't know what the null distribution would look like. What if the ratings data from our tea-tasting experiment was very skewed, such that there were many low ratings and a few very high ratings (as in Figure 6.10)?

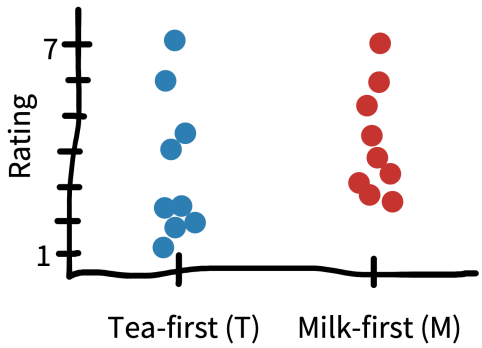


Figure 6.10: A small tea-tasting experiment with a skewed distribution of ratings.

With skewed data like this, we couldn't proceed with a t-test in good conscience because, with only $n = 18$, we can't necessarily trust that the Central Limit Theorem has "kicked in" sufficiently for the test to work despite the skewness. Put another way, we can't be sure that the null distribution is normal (Gaussian) in this case.

An alternative way to approximate a null distribution is through nonparametric resampling. **Resampling** means that we're going to draw new samples *from our existing sample*, and **nonparametric** means that we will do this in a way that obviates assumptions about the shape of the null distribution – in contrast to **parametric** approaches that do rely on such assumptions). These techniques are sometimes called "bootstrapping" techniques.

The idea is, if the treatment truly had no effect on the outcome, then the observations would be **exchangeable** between the treatment and control groups. That is, there would not be systematic differences between the treatment and control groups. This property may or may not be true in our observed sample (after all, that's why we're doing a hypothesis test in the first place), but we can draw new samples from our existing sample in a manner that forces exchangability.

To perform this kind of test with our tea-tasting data, we would randomly shuffle the ratings in our dataset while leaving the condition assignments fixed. If we did this thousands of times and computed the treatment effect in each case, the result would be a null distribution: what we might expect the treatment effect to look like if there was *no* condition effect. In essence we're using a simulated version of "random assignment" here to *break* the dependency between the condition manipulation and the observed data.

We can then compare our *actual* treatment effect to this nonparametric null distribution. If the actual treatment was smaller than the 2.5th percentile or larger than the 97.5th percentile in the null distribution, we would reject the null with $p < .05$, just the same as if we had used a t-test.

Resampling-based tests are extremely useful in a wide variety of cases. They can sometimes be less powerful than parametric approaches and they almost always require more computation, but their versatility makes them a great generic tool for data analysis.

6.4 Inference and its discontents

In earlier sections of this chapter, we reviewed NHST and Bayesian approaches to inference. Now it's time to step back and think about some of the ways that inference practices – especially those related to NHST – have been problematic for psychology research. We'll begin with some issues surrounding p-values and then give a specific accident report related to the process of "p-hacking" and some general philosophical discussion of how statistical testing relates to human reasoning.

6.4.1 Problems with the interpretation of p-values

p-values are basically likelihoods, in the sense we introduced in the previous chapter.[16] They are the likelihood of the data under the null hypothesis! This likelihood is a critical number to know – for computing the Bayes Factor among other reasons. But it doesn't tell us a lot of things that we might like to know!

[16] The only thing that is different is the idea that they are the likelihood of the observed data *or any more extreme*.

For example, p-values don't tell us the probability of the data under a specific alternative hypothesis that we might be interested in – that's the posterior probability $p(H_1|\text{data})$. When our tea-tasting t-test yielded $t(46) = 2.86$, $p = .006$, that p is *not* the probability of the null hypothesis being true! And it's definitely not the probability of milk-first tea being better.

What can you conclude when $p > .05$? According to the classical logic of NHST, the answer is "nothing"! A failure to reject the null hypothesis doesn't give you any additional evidence *for* the null. Even if the probability of the data (or some more extreme data) under H_0 is high, their probability might be just as high or higher under H_1.[17] But many practicing researchers make this mistake. Aczel et al. (2018) coded a sample of articles from 2015 and found that 72% of negative statements were inconsistent with the logic of their statistical paradigm of choice – most were cases where researchers said that an effect was not present when they had simply failed to reject the null.

[17] Of course, weighing these two against one another brings you back to the Bayes Factor.

These are not the only issues with p-values. In fact, people have so much trouble understanding what p-values *do* say that there are whole articles written about these misconceptions. Table 6.3 shows a set of misconceptions documented and refuted by S. N. Goodman (2008).

Let's take a look at just a few. Misconception 1 is that, if $p = .05$, the null has a 5% chance of being true. This misconception is a result of confusing $p(H_0|\text{data})$ (the posterior) and $p(\text{data}|H_0)$ (the likelihood – also known as the p-value). Misconception 2 – that $p > .05$ allows us to *accept* the null – also stems from this reversal of posterior and likelihood.

And misconception 3 is a misinterpretation of the p-value as an effect size (which we learned about in the last chapter): a large effect is likely to be clinically important, but with a large enough sample size, you can get a small p-value even for a very small effect. We won't go through all the misconceptions here, but we encourage you to challenge yourself to work through them (as in the exercise below).

Table 6.3: A "dirty dozen" p-value misconceptions. Adapted from S. N. Goodman (2008).

	Misconception
1	"If $p = .05$, the null hypothesis has only a 5% chance of being true."
2	"A nonsignificant difference (e.g., $p \geq .05$) means there is no difference between groups."
3	"A statistically significant finding is clinically important."
4	"Studies with p-values on opposite sides of .05 are conflicting."
5	"Studies with the same p-value provide the same evidence against the null hypothesis."
6	"p = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis."
7	"p = .05 and $p \leq .05$ mean the same thing."
8	"p-values are properly written as inequalities (e.g., '$p \leq .02$' when $p = .015$)"
9	"p = .05 means that if you reject the null hypothesis, the probability of a false positive error is only 5%."
10	"With a p = .05 threshold for significance, the chance of a false positive error will be 5%."
11	"You should use a one-sided p-value when you don't care about a result in one direction, or a difference in that direction is impossible."
12	"A scientific conclusion or treatment policy should be based on whether or not the p value is significant."

Beyond these misconceptions, there's another problem. The p-value is a probability of a certain set of events happening (corresponding to the observed data or any "more extreme" data, that is to say, data further from the null). Since p-values are probabilities, we can combine them together across different events. If we run a "null experiment" – an experiment where the true effect is zero – the probability of a dataset with $p < .05$ is of course .05. But if we run two such experiments, we can get $p < .05$ with probability 0.1. By the time we run 20 experiments, we have an 0.64 chance of getting a positive result.

It would obviously be a major mistake to run 20 experiments and then report only the positive ones (which, by design, are false positives) as though these still were "statistically significant." The same thing applies to doing 20 different statistical tests within a single experiment. There are many statistical corrections that can be made to adjust for this problem, which is known as the problem of **multiple comparisons**.[18] But the the broader issue is one of transparency: unless you *know* what the appropriate set of experiments or tests is, it's not possible to implement one of these corrections![19]

[18] The simplest and most versatile one, the Bonferroni correction, just divides .05 (or technically, whatever your threshold is) by the number of comparisons you are making. Using that correction, if you do 20 null experiments, you would have a 3% chance of a false positive.

[19] This issue is especially problematic with p-values because they are so often presented as an independent set of tests, but the problem of multiple comparisons comes up when you compute a lot of independent Bayes Factors as well. "Posterior hacking" via selective reporting of Bayes Factors is perfectly possible (Simonsohn 2014).

Accident report

Do extraordinary claims require extraordinary evidence?

In a blockbuster paper that may have inadvertently kicked off the replication crisis, Bem (2011) presented nine experiments he claimed provided evidence for precognition – that participants somehow had foreknowledge of the future. In the first of these experiments, Bem showed each of a group of 100 undergraduates 36 two-alternative forced choice trials in which they had to guess which of two locations on a screen would reveal a picture immediately before the picture was revealed. By chance, participants should choose the correct side 50% of the time of course. Bem found that, specifically for erotic pictures, participants' guesses were 53.1% correct. This rate of guessing was unexpected under the null hypothesis of chance guessing ($p = .01$). Eight other studies with a total of more than 1,000 participants yielded apparently supportive evidence, with participants appearing to show a variety of psychological effects even before the stimuli were shown!

Based on this evidence, should we conclude that precognition exists? Probably not. Wagenmakers et al. (2011) presented a critique of Bem's findings, arguing that 1) Bem's experiments were exploratory (not pre-registered) in nature, 2) that Bem's conclusions were *a priori* unlikely, and 3) that the level of statistical evidence from his experiments was quite low. We find each of these arguments alone compelling; together they present a knockdown case against Bem's interpretation.

First, we've already discussed the need to be skeptical about situations where experimenters have the opportunity for analytic flexibility in their choice of measures, manipulations, samples, and analyses. Flexibility leads to the possibility of cherry-picking those set of decisions from the "garden of forking paths" that lead to a positive outcome for the researcher's favored hypothesis (for more details, see Chapter 11). And there is plenty of flexibility on display even in Experiment 1 of Bem's paper. Although there were 100 participants in the study, they may have been combined post hoc from two distinct samples of 40 and 60, each of which saw different conditions. The 40 made guesses about the location of erotic, negative, and neutral pictures; the 60 saw erotic, positive non-romantic, and positive romantic pictures. The means of each of these conditions was presumably tested against chance (at least 6 comparisons, for a false positive rate of 0.26). Had positive romantic pictures been found significant, Bem certainly could have interpreted this finding the same way he interpreted the erotic ones.

Second, as we discussed, a p-value close to .05 does not necessarily provide strong evidence against the null hypothesis. Wagenmakers et al. computed the Bayes Factor for each of experiments in Bem's paper and found that, in many cases, the amount of evidence for H_1 was quite modest under a default Bayesian t-test. Experiment 1 was no exception: the BF was 1.64, giving only "anecdotal" support for the hypothesis of some non-zero effect, even before the multiple-comparisons problem mentioned above.

Finally, since precognition is not supported by any prior compelling scientific evidence (despite many attempts to obtain such evidence) and defies well-established physical laws, perhaps we should assign a low prior probability to Bem's H_1, a non-zero precognition effect. Taking a strong Bayesian position, Wagenmakers et al. suggest that we might do well to adopt a prior reflecting how unlikely precognition is, say $p(H_1) = 10^{-20}$. And if we adopt this prior, even a very well-designed, highly informative experiment (with a Bayes factor conveying substantial or even decisive evidence) would still lead to a very low posterior probability of precognition.

Wagenmakers et al. concluded that, rather than supporting precognition, the conclusion from Bem's paper should be psychologists should revise how they think about analyzing their data (and avoid p-hacking)!

6.4.1 Philosophical (and empirical) views of probability

Up until now we've presented Bayesian and frequentist tools as two different sets of computations. But in fact, these different tools derive from fundamentally different philosophical perspectives on what a probability even is. Very roughly, frequentist approaches tend to believe that probabilities quantify the long-run frequencies of certain events. So, if we say that some outcome of an event has probability .5, we're saying that if that event happened thousands of times, the long run frequency of the outcome would be 50% of the total events. In contrast, the Bayesian viewpoint doesn't depend on this sense that events could be exactly repeated. Instead, the **subjective Bayesian** interpretation of probability is that it quantifies a person's degree of belief in a particular outcome.[20]

[20] This is really a very rough description. If you're interested in learning more about this philosophical background, we recommend the Stanford Encyclopedia of Philosophy entry, "Interpretations of Probability" (https://plato.stanford.edu/entries/probability-interpret/).

You don't have to take sides in this deep philosophical debate about what probability is. But it's helpful to know that people actually seem to reason about the world in ways that are well described by the subjective Bayesian view of probability. Recent cognitive science research has made a lot of headway in describing reasoning as a process of Bayesian inference where probabilities describe degrees of belief in different hypotheses (for a textbook review of this approach, see N. D. Goodman, Tenenbaum, and Contributors 2016). These hypotheses in turn are a lot like the theories we described in Chapter 2: they describe the relationships between different abstract entities (Tenenbaum et al. 2011). You might think that scientists are different from lay-people in this regard, but one of the striking findings from research on probabilistic reasoning and judgment is that expertise doesn't matter that much. Statistically-trained scientists – and even statisticians – make many of the same reasoning mistakes as their un-trained students (Kahneman and Tversky 1979). Even children seem to reason intuitively in a way that looks a bit like Bayesian inference (Gopnik 2012).

These cognitive science findings help to explain some of the problems that people (scientists included) have in reasoning about p-values. If you are an intuitively Bayesian reasoner, the quantity that you're probably tracking is how much you believe in your hypothesis (its posterior probability). So, many people treat the p-value as the posterior probability of the null hypothesis.[21] That's exactly what fallacy #1 in Table 6.3 states – "If $p = .05$, the null hypothesis has only a 5% chance of being true." But this equivalence is incorrect! Written in math, $p(\text{data}|H_0)$ (the likelihood that lets us compute the p-value) is not the same thing as $p(H_0|\text{data})$ (the posterior that we want). Pulling from our accident report above, even if the *probability of the observed ESP data given the null hypothesis* is low, that doesn't mean that the *probability of ESP* is high.

[21] Cohen (1994) is a great treatment of this issue.

6.4.2 What framework to use?

The problem with binary inferences is that they enable behaviors that can introduce bias into the scientific ecosystem. By the logic of statistical significance, either an experiment "worked" or it didn't. Because everyone would usually rather have an experiment that worked than one that didn't, inference criteria like p-values often become a target for selection, as we discussed in Chapter 3.[22]

If you want to quantify evidence for or against a hypothesis, it's worth considering whether Bayes Factors address your question better than p-values. In practice, p-values are hard to understand and many people misuse them – though to be fair, BFs are misused plenty too. These issues may be rooted in basic facts about how human beings reason about probability.

Despite the reasons to be worried about p-values, for many practicing scientists (at least at time of writing) there is no one right answer about whether to use them or not. Even if we'd like to be Bayesian all the time, there are a number of obstacles. First, though new computational tools make fitting Bayesian models and extracting Bayes Factors much easier than before, it's still on average quite a bit harder to fit a Bayesian model than it is a frequentist one. Second, because Bayesian analyses are less familiar, it may be an uphill battle to convince advisors, reviewers, and funders to use them.

As a group of authors, some of us are more Bayesian than frequentist, while others are more frequentist than Bayesian – but all of us recognize the need to move between statistical paradigms depending on the problem we're working on. Furthermore, a lot of the time we're not so worried about which paradigm we're using. The paradigms are at their most divergent when making binary inferences, and they often look much more similar when they are used in the context of quantifying measurement precision.

6.5 Computing precision

Our last section presented an argument against using p-values for making *dichotomous* inferences. But we still want to move from what we know about our own limited sample to some inference about the population. How should we do this?

[22] More generally, this pattern is probably an example of Goodhart's law, which states that when a measure becomes a target, it ceases to be a good measure (Strathern 1997). Once the outcomes of statistical inference procedures become targets for publication, they are subject to selection biases – p-hacking, for example – that make them less meaningful.

6.5.1 Confidence intervals

One alternative to binary hypothesis testing is to ask about the precision of our estiamates, in particular 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 our 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 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 intervals $\hat{\beta} \pm 1.96\ \widehat{SE}$.[23] 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.[24]

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.

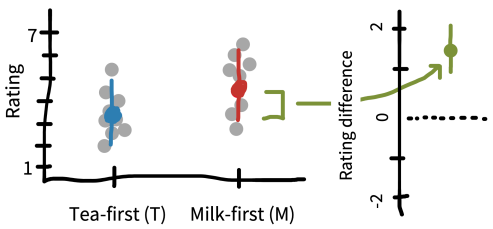


Figure 6.11: Confidence intervals on each of the two condition estimates, as well as on the difference between conditions.

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

[24] 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.

Code

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)
```

For visualization purposes, 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.

6.5.2 *Confidence in confidence intervals?*

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*.

Table 6.4: Confidence interval misconceptions for a confidence interval [0.1,0.4]. Adapted from Hoekstra et al. (2014).

	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."

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.[25] 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.

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

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 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 one another in many cases.[26]

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

6.6 Chapter summary: Inference

Inference tools help you move from characteristics of the sample to characteristics of the population. This move is a critical part of generalization from research data. But we hope we've convinced you that inference doesn't have to mean making a binary decision about the presence or absence of an effect. A strategy that seeks to estimate an effect and

its associated precision is often much more helpful as a building block for theory. As we move towards estimating causal effects in more complex experimental designs, the process will require more sophisticated models. Towards that goal, the next chapter provides some guidance for how to build such models.

Discussion questions

1. Can you write the definition of a p-value and a Bayes Factor without looking them up? Try this out – what parts of the definitions did you get wrong?
2. Take three of Goodman's (2008) "dirty dozen" in Table 6.3) and write a description of why each is a misconception. (These can be checked against the original article, which gives a nice discussion of each.

Readings

- Many of the concepts described here are illustrated beautifully via interactive visualizations. We recommend https://seeing-theory.brown.edu/ for a broad overview of statistical concepts and https://rpsychologist.com/viz for a number of interactives that specifically illustrate concepts discussed in this chapter and the previous one, including p-values, effect sizes, maximum likelihood estimation, confidence intervals, and Bayesian inference.
- A fun, polemical critique of NHST: Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49, 997–1003. https://doi.org/10.1037/0003-066X.49.12.997.
- A nice introduction to Bayesian data analysis: Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. Psychonomic bulletin & review, 25(1), 155-177. https://doi.org/10.3758/s13423-017-1272-1.

References

Aczel, Balazs, Bence Palfi, Aba Szollosi, Marton Kovacs, Barnabas Szaszi, Peter Szecsi, Mark Zrubka, Quentin F Gronau, Don van den Bergh, and Eric-Jan Wagenmakers. 2018. "Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation." *Advances in Methods and Practices in Psychological Science* 1 (3): 357–66.

Bem, Daryl J. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." *Journal of Personality and Social Psychology* 100 (3): 407.

Benjamin, Daniel J, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6–10.

Cohen, Jacob. 1990. "Things i Have Learned (so Far)." *American Psychologist* 45: 1304–12.

———. 1994. "The Earth Is Round (p<. 05)." *American Psychologist* 49 (12): 997.

Cumming, Geoff. 2014. "The New Statistics: Why and How." *Psychol. Sci.* 25 (1): 7–29.

Fisher, Ronald A. 1949. *The Design of Experiments*. 5th ed. Oliver & Boyd.

Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin. 1995. *Bayesian Data Analysis*. Chapman; Hall/CRC.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge University Press.

Gigerenzer, Gerd. 1989. *The Empire of Chance: How Probability Changed Science and Everyday Life*. 12. Cambridge University Press.

Goodman, Noah D, Joshua B. Tenenbaum, and The ProbMods Contributors. 2016. "Probabilistic Models of Cognition." https://probmods.org/.

Goodman, Steven N. 1999. "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor." *Annals of Internal Medicine* 130 (12): 1005–13.

———. 2008. "A Dirty Dozen: Twelve p-Value Misconceptions." In *Seminars in Hematology*, 45:135–40. 3. Elsevier.

Gopnik, Alison. 2012. "Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications." *Science* 337 (6102): 1623–27.

Hoekstra, Rink, Richard D Morey, Jeffrey N Rouder, and Eric-Jan Wagenmakers. 2014. "Robust Misinterpretation of Confidence Intervals." *Psychonomic Bulletin & Review* 21 (5): 1157–64.

Jeffreys, Harold. 1961. *The Theory of Probability*. 3rd ed. OUP Oxford.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica* 47 (2): 363–91.

Kruschke, John K. 2014. *Doing Bayesian Data Analysis: A Tutorial with r, JAGS, and Stan*. Academic Press.

Kruschke, John K, and Torrin M Liddell. 2018. "The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis from a Bayesian Perspective." *Psychon. Bull. Rev.* 25 (1): 178–206.

McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in r and Stan*. Chapman; Hall/CRC.

Morey, Richard D, Rink Hoekstra, Jeffrey N Rouder, Michael D Lee, and Eric-Jan Wagenmakers. 2016. "The Fallacy of Placing Confidence in Confidence Intervals." *Psychonomic Bulletin & Review* 23 (1): 103–23.

Morey, Richard D, and Jeffrey N Rouder. 2011. "Bayes Factor Approaches for Testing Interval Null Hypotheses." *Psychological Methods* 16 (4): 406.

Pratt, John Winsor, Howard Raiffa, Robert Schlaifer, et al. 1995. *Introduction to Statistical Decision Theory*. MIT press.

Simonsohn, Uri. 2014. "Posterior-Hacking: Selective Reporting Invalidates Bayesian Results Also." https://doi.org/https://dx.doi.org/10.2139/ssrn.2374040.

Strathern, Marilyn. 1997. "'Improving Ratings': Audit in the British University System." *European Review* 5 (3): 305–21.

Tenenbaum, Joshua B, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 331 (6022): 1279–85.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, and Han LJ Van Der Maas. 2011. "Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011)." *Journal of Personality and Social Psychology* 100 (3): 426–32. https://doi.org/https://doi.org/10.1037/a0022790.

7 Models

Learning goals

- Articulate a strategy for estimating experimental effects using statistical models
- Build intuitions about how classical statistical tests relate to linear regression models
- Explore variations of the linear model, including generalized linear models and mixed effects models
- Reason about tradeoffs and strategies for model specification, including the use of control variables

In the previous two chapters, we introduced concepts surrounding estimation of an experimental effect and inference about its relationship to the effect in the population. The tools we introduced there are for fairly specific research questions, and so are limited in their applicability. Once you get beyond the world of two-condition experiments in which each participant contributes one data point from a continuous measure, the simple t-test is not sufficient.

In some statistics textbooks, the next step would be to present a whole host of other statistical tests that are designed for other special cases. We could even show a decision-tree: you have repeated measures? Use Test X! Or categorical data? Use Text Y! Or three conditions? Use Test Z! But this isn't a statistics book, and even if it were, we don't advocate that approach. The idea of finding a specific narrowly-tailored test for your situation is part and parcel of the dichotomous NHST approach that we tried to talk you out of in the last chapter. If all you want is your $p < .05$, then it makes sense to look up the test that can allow you to compute a p value in your specific case. But we prefer an approach that is more focused on getting a good estimate of the magnitude of the causal effect.

In this chapter, we begin to explore how to select an appropriate **statistical model** to clearly and flexibly reason about these effects. A statistical model is a way of writing down a set of assumptions about how particular data are generated, the **data generating process**. Statistical models are the bread and butter tools for estimating particular **parameters** of interest from empirical data – like the magnitude of a causal effect associated with an experimental manipulation. They can also quantify MEASUREMENT PRECISION.

For example, suppose you watch someone tossing a coin and observe a sequence of heads and tails. A simple statistical model might assume that the observed data are generated via the flip of a weighted coin. From the perspective of the last two chapters, we could estimate a standard error for the estimated proportion of flips that are heads (e.g., for 6 heads out of 8 flips, we have $\hat{p} = 0.75 \pm 0.17$), or we could compare the observed proportion against a null hypothesis. From a model-based perspective, however, we instead begin by thinking about where the data came from: we might assume the coin being flipped has some weight (a **latent**, or unobservable, parameter of the data generating process), and our goal is to determine the most likely value of that weight given the observed data. This single unified model can then also be used to make inferences about whether the coin's weight differs from some null model (a fair coin, perhaps), or to predict future flips.

This example sounds a lot like the kinds of simple inferential tests we talked about in the previous chapter; not very "model-y." But things get more interesting when there are multiple parameters to be estimated, as in many real-world experiments. In the tea-tasting scenario we've belabored over the past two chapters, a real experiment might involve multiple people tasting different types of tea in different orders, all with some cups randomly assigned to be milk-first or tea-first. What we'll learn to do in this chapter is to make a model of this situation that allows us to reason about the magnitude of the milk-order effect while also estimating variation due to different people, orders, and tea types. This is the advantage of using models: once you are able to reason about estimation and inference in model-based terms, you will be set free from long decision trees and will be able to flexibly make the assumptions that make sense for your data.[1]

We'll begin by discussing the ubiquitous framework for building statistical models, **linear regression**.[2] We will then build connections between regression and the t-test. This section will discuss how to add covariates to regression models, and when linear regression does and doesn't work. In the following section, we'll discuss the **generalized linear model**, an innovation that allows us to make models of a broader range of data types, including **logistic regression**. We'll then briefly introduce **mixed models**, which allow us to model clustering in our datasets (such as clusters of observations from a single individual or single stimulus item). We'll end with some opinionated practical advice on model building.

If you're interested in building up intuitions about statistical model building, then we recommend reading this chapter all the way through. On the other hand, if you are already engaged in data analysis and want to see an example, we suggest that you skip to the last section,

[1] We won't explore the connection to DAGs and Bayesian models here, but one way to think of this model building is as creating a causal theory of the experiment. This approach, which is advocated by McElreath (2018), creates powerful connections between the ideas about theory we presented in Chapters 1 and 2 and the ideas about models here. If this sounds intriguing, we encourage you to go down the rabbit hole!

[2] The name regression originally comes from Galton (1877)'s work on heredity. He was looking at the relationship between the heights of parents and children. He found that children's heights *regressed*, and he did so by creating a *regression model*. Now we use the term "regression" to mean any model of this form.

where we give some opinionated practical advice on model building and provide a worked example of fitting a mixed effects model and interpreting it in context.

7.1 Regression models

There are many types of statistical models, but this chapter will focus primarily on regression, a broad and extremely flexible class of models. A regression model relates a dependent variable to one or more independent variables. Dependent variables are sometimes called **outcome variables**, and independent variables are sometimes called **predictor variables**, **covariates**, or **features**.[3] We will see that many common statistical estimators (like the sample mean) and methods of inference (like the t-test) are actually simple regression models. Understanding this point will help you see many statistical methods as special cases of the same underlying framework, rather than as unrelated, *ad hoc* tests.

7.1.1 Regression for estimating a simple treatment effect

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 $\theta_{\text{M}} = \theta_{\text{T}} + \beta$ and standard deviation σ.[4]

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.[5] 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 were using to denote the causal effect above and in the previous two chapters.

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

[3] The reverse is not true – not every predictor or covariate is an independent variable! One of the tricky things about relating regression models to causal hypotheses is that, just because something is on the right side of a regression equation, that doesn't mean it's a causal manipulation. And of course, just because you've got an estimate of some predictor in a regression, that doesn't mean the estimate tells you about the magnitude of the *causal* effect. It could, but it also might not!

[4] 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" – generate data through the process that's specified in those equations, e.g., 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.

[5] 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!

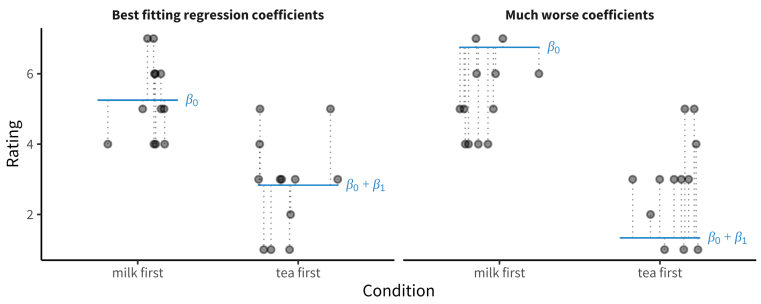
Table 7.1: Example tea tasting data.

id	condition	X	rating (Y)
1	milk first	0	6
2	milk first	0	4
3	milk first	0	5
4	tea first	1	1
5	tea first	1	3
6	tea first	1	5

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.[6] 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.[7]

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**, $\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i$ for each participant. Some regression coefficient estimates will yield better predictions than others. OLS 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.



[6] 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.

[7] 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.

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.

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.[8] 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.

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. As we said, we're not going to give a full treatment of all the relevant statistical topics. Instead we want to help you begin thinking about making models of your data.

[8] OLS minimizes squared error loss, in the sense that it will choose the regression coefficient estimates whose predictions minimize $\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$, where n is the sample size. A wonderful thing about OLS is that those optimal regression coefficients (generically termed $\widehat{\ }$) turn out to have a very simple closed form solution: $\widehat{\ } = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. We are using more general notation here that supports multiple independent variables: $\widehat{\ }$ 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 $\widehat{\ }$ has a similarly simple closed form!

Code

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)
```

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 the parameters of the model:

```
summary(mod)
```

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

7.1.2 Adding predictors

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.[9]

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

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 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 1 year of age. In "Practical modeling considerations" below, we discuss whether and when to "control for" additional variables (i.e., when to add them to your model).

7.1.3 Interactions

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 manipulationes 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.[10]

7.1.4 When does linear regression work?

Linear regression modeling with OLS is an incredibly powerful technique for creating models to estimate the influence of multiple predictors on a single dependent variable. In fact, OLS is in a mathematical sense the *best* way to fit a linear model![11] But OLS only "works" – in the sense of yielding good estimates – if three big conditions are met.

1. **The relationship between the predictor and outcome must be linear**. In our comparison of Alice's and Bob's expected outcomes based on their 1-year age difference, we were able to interpret the coefficient β_2 as the average difference in Y_i when comparing participants who differ by 1 year of age, *regardless* of whether those ages are 20 vs. 21 or 50 vs. 51. If we believed this relationship was **non-linear**, then we could transform our predictor – for example, including a **quadratic** effect of age by adding a $\beta_3 * X_2^2$ term. The *relationship* between this new predictor and the outcome would still be linear, however. It is always a good idea to use visualizations like scatter plots to look for possible problems with assuming a linear relationship between a predictor and your outcome.

2. **Errors must be independent**. In our example, observations in the regression model (i.e., rows in the dataset) were sampled independently: each participant was recruited independently to the study and each performed a single trial. On the other hand, suppose we have repeated-measures data in which we sample participants, and then obtained multiple measurements for each partici-

[10] We won't go into this topic here, but we do want to provide a pointer to one of the most persistent challenges that come 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 **contrast coding**. Davis (2010) gives a good tutorial on this tricky topic.

[11] There is a precise sense in which OLS gives the *very best* predictions we could ever get from any model that posits linear relationships between the independent variables and the outcome. That is, you can come up with any other linear, unbiased model you want, and yet if the assumptions of OLS are fulfilled, predictions from OLS will always be less noisy than those of your model. This is because of an elegant mathematical result called the Gauss-Markov Theorem.

pant. Within each participant, measurements would likely be correlated (perhaps because participants differ on their general level of tea enjoyment). This correlation can invalidate inferences from a model that does not accommodate the correlation. We'll discuss this problem in detail below.

3. **Errors must be normally distributed and unrelated to the predictor**. Imagine older people have very consistent tea-ordering preferences while younger people do not. In that case, the models' error term would be less variable for older participants than younger ones. This issue is called **heteroskedasticity**. It is a good idea to plot each independent variable versus the residuals to see if the residuals are more variable for certain values of the independent variable than for others.

If any of these three conditions are violated, it can undermine the estimates and inferences you draw from your model.

7.2 Generalized linear models

So far we have considered continuous outcome measures, like tea ratings. What if we instead had a binary outcome, such as whether a participant liked or didn't like the tea, or a count outcome, such as the number of cups a participant chose to drink? These and other non-continuous outcomes often violate the assumptions of OLS, in particular because they often induce heteroskedastic errors.

Binary outcomes inherently produce heteroskedastic errors because the variance of a binary variable depends directly on the outcome probability. Errors will be more variable when the outcome probability is closer to 0.50, and much less variable for when the probability is closer to 0 or 1.[12] This heteroskedasticity in turn means that inferences from the model (e.g., p-values) can be incorrect; sometimes just a little bit off but sometimes dramatically incorrect.[13]

Happily, **generalized linear models** (GLMs) are regression models closely related to OLS that can handle non-continuous outcomes. These models are called "generalized" because OLS is one of many members of this large class of models. To see the connection, let's first write an OLS model more generally in terms of what it says about the expected value of the outcome, which we notate as $E[Y_i]$:

$$E[Y_i] = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$$

[12] Specifically, the variance of a binary variable with probability p is simply $p(1-p)$, which is largest when $p = 0.50$.

[13] OLS can also be used with binary outcomes, in which case the coefficients represent differences in probabilities. However, the usual model-based standard errors will be incorrect.

where p is the number of independent variables, β_0 is the intercept, and β_j is the regression coefficient for the j^{th} independent variable. This equation is just a math-y way of saying that you predict from a regression model by adding up each of the predictors' contributions to the expected outcome ($\beta_j X_{ij}$).

The linear predictor of a GLM (i.e., $\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$) looks exactly the same as for OLS, but instead of modeling $E[Y_i]$, a GLM models some **transformation**, $g(.)$, of the expectation:

$$g(E[Y_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$$

GLMs involve transforming the *expectation* of the outcome, not the outcome itself! That is, in GLMs, we are not just taking the outcome variable in our dataset and transforming it before fitting an OLS model, but rather we are fitting a different model entirely, one that posits a fundamentally different relationship between the predictors and the expected outcomes. This transformation is called the **link function**. In other words, to fit different kinds of outcomes, all we need to do is construct a standard linear model and then just transform its output via the appropriate link function.

Perhaps the most common link function is the **logit** link, which is suitable for binary data. This link function looks like this, where w is any probability that is strictly between 0 and 1:

$$g(w) = \log\left(\frac{w}{1-w}\right)$$

The term $w/(1-w)$ is called the **odds** and represents the probability of an event occurring divided by the probability of its not occurring. The resulting model is called **logistic regression** and looks like:

$$\text{logit}(E[Y_{it}]) = \log\left(\frac{E[Y_i]}{1-E[Y_i]}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$$

Exponentiating the coefficients (i.e., e^{β}) would yield **odds ratios**, which are the *multiplicative* increase in the odds of $Y_i = 1$ that is associated with a one-unit increase in the relevant predictor variable.

Figure 7.2 shows the way that a logistic regression model transforms a predictor (X) into an outcome probability that is bounded at 0 and 1. Critically, although the predictor is still linear, the logit link means that the same change in X can result in a different change in the absolute probability of Y depending on where you are on the X scale. In this

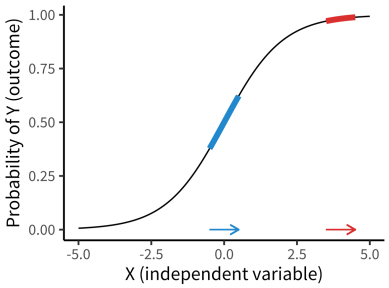


Figure 7.2: An example of how logistic regression transforms a change in the mean-centered predictor X into a change in the expected outcome Y. The same absolute change in X is associated in a large difference in the probability of the outcome when X is near its mean (blue) vs. a small change in the outcome when X is large (red) or small.

example, if you are in the middle of the predictor range, a one-unit change in X results in a 0.24 change in probability (blue). At a higher value, the change is much smaller (0.02). Notice how this is different from the linear regression model above, where the same change in age always resulted in the same change in preference!

Code

GLMs are as easy to fit in R as standard LMs. You simply need to call the `glm()` function – and to specify the link function. For our example above of a binary "liking" judgment, the call would be:

```
glm(liked_tea ~ condition, data = tea_data, family = "binomial")
```

The `family` argument specifies the type of distribution being used, where `binomial` is the logistic link function.

We have only scratched the surface of GLMs here. First, there are many different link functions that are suitable for different outcome types. And second, GLMs differ from OLS not only in their link functions, but also in how they handle the error terms. Our broader goal in this chapter is to show you how regression models are *models of data*. In that context, GLMs use link functions as a way to make models that generate many different times types of outcome data.[14]

[14] We sometimes think of linear models as a set of tinker toys you can snap together to stack up a set of predictors. In that context, link functions are an extra "attachment" that you can snap onto your linear model to make it generate a different response type.

7.3 Linear mixed effects models

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 effect. Here is the estimate and confidence interval for that coefficient: $b = -2.42$, 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![15] 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. 2014).

[15] 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.

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).[16]

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

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

able to attribute this difference to participant-level variation rather than to the treatment.[17]

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.

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

Code

Remarkably, GLMMs are not much harder to specify in R than standard LMs. One very popular package is `lme4` (Bates et al. 2014), which provides the `lmer()` and `glmer()` functions (the latter for generalized linear mixed effect models). For our example here, we'd write:

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

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.[18]

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.[19]

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

[19] 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.

Code

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

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

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.

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![20] 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).[21]

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.[22] 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 How do you use models to analyze data?

In the prior parts of this chapter, we've described a suite of regression-based techniques – standard OLS, the generalized linear model, and linear mixed effects models – that can be used to model the data resulting from randomized experiments (as well as many other kinds of data). The advantage of regression models over the simpler estimation and inference methods we described in the prior two chapters is that these models can more effectively take into account a range of different kinds of variation including covariates, multiple manipulations, and clustered structure. Further, when used appropriately to analyze a well-designed randomized experiment, regression models can give an unbiased estimate of a causal effect of interest, our main goal in doing experiments.

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

[21] Many R users may be familiar with the widely-used `lme4` package 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, which provides many powerful methods for specifying complex models.

[22] 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).

But – practically speaking – how should go you about building a model for your experiment? What covariates should you include and what should you leave out? There are many ways to use models to explore datasets, but in this section we will try to sketch a default approach for the use of models to estimate causal effects in experiments in the most straightforward way. Think of this as a starting point. We'll begin this section by giving a set of rules of thumb, then discuss a worked example. Our final subsections will deal with the issues of when you should include covariates in your model and how to check if your result is robust across multiple different model specifications.

Depth

An alternative approach: Generalized estimating equations

A second class of methods that helps resolve issues of clustering is **generalized estimating equations** (GEE). In this approach, we leave the linear predictor alone. We do not add random intercepts or slopes, nor do we assume anything about the distribution of the errors (i.e., we no longer assume that they are normal, independent, and homoskedastic).

In GEE, we instead provide the model with an initial "guess" about how we think the errors might be related to one another; for example, in a repeated-measures experiment, we might guess that the errors are exchangeable, meaning that they are correlated to the same degree within each participant but are uncorrelated across participants. Instead of *assuming* that our guess is correct, as do linear mixed models (LMM), GEE estimates the correlation structure of the errors empirically, using our guess as a starting point, and it uses this correlation structure to arrive at point estimates and inference for the regression coefficients. Remarkably, as the number of clusters and observations become very large, GEE will *always* provide unbiased point estimates and valid inference, *even if* our guess about the correlation structure was bad. Additionally, with simple finite-sample corrections (Mancl and DeRouen 2001), GEE seems to provide valid inference at smaller numbers of clusters than does LMM.

The price paid for these nice safeguards against model misspecification is that, in principle, GEE will typically have less statistical power than LMM *if* the LMM is in fact correctly specified, but the difference may be surprisingly slight in practice (Bie et al. 2021). For these reasons, some of this book's authors actually favor GEE with finite-sample corrections over LMM as the default model for clustered data, though they are much less common in psychology.

7.4.1 Modeling rules of thumb

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.[23]

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![24]

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.[25]
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 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 the 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

[23] 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.

[24] 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 sub-field, 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.

[25] 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?

missing data. The default model simply represents a very good starting point.

7.4.2 A worked example

Figure 7.3: Example stimulus materials analogous to those used in Stiller, Goodman, and Frank (2015).

All this advice may seem abstract, so let's put it into practice on a simple example. For a change, let's look at an experiment that's not about tea tasting. Here we'll consider data from an experiment testing preschool children's language comprehension (Stiller, Goodman, and Frank 2015). In this experiment, 2–5 year old children saw displays like the one in Figure 7.3. In the experimental condition, a puppet might say, for example, "My friend has glasses! Which one is my friend?" The goal was to measure how many children made the "pragmatic inference" that the puppet's friend was the face with glasses and *no* hat.

To estimate the effect, participants were randomly assigned to either the experimental condition or to a control condition in which the puppet had eaten too much peanut butter and couldn't talk, but they still had to guess which face was his friend. There were also three other types of experimental stimuli (houses, beds, and plates of pasta). Data from this experiment consisted of 588 total observations from 147 children, with all four stimuli presented to each child. The primary hypothesis of this experiment was that that preschool children could make pragmatic inferences by correctly inferring which of the three faces (for example) the puppet was describing.

Code

If you want to follow along with this example, you'll have to load the example data and do a little bit of preprocessing (also covered in Appendix D):

```
repo <- "https://raw.githubusercontent.com/langcog/experimentology/main/"
sgf <- read_csv(file.path(repo, "data/tidyverse/stiller_scales_data.csv")) |>
  mutate(age_group = cut(age, 2:5, include.lowest = TRUE),
         condition = condition |>
           fct_recode("Experimental" = "Label", "Control" = "No Label"))
```

All this advice may seem abstract, so let's put it into practice on a simple example. For a change, let's look at an experiment that's not about tea tasting. Here we'll consider data from an experiment testing preschool children's language comprehension , we also use these data in D. In this experiment, 2–5 year old children saw displays like the one in Figure 7.3. In the experimental condition, a puppet might say, for example, "My friend has glasses! Which one is my friend?" The goal was to

measure how many children made the "pragmatic inference" that the puppet's friend was the face with glasses and *no* hat.

To estimate the effect, participants were randomly assigned to either the experimental condition or to a control condition in which the puppet had eaten too much peanut butter and couldn't talk, but they still had to guess which face was his friend. There were also three other types of experimental stimuli (houses, beds, and plates of pasta). Data from this experiment consisted of 588 total observations from 147 children, with all four stimuli presented to each child. The primary hypothesis of this experiment was that that preschool children could make pragmatic inferences by correctly inferring which of the three faces (for example) the puppet was describing.

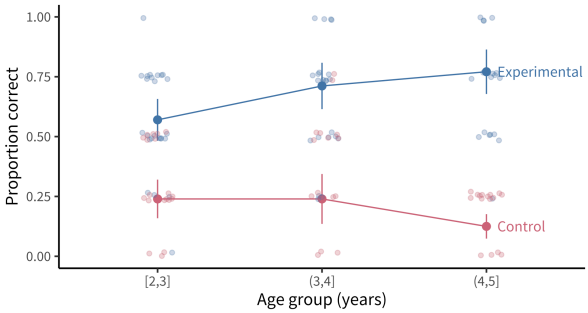


Figure 7.4: Data for Stiller, Goodman, and Frank (2015). Each point shows a single participant's proportion correct trials (out of 4 experimental stimuli) plotted by age group, jittered slightly to avoid overplotting. Larger points and associated confidence intervals show mean and 95% confidence intervals for each condition.

This experimental design looks a lot like some versions of our tea-tasting experiment. We have one primary condition manipulation (the puppet provides information versus does not), presented between-participants so that some participants are in the experimental condition and others are in the control condition. Our measurements are repeated within participants across different experimental stimuli. Finally, we have one important, pre-planned covariate: children's age. Experimental data are plotted in Figure 7.4.[26]

How should we go about making our default model for this dataset?[27] We simply include each of these design factors in a mixed effects model; we use a logistic link function for our mixed effects model (a **generalized linear mixed effects model**) because we would like to predict correct performance on each trial, which is a binary variable. So that gives us an effect of condition and age as a covariate. We further add an interaction between condition and age in case the condition effect varies meaningfully across groups. Finally, we add random effects of participant, γ_i, and experimental item, γ_t.[28]

The resulting model looks like this:

$$\text{logit}(E[Y_{it}]) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \gamma_i + \delta_t$$

[26] Our sampling plan for this experiment was actually **stratified** across age, meaning that we intentionally recruited the same number of participants for each one-year age group – because we anticipated that age was highly correlated with children's ability to succeed in this task. We'll describe this kind of sampling in more detail in Chapter 10.

[27] This experiment was not preregistered, but the paper includes a separate replication dataset with the same analysis.

[28] As discussed above, this is a tricky decision-point; we could very reasonably have added random slopes as well.

Let's break this complex equation down left to right:

- $\text{logit}(E[Y_{it}])$ says that we are predicting a logistic function of $E[Y_{it}]$ (where Y_{it} indicates whether child i was correct on trial t).
- β_0 is the **intercept**, our estimate of the average log-odds (i.e., the log of the odds ratio) of correct responses for participants in the control condition.
- $\beta_1 X_{i1}$ is the condition predictor. β_1 represents the change in log-odds associated with being in the experimental condition (the causal effect of interest!), and X_{i1} is an indicator variable that is 1 if child i is in the experimental condition and 0 for the control condition. Multiplying β_1 by this indicator means that the predictor has the value 0 for participants in the control condition and β_1 for those in the experimental condition.
- $\beta_2 X_{i2}$ is the age predictor. β_2 represents the difference in log-odds associated with one additional year of age for participants int he control condition[The age coefficient is a **simple effect**, meaning it is the effect of age in the control condition only. That's because we have dummy coded the condition predictor. If we wanted the average age effect (the **main effect**) then we would need to use contrast coding, per the note in the Interactions section above.], and X_{i2} is the age for each participant.[29]
- $\beta_3 X_{i1} * X_{i2}$ is the interaction between experimental condition and age. β_3 represents the difference in log odds (i.e., the log of the odds ratio) that is associated with being one year older *and* in the experimental condition versus the control condition. This term is multiplied by both each child's age *and* the condition indicator X_i.
- γ_i is the random intercept for participant i, capturing individual variation in the odds of success across trials.
- γ_t is the random intercept for stimulus t, capturing variation in the odds of success across the four different stimuli.

[29] We have **centered** our age predictor in this example so that all estimates from our model are for the average age of our participants. Centering is a good practice for modeling continuous predictors because it increases the interpretability of other parts of the model. For example, because age is centered in this model, the intercept β_0 can be interpreted as the predicted odds of a correct trial for a participant in the control condition at the average age.

Table 7.2: Estimated effects for our generalized linear mixed effects model on data from Stiller, Goodman, and Frank (2015).

term	estimate	conf.int	statistic	p.value
Control condition	0.80	[0.42, 1.18]	4.16	< .001
Age (years)	0.55	[0.21, 0.88]	3.19	.001
Expt condition	-2.26	[-2.70, -1.82]	-10.07	< .001
Age (years) * Expt condition	-0.92	[-1.43, -0.42]	-3.60	< .001

Code

To fit the model described above, the first step is to prepare your predictors. In this case, we center the age predictor.

```
sgf$age_centered <- scale(sgf$age, center = TRUE, scale = FALSE)
```

Again we use the `lme4` package, this time with the `glmer()` function. Again we have to specify our link function, just like in a standard GLM, by choosing the distribution family.

```
mod <- glmer(correct ~ age_centered * condition + (1|subid) + (1|item),
             family = "binomial", data = sgf)
```

You can see a summary of the fitted model using `summary(mod)` as before. The only big difference from `lm()` is that here you can extract both fixed and random effects (with `fixef(mod)` and `ranef(mod)` respectively).

Let's estimate this model and see how it looks. We'll focus here on interpretation of the so-called **fixed effects** (the main predictors), as opposed to the participant and item random effects.[30] Table 7.2 shows the coefficients. Again, let's walk through each.

- The **intercept** (control condition estimate) is $\hat{\beta} = 0.80$, 95% CI $[0.42, 1.18]$, $z = 4.16$, $p < .001$. This estimate reflects that the log-odds of a correct response for an average-age participant in the control condition is 0.8, which corresponds to a probability of 0.69. If we look at Figure 7.4, that estimate makes sense: 0.69 seems close to the average for the control condition.
- The **age effect** estimate is $\hat{\beta} = 0.55$, 95% CI $[0.21, 0.88]$, $z = 3.19$, $p = .001$. This means there is a slight decrease in the log-odds of a correct response for older children in the control condition. Again, looking at Figure 7.4, this estimate is interpretable: we see a small decline in the probability of a correct response for the oldest age group.
- The key experimental condition estimate then is $\hat{\beta} = -2.26$, 95% CI $[-2.70, -1.82]$, $z = -10.07$, $p < .001$. This estimate means that the log-odds of a correct response for an average-age participant in the experimental condition is the sum of the estimates for the control (intercept) and the experimental conditions: 0.8 + -2.26, which corresponds to a probability of 0.19. Grounding our interpretation in Figure 7.4, this estimate corresponds to the average value for the experimental condition.
- Finally, the **interaction** of age and condition is $\hat{\beta} = -0.92$, 95% CI $[-1.43, -0.42]$, $z = -3.60$, $p < .001$. This positive coefficient

[30] Participant means are estimated to have a standard deviation of 0.23 (in log-odds) while items have a standard deviation of 0.27. These indicate that both of our random effects capture meaningful variation.

reflects that with every year of age, the difference between control and experimental conditions grows.

In sum, this model suggests that there was a substantial difference in performance between experimental and control conditions, in turn supporting the hypothesis that children in the sampled age group can perform pragmatic inferences above chance.

This example illustrates the "default saturated model" framework that we recommend – the idea that a single regression model corresponding to the design of the experiment can yield an interpretable estimate of the causal effect of interest, even in the presence of other sources of variation.

Depth

When does it makes sense to include covariates in a model?

Let's come back to one piece of advice that we gave above about making a "default" model of an experiment: not including covariates. This advice can seem surprising. Many demographic factors are of interest to psychologists and other behavioral scientists, and in observational studies these factors will almost always be related to important life outcomes. So why not put them into our experimental models? After all, we did include age in our worked example above!

Well, if you have one or at most a small handful of covariates that you believe are meaningfully related to the outcome, you *can* plan in advance to put them in your model. If you think that your effect is likely to be **moderated** a specific demographic characteristic – as we did with age in our developmental example above – then this inclusion can be quite useful.

Further, including covariates can increase the precision of your estimates by reducing "noise" in your outcome, if you hypothesize that they interact. What's surprising though is how *little* this adjustment does to increase your overall precision unless the correlation between covariate and outcome is very strong.

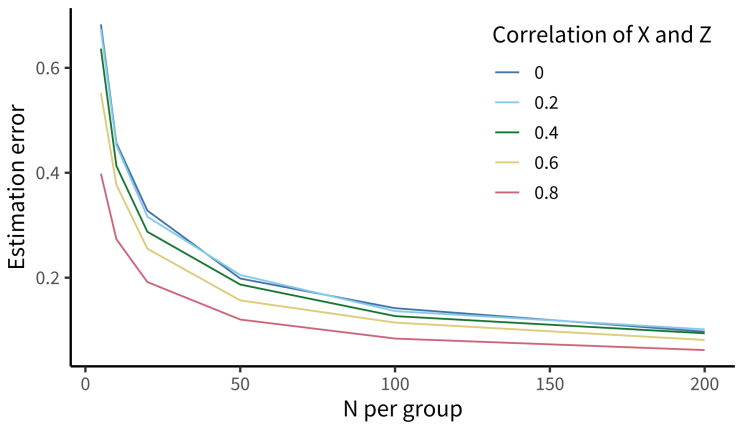


Figure 7.5: Decreases in estimation error due to adjusting for covariates, plotted by the N participants in each group and the correlation between the outcome (X) and the covariate (Z).

Figure 7.5 shows the results of a simple simulation investigating the relationship between estimation error and the inclusion of covariates. Only when the correlation between covariate and outcome (e.g., age and tea rating) is greater than $r = 0.6$ to $r = 0.8$ does this adjustment really help.

That said, there are quite a few reasons not to include covariates. These motivate our recommendation to skip them in your default model unless you have very strong theory-based expectations for either (A) a correlation with the outcome or (B) a strong moderation relationship.

The first reason not to include covariates is simply because we don't need to. Because randomization cuts causal links, our experimental estimate is an unbiased estimate of the causal effect of interest (at least for large samples). We are guaranteed that, in the limit of many different experiments, even though people with different ages will be in the different tea tasting conditions, this source of variation will be averaged out. Actually, including unnecessary covariates into models (slightly) decreases the probability that the model can detect a true effect (that is, it decreases statistical precision and power). Just by chance, covariates can "soak up" variation in the outcome, leaving less to be accounted for by the true effect!

The second reason is that you can actually compromise your causal inference by including some covariates, particularly those that are collected *after* randomization. The logic of randomization is that you cut all causal links between features of the sample and the condition manipulation. But you can "uncut" these links by accident by adding variables into your model that are related to group status. This problem is generically called **conditioning on post-treatment variables** and a full discussion of is out of the scope of this book, but it's something to avoid (and read up on if you're worried about it, see Montgomery, Nyhan, and Torres 2018).

Finally, one of the standard justifications for adding covariates – because your groups are unbalanced – is actually ill-founded as well. People often talk about "unhappy randomization": you randomize to the different tea-tasting groups, for example, but then it turns out the mean age is a bit different between groups. Then you do a t-test or some other statistical test and find out that you actually have a significant age difference. This practice makes no sense! Because you randomized, you know that the difference in ages occurred by chance. Further, incidental demographic differences between groups are unlikely to be important unless that characteristic is highly correlated with the outcome (see above). Instead, if the sample size is small enough that meaningfully large incidental differences could arise in important confounders, then it is preferable to **stratify** on that confounder at the outset – we'll have lot more to say about this issue in Chapter 10.

So these are our options: if a covariate is known to be very strongly related to our outcome, we can include it in our default model. Otherwise, we avoid a lot of trouble by leaving covariates out.

7.4.1 Robustness checks and the multiverse

Using the NHST statistical testing approach that has been common in the psychology literature, even a simple two factor experimental design affords a host of different t-tests and ANOVAs,[31] offering many opportunities for p-hacking and selective reporting. We've been advocating here instead for a "default model" approach in which you pre-plan and pre-register a single regression model that captures the planned features of your experimental design including manipulations and sources of clustering. This approach can help you to navigate some of the complexity of data analysis by having a standard approach that you take in almost every case.

Not every dataset will be amenable to this approach, however. For complex experimental designs or unusual measures, sometimes it can be hard to figure out how to specify or fit the default saturated model. And especially in these cases, the choice of model can make a big difference to the magnitude of the reported effect. To quantify variability in effect size due to model choice, "Many Analysts" projects have asked a set of teams to approach a dataset using different analysis methods. The result from these projects has been that there is substantial variability in outcomes depending on what approach is taken (Silberzahn et al. 2018; Botvinik-Nezer et al. 2020).[32]

[32] To be fair, often the analytic questions being investigated in "Many Analysts" projects are more complex than the simple experiments we recommend doing, and there is debate about how much true variability these investigations reveal (Breznau et al. 2022; Mathur, Covington, and VanderWeele 2023).

Robustness analysis (also sometimes called "sensitivity analysis" or "multiverse analysis", which sounds cooler) is a technique for addressing the possibility that an individual analysis over- or under-estimates a particular effect by chance (Steegen et al. 2016). The general idea is that analysts explore a space of different possible analyses. In its simplest form, alternative model specifications can be reported in a supplement; more sophisticated versions of the idea call for averaging estimates across a range of possible specifications and reporting this average as the primary effect estimate.

The details of this kind of analysis will vary depending on what you are worried about your model being sensitive to. One analyst might be concerned about the effects of adding different covariates; another might be using a Bayesian framework and be concerned about sensitivity to particular prior values. If you get similar results across many different specifications, you can sleep better at night. The primary principle to take home is a bit of humility about our models. Any given model is likely wrong in some of its details. Investigating the sensitivity of your estimates to the details of your model specification is a good idea.

7.5 Chapter summary: Models

In the last three chapters, we have spelled out a framework for data analysis that focuses on our key experimental goal: a measurement of a particular causal effect. We began with basic techniques for estimating effects and making inferences about how these effects estimated from a sample can be generalized to a population. This chapter showed how these ideas naturally give rise to the idea of making models of data, which allow estimation of effects in more complex designs. Simple regression models, which are formally identical to other inference methods in the most basic case, can be extended with the generalized linear model as well as with mixed effects models. Finally, we ended with

some guidance on how to build a "default model" – an (often pre-registered) regression model that maps onto your experimental design and provides the primary estimate of your key causal effect.

Discussion questions

1. Choose a paper that you have read for your research and take a look at the statistical analysis. Does the reporting focus more on hypothesis testing or on estimating effect sizes?
2. We focused here on the linear model as a tool for building models, contrasting this perspective with the common "statistical testing" mindset. But – here's the mind-blowing thing – most of those statistical tests are special cases of the linear model anyway. Take a look at this extended meditation on the equivalences between tests and models: https://lindeloev.github.io/tests-as-linear/#9_teaching_materials_and_a_course_outline. If the paper you chose for question 1 used tests, could their tests be easily translated to models? How would the use of a model-based perspective change the results section of the paper?
3. Take a look at this cool visualization of hierarchical (mixed effect) models: http://mfviz.com/hierarchical-models/. In your own research, what are the most common units that group together your observations?

Readings

– An opinionated practical guide to regression modeling and data description: Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press. Free online at https://avehtari.github.io/ROS-Examples/.

– A more in-depth introduction to the process of developing Bayesian models of data that allow for estimation and inference in complex datasets: McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC. Free materials available at https://xcelab.net/rm/statistical-rethinking/.

Part III

Planning

References

Barr, Dale J, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. "Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal." *Journal of Memory and Language* 68 (3): 255–78.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2014. "Fitting Linear Mixed-Effects Models Using Lme4." *arXiv Preprint*. https://doi.org/https://doi.org/10.48550/arXiv.1406.5823.

Bie, Ruofan, Sebastien Haneuse, Nathan Huey, Jonathan Schildcrout, and Glen McGee. 2021. "Fitting Marginal Models in Small Samples: A Simulation Study of Marginalized Multilevel Models and Generalized Estimating Equations." *Statistics in Medicine* 40 (24): 5298–5312.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, et al. 2020. "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams." *Nature* 582 (7810): 84–88.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung HV Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, et al. 2022. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty." *Proceedings of the National Academy of Sciences* 119 (44): e2203150119.

Davis, Matthew J. 2010. "Contrast Coding in Multiple Regression Analysis: Strengths, Weaknesses, and Utility of Popular Coding Structures." *Journal of Data Science* 8 (1): 61–73.

Galton, Francis. 1877. *Typical Laws of Heredity*. Royal Institution of Great Britain.

Mancl, Lloyd A, and Timothy A DeRouen. 2001. "A Covariance Estimator for GEE with Improved Small-Sample Properties." *Biometrics* 57 (1): 126–34.

Mathur, Maya B, Christian Covington, and Tyler J VanderWeele. 2023. "Variation Across Analysts in Statistical Significance, yet Consistently Small Effect Sizes." *Proceedings of the National Academy of Sciences* 120 (3): e2218957120.

McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in r and Stan*. Chapman; Hall/CRC.

Montgomery, Jacob M, Brendan Nyhan, and Michelle Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62 (3): 760–75.

Silberzahn, Raphael, Eric L Uhlmann, Daniel P Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahnı□k, et al. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1 (3): 337–56.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11 (5): 702–12. https://doi.org/10.1177/1745691616658637.

Stiller, Alex J, Noah D Goodman, and Michael C Frank. 2015. "Ad-Hoc Implicature in Preschool Children." *Language Learning and Development* 11 (2): 176–90.

8 Measurement

Learning goals

- Discuss the reliability and validity of psychological measures
- Reason about tradeoffs between different measures and measure types
- Identify the characteristics of well-constructed survey questions
- Articulate risks of measurement flexibility and the costs and benefits of multiple measures

In the previous section of the book, we described a set of measurement-focused statistical techniques for quantifying (and maximizing) our precision. In this next set of three chapters focusing on planning experiments, we will develop our toolkit for designing the measures (this chapter), design manipulations (Chapter 9), and sampling (Chapter 10) strategies that will allow us to create and evaluate experiments. These chapters form a core part of our approach to "experimentology": a set of decisions to REDUCE BIAS, maximize MEASUREMENT PRECISION, and assess GENERALIZABILITY. Let's begin with measurement.

Throughout the history of science, advances in measurement have gone hand in hand with advances in knowledge.[1] Telescopes revolutionized astronomy, microscopes revolutionized biology, and patch clamping revolutionized physiology. But measurement isn't easy. Even the humble thermometer, allowing reliable measurement of temperature, required centuries of painstaking effort to perfect (Chang 2004). Psychology and the behavioral sciences are no different – we need reliable instruments to measure the things we care about. In this next section of the book, we're going to discuss the challenges of measurement in psychology, and the properties that distinguish good instruments from bad.

What does it mean to measure something? Intuitively, we know that a ruler measures the quantity of length, and a scale measures the quantity of mass (Kisch 1965). As we discussed in Chapter 2, those quantities are **latent** (unobserved). Individual measurements, in contrast, are **manifest**: they are observable to us. What does it mean to measure a psychological construct – a hypothesized theoretical quantity inside the head?

[1] As such, measurement is a perennially controversial topic in philosophy of science. For an overview of competing frameworks, see Tal (2020) or Maul, Irribarra, and Wilson (2016), which focuses specifically on measurement in psychology.

We first have to keep in mind that not every measure is equally precise. This point is obvious when you think about physical measurement instruments: a caliper will give you a much more precise estimate of thickness than a ruler will. One way to see that the measurement is more precise is by repeating it a bunch of times. The measurements from the caliper will likely be more similar to one another, reflecting the fact that the amount of error in each individual measurement is smaller. We can do the same thing with a psychological measurement – repeat and assess variation – though as we'll see below it's a little trickier. Measurement instruments that have less error are called more **reliable** instruments.[2]

Second, psychological measurements do not directly reflect latent theoretical constructs of interest, quantities like happiness, intelligence, or language processing ability. And unlike quantities like length and mass, there is often disagreement in psychology about what the right theoretical quantities are. Thus, we have to measure an observable behavior – our operationalization of the construct – and then make an argument about how the measure relates to a proposed construct of interest (and sometimes whether the construct really exists at all!). This argument is about the **validity** of our measurements.[3]

These two concepts, reliability and validity, provide a conceptual toolkit for assessing a psychological measurement and how well it serves the researcher's goal.

8.1 Reliability

Reliability is a way of describing the extent to which a measure yields signal relative to noise. Intuitively, if there's less noise, then there will be more similarity between different measurements of the same quantity, illustrated in Figure 8.1 as a tighter grouping of points on the bulls-eye. But how do we measure signal and noise?

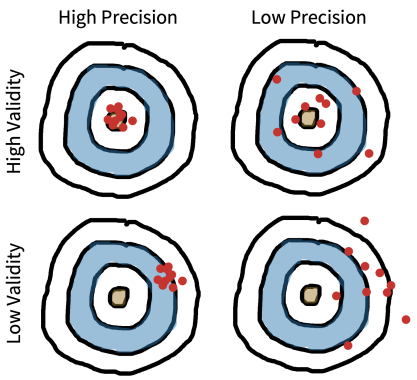


Figure 8.1: Reliability and validity visualized. The reliability of an instrument is its expected precision. The bias of measurements from an instrument also provide a metaphor for its validity.

[2] Is reliability the same as **precision**? Yes, more or less. Confusingly, different fields call these concepts different things (there's a helpful table of these names in Brandmaier et al. 2018). Here we'll talk about reliability as a property of instruments specifically while using the term precision to talk about the measurements themselves.

[3] We are also going to talk in Chapter 9 about the validity of manipulations. The way you identify a causal effect on some measure is by operationalizing some construct as well. To identify causal effects, we must link a particular construct of interest to something we can concretely manipulate in an experiment, like the stimuli or instructions.

Case study

A reliable and valid measure of children's vocabulary

Anyone who has worked with little children, or had children of their own, can attest to how variable their early language is. Some children speak clearly and produce long sentences from an early age, while others struggle; this variation appears to be linked to later school outcomes (Marchman and Fernald 2008). Thus, there are many reasons why you'd want to make precise measurements of children's early language ability as a latent construct of interest.

Because bringing children into a lab can be expensive, one popular alternative option for measuring child language is the MacArthur Bates Communicative Development Inventory (CDI for short), a form which asks parents to mark words that their child says or understands. CDI forms are basically long checklists of words. But is parent report a reliable or valid measure of children's early language?

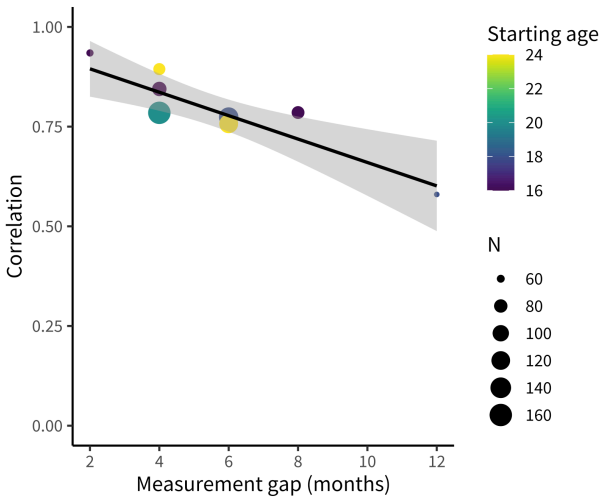


Figure 8.2: Longitudinal (test-retest) correlations between a child's score on one administration of the CDI and another one several months later. Based on Frank et al. (2021).

As we'll see below, one way to measure the reliability of the CDI to compute the correlation between two different administrations of the form for the same child. Unfortunately, this analysis has one issue: the longer you wait between observations the more the child has changed! Figure 8.2 displays these correlations for two CDIs, showing how correlations start off high and drop off as the gap between observations increases (Frank et al. 2021).

Given that CDI forms are relatively reliable instruments, are they valid? That is, do they really measure the construct of interest, namely children's early language ability? Bornstein and Haynes (1998) collected many different measures of children's language – including the ELI (an early CDI form) and other "gold standard" measures like transcribed samples of children's speech. CDI scores were highly correlated with all the different measures, suggesting that the CDI was a valid measure of the construct.

The combination of reliability and validity evidence suggests that CDIs are a useful (and relatively inexpensive source) of data about children's early language, and indeed they have become one of the most common assessments for this age group!

8.1.1 Measurement scales

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

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

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).

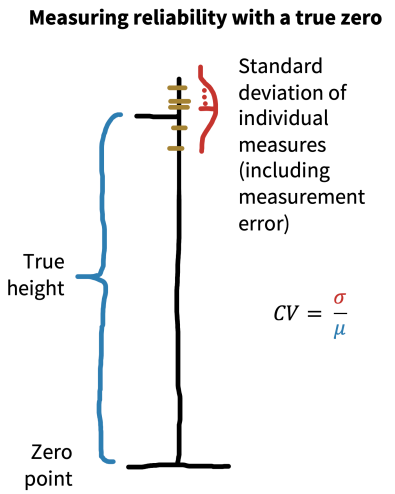


Figure 8.3: Computing the coefficient of variation (CV).

Imagine we measure the height of a person five times, resulting in measurements of 171cm, 172cm, 171cm, 173cm, and 172cm. 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 0 is meaningless – an IQ of 0 does not correspond to any particular interpretation.

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' 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.

Table 8.1: Scale types and their associated operations and statistics (Stevens 1946).

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

Critically, different summary measures work for each scale type. If you have an unordered list like a list of options for a question about race on a survey, you can present the modal response (the most likely one). 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.[4]

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 4 points, then we could estimate the precision of their measurements as "CV" of 4/100 = .04. But since IQ of 0 isn't meaningful, we could just set the mean IQ for the population to 200. Our test would be the same, and so the CV would be 4/200 = .02. On that logic we just doubled the precision of our measurements by rescaling the test! That doesn't make any sense.

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).

[4] 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"?

Depth

Early controversies over psychological measurement

> "Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of […] measurement" (Cattel 1890).

It is no coincidence that the founders of experimental psychology were obsessed with measurement (Heidelberger 2004). It was viewed as the primary obstacle facing psychology on its road to becoming a legitimate quantitative science. For example, one of the final pieces written by Hermann von Helmholtz (Wilhelm Wundt's doctoral advisor), was a 1887 philosophical treatise entitled "Zahlen und Messen" ("Counting and Measuring," see Darrigol 2003). In the same year, Fechner (1987) explicitly grappled with the foundations of measurement in "Uber die psychischen Massprincipien" ("On Psychic Measurement Principles").

Many of the early debates over measurement revolved around the emerging area of *psychophysics*, the problem of relating objective, physical stimuli (e.g. light, sound, pressure) to the subjective sensations they produce in the mind. For example, Fechner (1860) was interested in a quantity called the "just noticeable difference", the smallest change in a stimulus that can be discriminated by our senses. He argued for a lawful (logarithmic) relationship: a logarithmic change in the intensity of, say, brightness corresponded to a linear change in the intensity people reported (up to some constant). In other words, sensation was *measurable* via instruments like just noticeable difference.

It may be surprising to modern ears that the basic claim of measurability was controversial, even if the precise form of the psychophysical function would continue to be debated. But this claim led to a deeply rancorous debate, culminating with the so-called Ferguson Committee, formed by the British Association for the Advancement of Science in 1932 to investigate whether such psychophysical procedures could count as quantitative 'measurements'

of anything at all (Moscati 2018). It was unable to reach a conclusion, with physicists and psychologists deadlocked:

> Having found that individual sensations have an order, they [some psychologists] assume that they are *measurable*. Having travestied physical measurement in order to justify that assumption, they assume that their sensation intensities will be related to stimuli by numerical laws […] which, if they mean anything, are certainly false. (Ferguson et al. 1940)

The heart of the disagreement was rooted in the classical definition of quantity requiring strictly *additive* structure. An attribute was only considered measurable in light of a meaningful concatenation operation. For example, weight was a measurable attribute because putting a bag of three rocks on a scale yields the same number as putting each of the three rocks on separate scales and then summing up those numbers (in philosophy of science, attributes with this concatenation property are known as "extensive" attributes, as opposed to "intensive" ones). Norman Campbell, one of the most prominent members of the Ferguson Committee, had recently defined *fundamental* measurement in this way (e.g., Campbell 1928), contrasting it with *derived measurement*, which involved computing some function based on one or more fundamental measures. According to the physicists on the Ferguson Committee, measuring mental sensations was impossible because they could never be grounded in any *fundamental* scale with this kind of additive operation. It just didn't make sense to break up people's sensations into parts the way we would weights or lengths: they didn't come in "amounts" or "quantities" that could be combined (Cattell 1962). Even the intuitive additive logic of Donders (1868/1969)'s "method of subtraction" for measuring the speed of mental processes was viewed skeptically on the same grounds by the time of the committee (e.g., in an early textbook, Woodworth (1938) claimed "we cannot break up the reaction into successive acts and obtain the time for each act.")

The primary target of the Ferguson Committee's investigation was the psychologist S. S. Stevens, who had claimed to measure the sensation of loudness using psychophysical instruments. Exiled from classical frameworks of measurement, he went about developing an alternative "operational" framework (Stevens 1946), where the classical ratio scale recognized by physicists was only one of several ways of assigning numbers to things (see Table 8.1 above). Stevens' framework quickly spread, leading to an explosion of proposed measures. However, operationalism remains controversial outside psychology (Michell 1999). The most extreme version of Steven's stance ("measurement is the assignment of numerals to objects or events according to rule") permits researchers to *define* constructs operationally in terms of a measure (Hardcastle 1995). For example, one may say that the construct of intelligence is simply *whatever it is* that IQ measures. It is then left up to the researcher to decide which scale type their proposed measure should belong to.

In Chapter 2, we outlined a somewhat different view, closer to a kind of constructive realism (Giere 2004; Putnam 2000). Psychological constructs like happiness are taken to exist independent of any given operationalization, putting us on firmer ground to debate the pros and cons associated with different ways of measuring the same construct. In other words, we are not free to assign numbers however we like. Whether a particular construct or quantity *is measurable* on a particular scale should be treated as an empirical question.

The next major breakthrough in measurement theory emerged with the birth of mathematical psychology in the 1960s, which aimed to put psychological measurement on more rigorous foundations. This effort culminated in the three-volume Foundations of Measurement series (Krantz et al. 1971; Suppes et al. 1989; Robert Duncan Luce et al. 1990), which has become the canonical text for every psychology student seeking to understand measurement in the non-physical sciences. One of the key breakthroughs was to shift the burden from measuring (additive) constructs themselves to measuring (additive) *effects* of constructs in conjunction with one another:

> When no natural concatenation operation exists, one should try to discover a way to measure factors and responses such that the 'effects' of different factors are additive. (R. Duncan Luce and Tukey 1964).

This modern viewpoint broadly informs the view we describe here.

8.1.1 Measuring reliability

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.[5]

[5] 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.

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).

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 sub-parts 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**.[6]

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

One final note: these tools often get used for 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 Depth box below) for continuous data as well as many other measures.

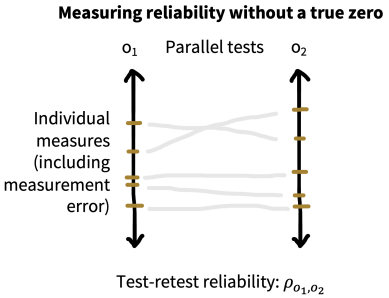


Figure 8.4: Computing test-retest reliability.

[6] 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.

Depth

Reliability paradoxes!

There's a major issue with calculating reliabilities using the approaches we described here: because reliability is defined as a ratio of two measures of variation, it will always be relative to the variation in the sample. So if a sample has less variability, reliability will decrease!

One way to define reliability formally is by using the intra-class correlation coefficient (ICC):

$$ICC = \frac{\sigma_b^2}{\sigma_w^2 + \sigma_b^2}$$

where σ_w^2 is the within-subject variance in measurements and σ_b^2 is the between-subject variance in the measure-

ments. (The denominator of the ICC comes from partitioning the total observed variance σ_o^2 in the reliability formula above).

So now instead of comparing variation to the mean, we're comparing variation on one dimension (between person) to total variation (within and between person). ICCs are tricky and there are several different flavors available depending on the structure of your data and what you're trying to do with them. McGraw and Wong (1996) and Gwet (2014) provide extensive guidance on how to compute and interpret this statistic in different situations.

Let's think about the CDI data in our case study, which showed high reliability. Now imagine we restricted our sample to only change scores between 16 – 18-month-olds (our prior sample had 16 – 30-month-olds). Within this more restricted subset, overall vocabularies would be lower and more similar to one another, and so the average amount of change *within* a child (σ_w) would be larger relative to the differences *between* children (σ_b). That would make our reliability go *down*, even though we would be computing it on a subset of the exact same data.

That doesn't sound so bad. But we can construct a much more worrisome version of the same problem. Say we are very sloppy in our administration of the CDI and create lots of between-participants variability, perhaps by giving different instructions to different families. This practice will actually *increase* our estimate of split-half reliability (by increasing σ_b). While the within-participant variability will remain the same, the between-participant variability will go up! You could call this a "reliability paradox" – sloppier data collection can actually lead to higher reliabilities.

We need to be sensitive to the sources of variability we're quantifying reliability over – both the numerator and the denominator. If we're computing split-half reliabilities, typically we're looking at variability across test questions (from some question bank) vs. across individuals (from some population). Both of these sampling decisions affect reliability – if the population is more variable *or* the questions are less variable, we'll get higher reliability. In sum, *reliability is relative*: reliability measures depend on the circumstances in which they are computed.

8.1.1 Practical advice for computing reliability

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.[7] It's hard work to measure test-retest 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).

Accident report

Wasted effort

Low-reliability measures limit your ability to detect correlations between measurements. Mike spent several fruitless months in graduate school running dozens of participants through batteries of language processing tasks and correlating the results across tasks. Every time data collection finished, one or the other (spurious) correlation would show up in the data analysis. Something was always correlated with something else. Thankfully, he would always attempt to replicate the correlation in a new sample – and in that next dataset, the correlation we were trying to replicate would be null but another (again likely spurious) correlation would show up.

This exercise was a waste of time because most of the tasks were of such low reliability that, even had they been highly correlated with one another, relationship would have been almost impossible to detect without a huge sample size. (It also would have been helpful if someone had mentioned multiplicity corrections (Chapter 6) to him.)

One rule of thumb that's helpful for individual difference designs of this sort is that the maximal correlation that can be observed between two variables x and y is the square root of the product of their reliabilities: $\sqrt{r_x r_y}$. So if you have two measures that are reliable at .25, the maximal measured correlation between them is .25 as well! This kind of method is now frequently used in cognitive neuroscience (and other fields as well) to compute the so-called **noise ceiling** for a measure: the maximum amount of signal that in principle *could* be predicted (Lage-Castellanos et al. 2019). If your sample size is too small to detect correlations at the noise ceiling (see Chapter 10), then the study is not worth doing.

8.2 Validity

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.[8]

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.[9] 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 that 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 pre-theoretic intuitions rather than any quantitative assessment. For example, reaction time is typically correlated with intelligence test results (e.g., Jensen and Munro 1979), but 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.[10]

[8] Some authors have treated "validity" as a broader notion that can include, for example, statistical issues (Shadish, Cook, and Campbell 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.

[9] 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).

[10] Often this concept is described as being relevant to the validity of a *manipulation* also, e.g. 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.

- **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, Diener, and Seidlitz 1993).[11]
- **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, Diener, and Suh 1996).

[11] 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 above Depth Box on the history of measurement). We don't ever have an objective starting point for the study of the human mind.

8.2.1 Validity arguments in practice

Let's take a look at how we might make an argument about the validity of the CDI, the vocabulary instrument from our case study.

First, the CDI is face valid – it is clearly about early language ability. In contrast, even though a child's height would likely be correlated with their early language ability, we should be skeptical of this measure due to its lack of face validity. In addition, the CDI shows good convergent and predictive validity. Concurrently, the CDI correlates well with evidence from transcripts of children's actual speech and from standardized language assessments (as discussed in the case study above). And predictively, CDI scores at age 2 relate to reading scores during elementary school (Marchman and Fernald 2008).

On the other hand, users of the CDI must avoid challenges to the internal validity of the data they collect. For example, some CDI data are compromised by confusing instructions or poor data collection processes (Frank et al. 2021). Further, advocates and critics of the CDI argue about its ecological validity. There is something quite ecologically valid about asking parents and caregivers – who are experts on their

own child – to report on their child's abilities. On the other hand, the actual experience of filling out a structured form estimating language ability might be more familiar to some families from high education backgrounds than for others from lower education backgrounds. Thus, a critic could reasonably say that comparisons of CDI scores across socioeconomic strata would be an invalid usage (Feldman et al. 2000).

8.2.2 Avoid questionable measurement practices!

Experimentalists sometimes have a tendency to make up ad hoc measures on the fly. It's fine to invent new measures, but the next step is to think about what evidence there is that it's valid! Table 8.2 gives a set of questions to guide thoughtful reporting of measurement practices (adapted from Flake and Fried 2020).

Table 8.2: Questions about measurement that every reseacher should answer in their paper. Adapted from Flake and Fried (2020).

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 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.

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 instrument can sometimes be justified by features of 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.

Accident report

Talk about flexible measurement!

The Competitive Reaction Time Task (CRTT) is a lab-based measure of aggression. Participants are told that they are playing a reaction-time game against another player and are asked to set the parameters of a noise blast that will be played to their opponent. Unfortunately, in an analysis of the literature using CRTT, Elson et al. (2014) found that different papers using the CRTT use dramatically different methods for scoring the task. Sometimes the analysis focused on the volume of the noise blast and sometimes it focused on the duration. Sometimes these scores were transformed (via logarithms) or thresholded. Sometimes they were combined into a single score. Elson was so worried by this flexibility, he created a website, https://flexiblemeasures.com, to document the variation he observed.

As of 2016, Elson had found 130 papers using the CRTT. And across these papers, he documented an astonishing 157 quantification strategies. One paper reported ten different strategies for extracting numbers from this measure! More worrisome still, Elson and colleagues found that when they tried out some of these strategies on their own data, different strategies led to very different effect sizes and levels of statistical significance. They could effectively make a finding appear bigger or smaller depending on which scoring they chose.

Triangulating a construct through multiple pre-specified measurements can be a good thing. But the issue with the CRTT analysis was that changes in the measurement strategy appeared to be made in a *post hoc*, data-driven way so as to maximize the significance of the experimental manipulation (just like the *p*-hacking we discussed in Chapters 3 and 6).

This examination of the use of the CRTT measure has several implications. First, and most troublingly, there may have been undisclosed flexibility in the analysis of CRTT data across the literature, with investigators taking advantage of the lack of standardization to try many different analysis variants and report the one most favorable to their own hypothesis. Second, it is unknown which quantification of CRTT behavior is in fact most reliable and valid. Since some of these variants are presumably better than others, researchers are effectively "leaving money on the table" by using suboptimal quantifications. As a consequence, if researchers adopt the CRTT, they find much less guidance from the literature on what quantification to adopt.

8.3 How to select a good measure?

Ideally you want a measure that is reliable and valid. How do you get one? An important first principle is to use a pre-existing 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, you also have an important index of population variability by comparing absolute scores to other reports.[12]

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

[12] 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.

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.[13] 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.

[13] 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, Goodman, and Frank 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, e.g.:

- 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.

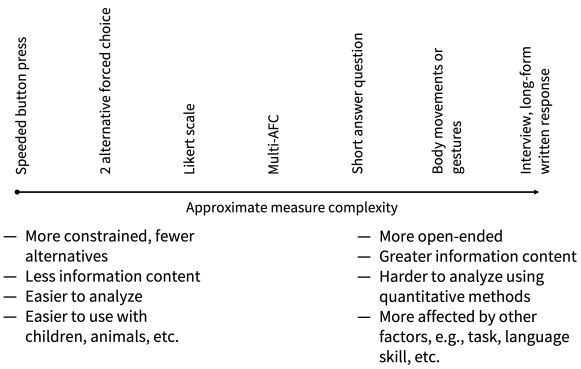


Figure 8.5: Often choosing a measure can be consolidated into a choice along a continuum from simple measures that provide a small amount of information but are quick and easy to repeat and those that provide much richer information but require more time.

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 non-human 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, e.g. 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 – e.g., 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 set of experiments. 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.[14] 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.

[14] 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 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.

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 – e.g., the answer to a question. In contrast, implicit measures provide measurements of psychological processes that participants are unable to report (or occasionally, unwilling to).[15] Implicit measures, especially reaction time, have long been argued to reflect internal psychological processes (Donders 1868/1969). They also have been proposed as measures of qualities such as racial bias that participants may have motivation not to disclose (Greenwald, McGhee, and Schwartz 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, MEG, NIRS, and EEG 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, Xu, and Milham 2019).

[15] 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.

Many tasks produce both accuracy and reaction time data. Often these trade off with one another in a classic **speed-accuracy tradeoff**: 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 either 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 manipulation 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 tradeoff, yielding effects on different measures.[16]

[16] One way of describing the information processing underlying this tradeoff is given by drift diffusion models, which allow joint analysis of accuracy and reaction time (Voss, Nagler, and Lerche 2013). Used appropriately, drift diffusion models can provide a way to remove speed-accuracy tradeoffs and extract more reliable signals from tasks where accuracy and reaction time are both measured (see Johnson et al. 2017 for an example of 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

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, reliable and valid measure.[17] If you are tempted to include more than one measure, see if we can talk you out of it.[18]

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.[19]

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.

[17] 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.

[18] 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, Mathur, and Chen 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.

[19] 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.

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, etc. Further, such trials are extremely expensive and difficult to repeat. Thus, there is a good reason for including more measures in such studies.

Depth

Survey measures

Sometimes the easiest way to elicit information from participants is simply to ask. Surveys are an important part of experimental measurement, so we'll share a few best practices, primarily derived from Krosnick and Presser (2010).

Treat survey questions as a conversation. The easier your items are to understand, the better. Don't repeat variations on the same question unless you want different answers! Try to make the order reasonable, for example by grouping together questions about the same topic and moving from more general to more specific questions. The more you include "tricky" items the more you invite tricky answers to straightforward questions. One specific kind of tricky questions are "check" questions that evaluate participant compliance. We'll talk more in Chapter 12 about various ways of evaluating compliance and their strengths and weaknesses.

Open-ended survey questions can be quite rich and informative, especially when an appropriate coding (classification) scheme is developed in advance and responses are categorized into a relatively small number of types. On the other hand, they present practical obstacles because they require coding (often by multiple coders to ensure reliability of the coding). Further, they tend to yield nominal data, which are often less useful for quantitative theorizing. Open-ended questions are a useful tool to add nuance and color to the interpretation of an experiment.

One common mistake that survey developers make is trying to put too much into one question. Imagine asking a restaurant-goer for a numerical ranking on the question, "How do you like our food and service?" What if they loved the food but hated the service, or vice versa – would they choose an intermediate option? Items that ask about more than one thing at once are known as **double-barreled** questions. They can confuse and frustrate participants as well as leading to uninterpretable data.

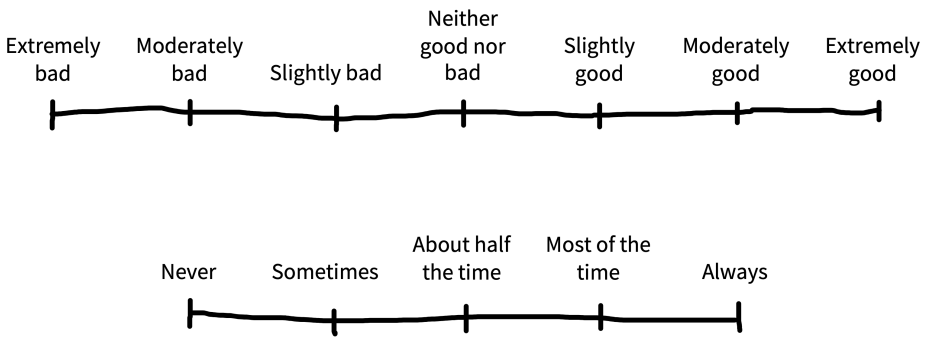


Figure 8.6: Likert scales based on survey best practices: a bipolar opinion scale with seven points and a unipolar frequency scale with five points. Both have all points labeled.

Especially given their ubiquity in commercial survey research, **Likert scales** – scales with a fixed number of ordered, numerical response options – are a simple and conventional way of gathering data on attitude and judgment questions (Figure 8.6). Bipolar scales are those in which the endpoints represent opposites, for example the continuum between "strongly dislike" and "strongly like." Unipolar scales have one neutral endpoint, like the continuum between "no pain" and "very intense pain." Survey methods research suggests that reliability is maximized when bipolar scales have seven points and unipolar scales have five. Labeling every point on the scale with verbal labels is preferable to labeling only the endpoints.

One important question is whether to treat data from Likert scales as ordinal or interval. It's extremely common (and convenient) to make the assumption that Likert ratings are interval, allowing the use of standard statistical tools like means, standard deviations, linear regression, etc. The risk in this practice comes from the possibility that scale items are not evenly spaced – for example, on a scale labeled "never","seldom", "occasionally","often","always," the distance from "often" to "always" may be larger than the distance from "seldom" to "occasionally."

In practice, you can choose to use regression variants that are appropriate, e.g. ordinal logistic regression and its variants, or they can attempt to assess and mitigate the risks of treating the data as interval. If you choose the second option, it's definitely a good idea to look carefully at the raw distributions for individual items to see if their distribution appears approximately normal (see Chapter 15).

Recently some researchers have begun to use "visual analog scales" (or sliders) as a solution. We don't recommend these – the distribution of the resulting data is often anchored at the starting point or endpoints (Matejka et al. 2016), and a meta-analysis shows they're a lot lower than Likert scales in reliability (Krosnick and Presser 2010).

It rarely helps matters to add a "don't know" or "other" option to survey questions. These are some of a variety of practices that encourage **satisficing**, where survey takers give answers that are good enough but don't reflect substantial thought about the question. Another behavior that results from satisficing is "straight-lining" – that is, picking the same option for every question. In general, the best way to prevent straight-lining is to make surveys relatively short, engaging, and well-compensated. The practice of "reverse coding" to make the expected answers to some questions more negative can block straight-lining, but at the cost of making items more confusing. Some obvious formatting options can reduce straight-lining as well, for example placing scales further apart or on subsequent (web) pages.

In sum, survey questions can be a helpful tool for eliciting graded judgments about explicit questions. The best way to execute them well is to try and make them as clear and easy to answer as possible.

8.5 Chapter summary: Measurement

In olden times, all the psychologists went to the same conferences and worried about the same things. But then a split formed between different groups. Educational psychologists and psychometricians thought a lot about how different problems on tests had different measurement properties. They began exploring how to select good and bad items, and how to figure out people's ability abstracted away from specific items. This research led to a profusion of interesting ideas about measurement, but these ideas rarely percolated into day-to-day practice in other areas of psychology. For example, cognitive psychologists collected lots of trials and measured quantities of interest with high precision, but worried less about measurement validity. Social psychologists spent more

time worrying about issues of ecological validity in their experiments, but often used *ad hoc* scales with poor psychometric properties.

These sociological differences between fields has led to an unfortunate divergence, where experimentalists often don't recognize the value of the conceptual tools developed to aid measurement, and so fail to reason about the reliability and validity of their measures in ways that can help them make better inferences. As we said in our discussion of reliability, ignorance is not bliss. Much better to think these choices through!

Discussion questions

1. Let's go back to our example on the relationship between money and happiness from Chapter 1. How many different kinds of measures of happiness can you come up with? Make a list with at least five.
2. Choose one of your measures of happiness and come up with a validation strategy for it, making reference to at least three different types of validity. What data collection would this validation effort require?

Readings

- A classic textbook on psychometrics that introduces the concepts of reliability and validity in a simple and readable way: Furr, R. M. (2021). *Psychometrics: an introduction*. SAGE publications.
- A great primer on questionnaire design: Krosnick, J.A. (2018). Improving Question Design to Maximize Reliability and Validity. In: Vannette, D., Krosnick, J. (eds) The Palgrave Handbook of Survey Research. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-54395-6_13.
- Introduction to general issues in measurement and why they shouldn't be ignored: Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. Advances in Methods and Practices in Psychological Science, 3(4), 456-465. https://doi.org/10.1177/2515245920952393.
- An accessible popular book on scientific measurement: Vincent, J. (2022). *Beyond Measure: The Hidden History of Measurement from Cubits to Quantum Constants*. W. W. Norton.

References

Bornstein, Marc H, and O Maurice Haynes. 1998. "Vocabulary Competence in Early Childhood: Measurement, Latent Construct, and Predictive Validity." *Child Development* 69 (3): 654–71.

Brandmaier, Andreas M, Elisabeth Wenger, Nils C Bodammer, Simone Kühn, Naftali Raz, and Ulman Lindenberger. 2018. "Assessing Reliability in Neuroimaging Research Through Intra-Class Effect Decomposition (ICED)." *Elife* 7: e35718.

Campbell, Norman Robert. 1928. *An Account of the Principles of Measurement and Calculation*. Longmans, Green; Company, Limited.

Cattel, J McK. 1890. "Mental Tests and Measurements." *Mind* 15: 373–80.

Cattell, Raymond B. 1962. "The Relational Simplex Theory of Equal Interval and Absolute Scaling." *Acta Psychologica* 20: 139–58.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.
Cohen, Jacob. 1990. "Things i Have Learned (so Far)." *American Psychologist* 45: 1304–12.
Cronbach, L J, and P E Meehl. 1955. "Construct Validity in Psychological Tests." *Psychol. Bull.* 52 (4): 281–302.
Darrigol, Olivier. 2003. "Number and Measure: Hermann von Helmholtz at the Crossroads of Mathematics, Physics, and Psychology." *Studies in History and Philosophy of Science Part A* 34 (3): 515–73.
Donders, Franciscus Cornelis. 1868/1969. "On the Speed of Mental Processes." Translated by W G Koster. *Acta Psychologica* 30 (1868/1969): 412–31. https://doi.org/https://doi.org/10.1016/0001-6918(69)90065-1.
Elson, Malte, M. Rohangis Mohseni, Johannes Breuer, Michael Scharkow, and Thorsten Quandt. 2014. "Press CRTT to Measure Aggressive Behavior: The Unstandardized Use of the Competitive Reaction Time Task in Aggression Research." *Psychological Assessment* 26 (2): 419–32. https://doi.org/10.1037/a0035569.
Embretson, Susan E, and Steven P Reise. 2013. *Item Response Theory*. Psychology Press.
Fechner, Gustav Theodor. 1860. *Elemente Der Psychophysik*. Vol. 2. Breitkopf u. Härtel.
———. 1987. "My Own Viewpoint on Mental Measurement (1887)." *Psychological Research* 49 (4): 213–19.
Feldman, Heidi M, Christine A Dollaghan, Thomas F Campbell, Marcia Kurs-Lasky, Janine E Janosky, and Jack L Paradise. 2000. "Measurement Properties of the MacArthur Communicative Development Inventories at Ages One and Two Years." *Child Development* 71 (2): 310–22.
Ferguson, A., C. S. Myers, R. J. Bartlett, H. Banister, F. C. Bartlett, Brown W., and W. S. Tucker. 1940. "Quantitative Estimates of Sensory Events, Final Report." *Report of the British Association for the Advancement of Science*, 331–49.
Flake, Jessica Kay, and Eiko I Fried. 2020. "Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them." *Advances in Methods and Practices in Psychological Science* 3 (4): 456–65.
Frank, Michael C, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. 2021. *Variability and Consistency in Early Language Learning: The Wordbank Project*. MIT Press.
Furr, R Michael. 2021. *Psychometrics: An Introduction*. SAGE publications.
Giere, Ronald N. 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71 (5): 742–52. https://doi.org/https://doi.org/10.1086/425063.
Greenwald, Anthony G, Debbie E McGhee, and Jordan LK Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74 (6): 1464.
Gwet, Kilem L. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC.
Hardcastle, Gary L. 1995. "SS Stevens and the Origins of Operationism." *Philosophy of Science*, 404–24.
Heidelberger, Michael. 2004. *Nature from Within: Gustav Theodor Fechner and His Psychophysical Worldview*. University of Pittsburgh Pre.
Herzog, A Regula, and Jerald G Bachman. 1981. "Effects of Questionnaire Length on Response Quality." *Public Opinion Quarterly* 45 (4): 549–59.
Jensen, Arthur R, and Ella Munro. 1979. "Reaction Time, Movement Time, and Intelligence." *Intelligence* 3 (2): 121–26.
Johnson, David J, Christopher J Hopwood, Joseph Cesario, and Timothy J Pleskac. 2017. "Advancing Research on Cognitive Processes in Social and Personality Psychology: A Hierarchical Drift Diffusion Model Primer." *Social Psychological and Personality Science* 8 (4): 413–23.
Kane, Michael T. 1992. "An Argument-Based Approach to Validity." *Psychological Bulletin* 112 (3): 527.
Kisch, B. 1965. *Scales and Weights: A Historical Outline*. Yale Studies in the History of Science and Medicine. Yale University Press.
Krantz, David H, R Duncan Luce, Patrick Suppes, and Amos Tversky. 1971. *Foundations of Measurement i: Additive and Polynomial Representations*. Courier Corporation.
Krosnick, Jon A, and Stanley Presser. 2010. "Question and Questionnaire Design." *Handbook of Survey Research*, 263.
Lage-Castellanos, Agustin, Giancarlo Valente, Elia Formisano, and Federico De Martino. 2019. "Methods for Computing the Maximum Performance of Computational Models of fMRI Responses." *PLoS Computational Biology* 15 (3): e1006397.
Lieberman, Joel D, Sheldon Solomon, Jeff Greenberg, and Holly A McGregor. 1999. "A Hot New Way to Measure Aggression: Hot Sauce Allocation." *Aggressive Behavior: Official Journal of the International Society for Research on*

Aggression 25 (5): 331–48.
Lucas, Richard E, Ed Diener, and Eunkook Suh. 1996. "Discriminant Validity of Well-Being Measures." *Journal of Personality and Social Psychology* 71 (3): 616.
Luce, R Duncan, and John W Tukey. 1964. "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement." *Journal of Mathematical Psychology* 1 (1): 1–27.
Luce, Robert Duncan, David H Krantz, Patrick Suppes, and Amos Tversky. 1990. *Foundations of Measurement III: Representation, Axiomatization, and Invariance*. Courier Corporation.
Marchman, Virginia A, and Anne Fernald. 2008. "Speed of Word Recognition and Vocabulary Knowledge in Infancy Predict Cognitive and Language Outcomes in Later Childhood." *Developmental Science* 11 (3): F9–16.
Matejka, Justin, Michael Glueck, Tovi Grossman, and George Fitzmaurice. 2016. "The Effect of Visual Appearance on the Performance of Continuous Sliders and Visual Analogue Scales." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5421–32.
Maul, Andrew, David Torres Irribarra, and Mark Wilson. 2016. "On the Philosophical Foundations of Psychological Measurement." *Measurement* 79: 311–20.
McGraw, Kenneth O, and Seok P Wong. 1996. "Forming Inferences about Some Intraclass Correlation Coefficients." *Psychological Methods* 1 (1): 30.
Michell, Joel. 1999. *Measurement in Psychology: A Critical History of a Methodological Concept*. Vol. 53. Cambridge University Press.
Moffitt, Terrie E, Louise Arseneault, Daniel Belsky, Nigel Dickson, Robert J Hancox, HonaLee Harrington, Renate Houts, et al. 2011. "A Gradient of Childhood Self-Control Predicts Health, Wealth, and Public Safety." *Proceedings of the National Academy of Sciences* 108 (7): 2693–98.
Moscati, Ivan. 2018. *Measuring Utility: From the Marginal Revolution to Behavioral Economics*. Oxford University Press.
Narens, Louis, and R Duncan Luce. 1986. "Measurement: The Theory of Numerical Assignments." *Psychological Bulletin* 99 (2): 166.
Payne, B Keith. 2001. "Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon." *Journal of Personality and Social Psychology* 81 (2): 181.
Putnam, Hilary. 2000. *The Threefold Cord: Mind, Body, and World*. Columbia Univ. Press.
Ross, Hildy S, and Susan P Lollis. 1989. "A Social Relations Analysis of Toddler Peer Relationships." *Child Development*, 1082–91.
Rumelhart, David E, and Donald A Norman. 1982. "Simulating a Skilled Typist: A Study of Skilled Cognitive-Motor Performance." *Cognitive Science* 6 (1): 1–36.
Sandvik, Ed, Ed Diener, and Larry Seidlitz. 1993. "Subjective Well-Being: The Convergence and Stability of Self-Report and Non-Self-Report Measures." *Journal of Personality* 61 (3): 317–42.
Schmuckler, Mark A. 2001. "What Is Ecological Validity? A Dimensional Analysis." *Infancy* 2 (4): 419–36.
Shadish, William, Thomas D Cook, and Donald Thomas and Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.
Shafto, Patrick, Noah D Goodman, and Michael C Frank. 2012. "Learning from Others: The Consequences of Psychological Reasoning for Human Learning." *Perspectives on Psychological Science* 7 (4): 341–51.
Sijtsma, Klaas. 2009. "On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha." *Psychometrika* 74 (1): 107.
Stevens, S S. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684): 677–80.
Suppes, Patrick, David H Krantz, Robert Duncan Luce, and Amos Tversky. 1989. *Foundations of Measurement II: Geometrical, Threshold, and Probabilistic Representations*. Courier Corporation.
Tal, Eran. 2020. "Measurement in Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2020. https://plato.stanford.edu/archives/fall2020/entries/measurement-science/; Metaphysics Research Lab, Stanford University.
VanderWeele, Tyler J, Maya B Mathur, and Ying Chen. 2020. "Outcome-Wide Longitudinal Designs for Causal Inference: A New Template for Empirical Studies." *Statistical Science* 35 (3): 437–66.
Voss, Andreas, Markus Nagler, and Veronika Lerche. 2013. "Diffusion Models in Experimental Psychology." *Experimental Psychology* 60 (6): 385–402. https://doi.org/10.1027/1618-3169/a000218.

Woodworth, R. S. 1938. *Experimental Psychology*. Holt.
Zuo, Xi-Nian, Ting Xu, and Michael Peter Milham. 2019. "Harnessing Reliability for Neuroscience Research." *Nature Human Behaviour* 3 (8): 768–71. https://doi.org/10.1038/s41562-019-0655-x.

9 Design

Learning goals

- Describe key elements of experimental design
- Define randomization and counterbalancing strategies for removing confounds
- Discuss strategies to design experiments that are appropriate to the populations of interest

The key thesis of our book is that experiments should be designed to yield precise and unbiased measurements of a causal effect. But the causal effect of what? The manipulation! In an experiment we manipulate (intervene on) some aspect of the world and measure the effects of that manipulation. We then compare that measurement to a situation where the intervention has not occurred.

We refer to different intervention states as **conditions** of the experiment. The most common experimental design is the comparison between a **control** condition, in which the intervention is not performed, and an **experimental** (sometimes called **treatment**) condition in which the intervention is performed.

But many other experimental designs are possible. In more complex experiments, manipulations along different dimensions (sometimes called **factors** in this context) can be combined. In the first part of the chapter, we'll introduce some common experimental designs and the vocabulary for describing them. Our focus here is in identifying designs that maximize MEASUREMENT PRECISION.

A good experimental measure must be a valid measure of the construct of interest. The same is true for a manipulation – it must validly relate to the causal effect of interest. In the second part of the chapter, we'll discuss issues of **manipulation validity**, including both issues of ecological validity and **confounding**. We'll talk about how practices like **randomization** and **counterbalancing** can help remove nuisance confounds, an important part of BIAS REDUCTION for experimental designs.[1]

To preview our general take-home points from this chapter: we think that your default experiment should manipulate one or two factors – usually not more – and should manipulate those factors continuously and within-participants. Although such designs are not always possible,

[1] This section will draw on our introduction to causal inference in Chapter 1, so if you haven't read that, now's the time.

they are typically the most likely to yield precise estimates of a particular effect that can be used to constrain future theorizing. We'll start by considering a case study in which a subtle confound led to difficulties interpreting an experimental result.

Case study

Automatic theory of mind?

In an early version of our course, student Desmond Ong set out to replicate a thought-provoking finding: both infants and adults seemed to show evidence of tracking other agents' belief state, even when it was irrelevant to the task at hand (Kovács, Téglás, and Endress 2010). In the paradigm, an animated Smurf character would watch as a self-propelled ball came in and out from behind a screen. At the end of the video, the screen would swing down and the participant had to respond whether the ball was present or absent. Reaction time for this decision was the key dependent variable.

The experimental design investigated two factors: whether the participant believed the ball was present or absent (P+/P-) and whether the animated agent *would have believed* the ball was present or absent (A+/A-) based on what it saw. The result was four conditions: P+/A+, P+/A-, P-/A+, and P-/A-. (We could call this a **fully crossed** design because each level of one factor was presented with each level of the other).

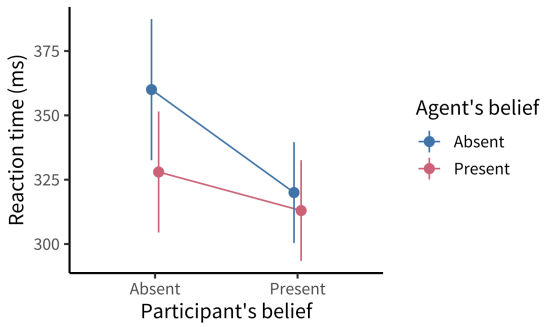


Figure 9.1: Original data from Kovács, Téglás, and Endress (2010). Error bars show 95% confidence intervals. Based on Phillips et al. (2015).

Both the original experiments and the replication that Desmond ran showed a significant effect of the agent's beliefs on participants' reaction times, suggesting that what the – totally irrelevant – agent thought about the ball was leading them to react more or less quickly to the presence of the ball. Figure 9.1 shows the original data (N=24). But although both studies showed an effect of agent belief, the replication and several variations also showed a crossover **interaction** of participant and agent belief. The participants were slower when the agents *and* the participants believed that the ball was behind the screen (Figure 9.2). That finding wasn't consistent with the theory that tracking inconsistent beliefs slowed down reaction times. If participants were tracking their own beliefs about the ball *and* the agent's, they should have been fastest in the P+/A+ condition, not slower.

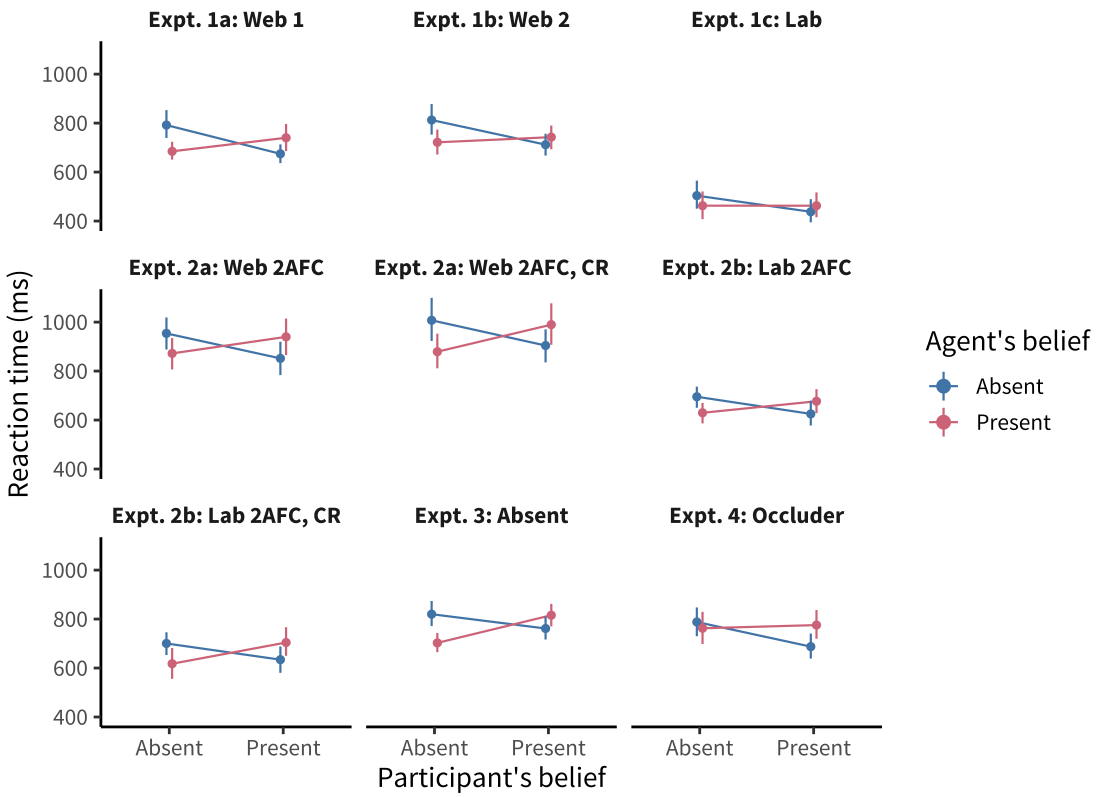


Figure 9.2: Data from a series of replications of Kovács, Téglás, and Endress (2010), including versions on the web (Experiments 1a and 1b) and in lab (Experiment 1c), as well as several variations on the format of responding (Experiments 2 and 3; 2AFC = two alternative forced choice) and an experiment where a large wall kept the agent from seeing the ball at all (Experiment 4). "Hits" and "CRs" panels refer to different subsets of trials where participants responded "present" when the ball was present and "absent" when the ball was absent. Error bars are 95% confidence intervals. Based on Phillips et al. (2015).

A collaborative team working on this paradigm identified a key issue (Phillips et al. 2015). There was a **confound** in the experimental design – another factor that varied across conditions besides the target factors. In other words, something was changing between conditions other than the agent's and participant's belief states. The confound was an attention check (discussed further in Chapter 12): participants had to press a key when the agent left the scene to show that they were paying attention. This attention check appeared a few seconds later in the videos for the P+/A+ and P-/A- trials – the ones that yielded the slow reaction times – than it did for the other two. When the attention check was removed or when its timing was equalized across conditions, reaction time effects were eliminated, suggesting that the original pattern of findings may have been due to the confound.

If the standard for replication is significance of particular statistical tests at $p<.05$, then this experiment replicated successfully. But the effect estimates were inconsistent with the proposed theoretical explanation. A finding can be replicable without providing support for the underlying theory!

There's an important caveat to this story. The followup work *only* revealed that there was a confound in one particular experimental operationalization, and did not provide evidence against automatic theory of mind in general. Indeed, others have suggested that different versions of this paradigm *do* reveal evidence for theory of mind processing once the confound is eliminated (El Kaddouri et al. 2020).

9.1 Experimental designs

Experimental designs are fundamental to many fields; unfortunately the terminology used to describe them can vary, which can get quite confusing! Here we will mostly describe an experiment as a relationship between some manipulation(s), in which participants are randomly assigned to experimental conditions to estimate effects on some measure. Factors are the dimensions along which manipulations vary. For example, in our case study above, the two factors were participant belief and agent belief. Another terminology it's good to be familiar with is the terms used in Chapters 5–7, which are often used in econometrics and statistics: **treatment** (manipulation) and **outcome** (measure).[2]

In this section, we'll discuss key dimensions on which experiments vary: 1) how many factors they incorporate and how these factors are crossed, 2) how many conditions and measures are given to each participant, and 3) if manipulations have discrete levels or fall on a continuous scale.

9.1.1 A two-factor experiment

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 θs indicate the estimates for each 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 $\beta_{T_2} = \theta_{T_2} - \theta_C$.

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

[2] Terminology here is hard. In psychology people sometimes say there's an **independent variable** (the manipulation, which is causally prior and hence "independent" of other causal influences) and a **dependent variable** (the measure, which causally depends on the manipulation, or so we hypothesize). We find this terminology to be hard to remember because the terms are so different from the actual concepts being described.

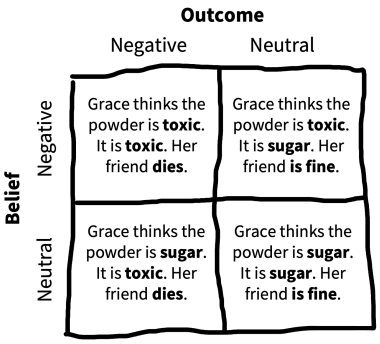


Figure 9.3: The 2x2 crossed design used in Young et al. (2007)

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.

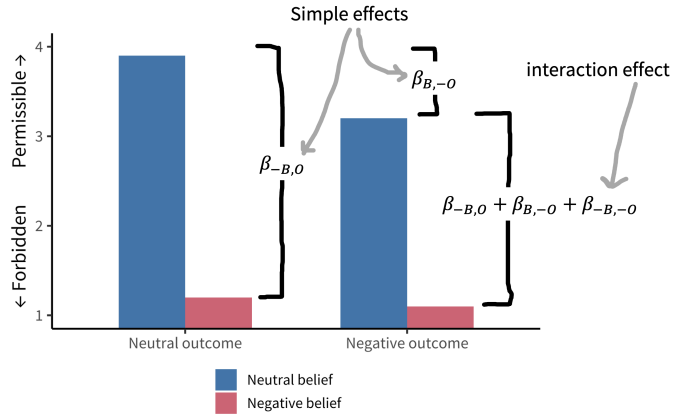


Figure 9.4: Moral permissability as a function of belief and outcome. Results from Young et al. (2007), annotated with the estimated effects. Simple effects measure differences between the individual conditions and the neutral belief, neutral outcome condition. The interaction measures the difference 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).[3] 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 $\beta_{B,-O}$.[4] 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 minimum 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: $\beta_{-B,-O}$.[5] 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.

[3] 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.

[4] 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 $\beta_{-B} = \frac{(\theta_{-O,-B}+\theta_{O,-B})-(\theta_{-O,B}+\theta_{O,B})}{2}$.

[5] If you're reading carefully, you might be thinking that this all sounds like we're talking about the 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

Young et al.'s design, in which there are two factors with two levels each, is called a **2x2 design** (pronounced "two by two"). 2x2 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 towards her friend or is angry on that day. If we fully crossed this third affective factor with the other two (belief and outcome), we'd have a 2x2x2 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 in turn allow us to estimate both two-way and three-way interactions, enumerated in Table 9.1.

Table 9.1: Effects in a 2x2x2 design with affect, belief, and outcome as factors.

Effect	Term Type
Affect	Main effect
Belief	Main effect
Outcome	Main effect
Affect X Belief	2-way interaction
Affect X Outcome	2-way interaction
Belief X Outcome	2-way interaction
Affect X Belief X Outcome	3-way interaction

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 7 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![6] 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.

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

Depth

Estimation strategies for generalized factorial designs

So what should you do if you really do care about four or more factors – in the sense that you want to estimate their effects and include them in your theory? The simplest strategy is to start your research off by measuring them independently in a series of single-factor experiments. This kind of setup is natural when there is a single reference level for each factor of interest, and such experiments can provide a basis for judging which factors are most important for your outcome and hence which should be prioritized for experiments to estimate interactions.

On the other hand, sometimes there is no reference level for a factor. For example, in the Kovács, Téglás, and Endress (2010) paradigm, it's not clear whether a positive or negative belief is the reference level. That's not a problem in a fully-crossed design like theirs, but this situation can pose a problem if you have more than two such factors. Ideally you would want to run independent experiments, but you have to choose some level for all of the other variables – you can't just assume that one level is "neutral."

One solution that lets you compute main effects but not interactions is called a **Latin square**. Latin squares are a good solution for three-factor designs, which is the level at which a fully-crossed design typically gets overwhelming. A Latin square is an nxn matrix in which each number occurs exactly once in each row and column, e.g.

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix}$$

This Latin square for $n = 3$ gives the solution for how to balance factors across a 3x3x3 experiment. The row number is one factor, the column number is the second factor, and the number in the cell is the third factor. So one condition would be (1,1,1), the first level of all factors, shown in the upper left cell. Another would be (3,3,2), the lower right cell. Although a fully-crossed design would require 27 cells to be run, the Latin square has only nine. Critically, the combinations of factors are balanced across the nine cells so that the average effect of each level of the three factors can be estimated.

There are also fancier methods available. For example, the literature on **optimal experiment design** contains methods for choosing the most informative sequence of experiments to run in order to estimate the parameters in a model that can include many factors and their interactions (Myung and Pitt 2009). Going down this road typically means having an implemented computational theory of your domain, but it can be a very productive strategy for exploring a complex experimental space with many factors.

9.1.1 Between- vs. within-participant designs

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

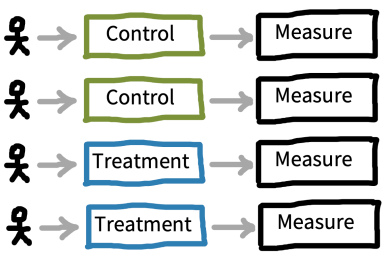


Figure 9.5: A between-participants design.

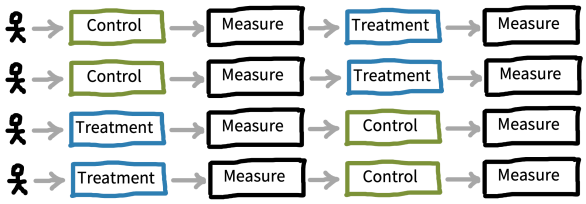


Figure 9.6: A within-participants design, counterbalanced for order (discussion of counterbalancing below).

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![7]

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:[8]

- **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!

[7] 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 that $N_{within} = N_{between}(1 - \rho)/2$ where ρ is the correlation between the measurement of the two conditions within individuals. The more correlated they are, the smaller your within-participants N.

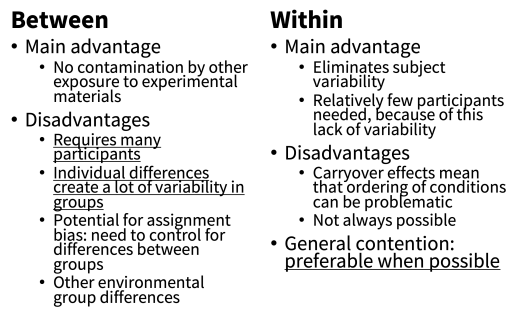
[8] We tend to think of all of these as being forms of carry-over effect, and sometimes use this label as a catch-all description. Some people also use the picturesque description "poisoning the well" (Gelman 2017) – earlier conditions "ruin" the data for later conditions.

- **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 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!
- **Carry-over 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 spill-over of effects between conditions.

All of these issues can lead to real concerns with respect to 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 we 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 carry-over 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 carryover effects these within your statistical model (for example by estimating the interaction of condition and order).[9]

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, Mathur, and Frank 2023).[10]



[9] 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.

[10] Caveat: this study used an observational design, so no causal inference is possible.

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

9.1.2 Repeated measurements and experimental items

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

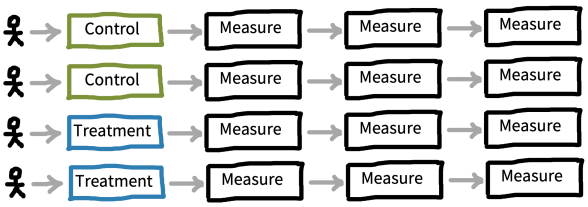


Figure 9.8: A between-participants, repeated-measures design.

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.[11] 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 depend 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 measurements 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.

[11] 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.

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).

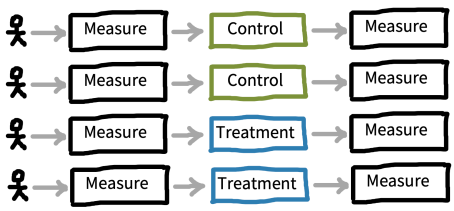


Figure 9.9: A between-participants, pre-post design.

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 pre measurements 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 $\beta = (\theta_{T_{post}} - \theta_{T_{pre}}) - (\theta_{C_{post}} - \theta_{C_{pre}})$. This equation says "how much more did the treatment group go up than the control group?[12]

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.

[12] 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.

Accident report

Stimulus-specific effects

Imagine you're a psycholinguist who has the hypothesis that nouns are processed faster than verbs. You run an experiment where you pick out ten verbs and ten nouns, then measure a large sample of participants' reading time for each of these. You find strong evidence for the predicted effect and publish a paper on your claim. The only problem is that, at the same time, someone else has done exactly the same study – with different nouns and verbs – and published a paper making the opposite claim. When this happens, it is possible that each effect is driven by the specific experimental items that were chosen, rather than a generalization that is true of nouns and verbs in general (Clark 1973).

The problem of generalization from sample to population is not new – as we discussed in Chapter 6, we are constantly making this kind of inference with the samples of people that participate in our experiments. Our classic statistical techniques are designed to quantify our ability to generalize from a sample of participants to a population, so we recognize that a very small sample size leads to a weak generalization. The exact same issue comes up with *items*: a very small sample of experimental items leads to a weak generalization to the population of items.

Item effects are kind of like accidentally finding a group of ten people whose left toes are longer than their right ones. If you continued to measure the same group's toes, you could continue to replicate the difference in length. But that doesn't mean it's true of the population as a whole.

This kind of **stimulus generalizability** problem comes up across many different areas of psychology. In one example, hundreds of papers were written about a phenomenon called the "risky shift" – in which groups deliberating about a decision would produce riskier decisions than individuals. Unfortunately, this phenomenon appeared to be completely driven by the specific choice of vignettes that groups deliberated about, with some stories producing a risky shift and others producing a more conservative shift (Westfall, Judd, and Kenny 2015).

Another example comes from the memory literature, where a classic paper by Baddeley, Thomson, and Buchanan (1975) suggested that words that take longer to pronounce ("tycoon" or "morphine") would be remembered worse than words that took a shorter amount of time ("ember" or "wicket") even when they had the same number of syllables. This effect also appears to be driven by the specific sets of words chosen in the original paper. It's very replicable with that particular stimulus set but not generalizable across other sets (Lovatt, Avons, and Masterson 2000).

The implication of these examples is clear: experimenters need to take care in both their experimental design and analysis to avoid overgeneralizing from their stimuli to a broader construct. Three primary steps can help experimenters avoid this pitfall:

1. To maximize generality, use samples of experimental items – words, pictures, or vignettes – that are comparable in size to your samples of participants.
2. When replicating an experiment, consider taking a new sample of items as well as a new sample of participants. It's more work to draft new items, but it will lead to more robust conclusions.
3. When experimental items are sampled at random from a broader population, use a statistical model that includes this sampling process (e.g., mixed effects models with random intercepts for items from Chapter 7).

9.1.1 Discrete and continuous experimental manipulations

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.

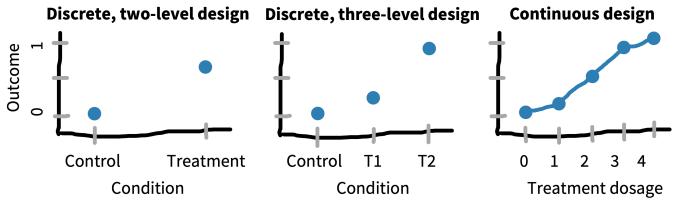


Figure 9.10: Three schematic designs. (left) Control and treatment are two levels of a nominal variable. (middle) Control is compared to ordered levels of a treatment. (right) Treatment level is an interval or ratio variable such that points can be connected and a parametric curve can be extrapolated.

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.[13]

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. 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.[14] 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!

[13] 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!

[14] 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 important point is that choosing a function makes explicit your assumptions about the nature of the treatment-effect relationship.

Depth

Tradeoffs associated with titrated designs

Like adults, babies like to look at more interesting, complex stimuli. But do they uniformly prefer complex stimuli, or do they search for stimuli at an appropriate level of complexity for their processing abilities? To test this hypothesis, Brennan, Ames, and Moore (1966) exposed infants in three different age groups (3, 8, and 14 weeks, N=30) to black and white checkerboard stimuli with three different levels of complexity (2x2, 8x8, and 24x24).

Their findings are plotted in Figure 9.11: the youngest infants preferred the simplest stimuli, while infants at an intermediate age preferred stimuli of intermediate complexity, and the oldest infants preferred the most complex stimuli. These findings help to motivate the theory that infants attend preferentially to stimuli that provide appropriate learning input for their processing ability (Kidd, Piantadosi, and Aslin 2012).

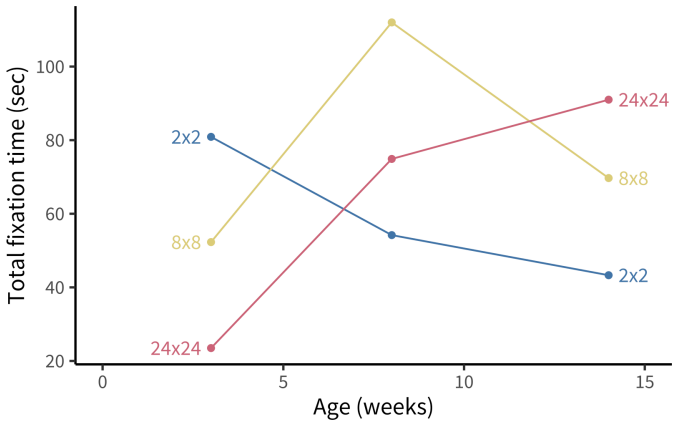


Figure 9.11: Infants' looking time, plotted by stimulus complexity and infant age. Data from Brennan, Ames, and Moore (1966).

If your goal is simply to detect whether an effect is zero or non-zero, then dose-response designs do not achieve the maximum statistical power. For example, if Brennan, Ames, and Moore (1966) simply wanted to achieve maximal statistical power, they probably should have only tested two age groups and two levels of complexity (say, 3 and 14 week infants and 2x2 and 24x24 checkerboards). That would have been enough to show an interaction of complexity and age, and their greater resources devoted to these four (as opposed to nine) conditions would mean more precise estimates of each. But their findings would be less clearly supportive of the view that infants prefer stimuli that are appropriate to their processing ability, because no group would have preferred an intermediate level of complexity (as the 9-week-olds apparently did). By seeking to measure intermediate conditions, they provided a stronger test of their theory.

9.2 Choosing your manipulation

In the previous section, we reviewed a host of common experimental designs. These designs provide a palette of common options for combining manipulations and measures. But your choice must be predicated on the specific manipulation you are interested in! In this section, we discuss considerations for experimenters as they design manipulations.

In Chapter 8, we talked about *measurement* validity, but the idea of validity concept can be applied to manipulations as well as measures. In particular, a manipulation is valid if it corresponds to the construct that the experimenter intends to intervene on. In this context, *internal* validity threats to manipulations tend to refer to cases where factors in the experimental design keep the intended manipulation from actually intervening on the construct of interest. In contrast, *external* validity threats to manipulations tend to be cases where the manipulation simply doesn't line up well with the construct of interest.

9.2.1 Internal validity threats: Confounding

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 $1000, 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 of the monetary gifts. Maybe happiness just went up more over time, irrespective of getting the second gift.

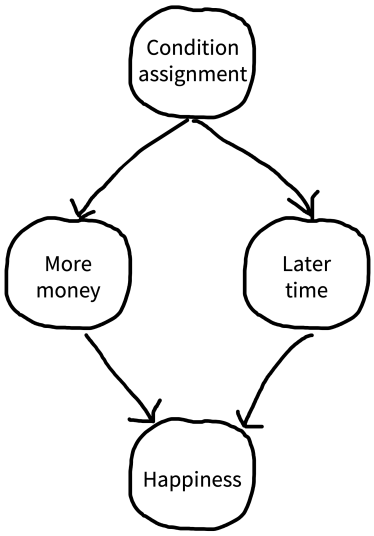


Figure 9.12: Confounding order and condition assignment means that you can't make an inference about the link between money and happiness.

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 $1000 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).[15] 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.

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 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.

[15] 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.

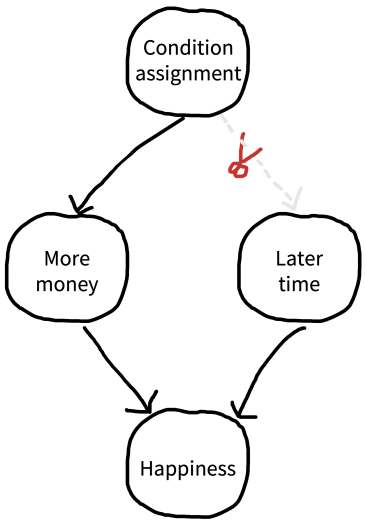


Figure 9.13: Confounding between a specific condition and the time at which it's administered can be removed by counterbalancing or randomization of order.

9.2.2 *Internal validity threats: Placebo, demand, and expectancy*

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 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.

Accident report

Brain training?

Can doing challenging cognitive tasks make you smarter? In the late 2000s and early 2010s, a large industry for "brain training" emerged. Companies like Lumos Labs, CogMed, BrainHQ, and CogniFit offered games, often modeled on cognitive psychology tasks, that claimed to lead to gains in memory, attention, and problem solving.

These companies were basing their claims in part on a scientific literature reporting that concerted training on difficult cognitive tasks could lead to benefits that **transferred** to other cognitive domains. Among the most influential of these was a study by Jaeggi et al. (2008). They conducted four experiments in which participants (N=70 across the studies) were assigned to either working memory training via a difficult working memory task (the "dual N-back") or a no-training control, with training varying from 8 days all the way to 19 days.

The finding from this study excited a tremendous amount of interest because they reported not only gains in performance on the specific training task but also on a general intelligence task that the participants had trained on. While the control group's scores on these tasks improved, presumably just from being tested twice, there was a condition by time (pre- vs. post) interaction such that the scores of the trained groups (consolidated across all four training experiments) grew significantly more over the training period (Figure 9.14). These results were interpreted as supporting transfer – whereby training on one task leads to broader gains – a key goal for "brain training."

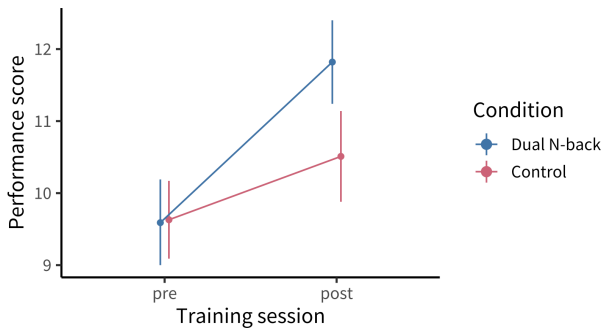


Figure 9.14: The primary outcome graph for data from Jaeggi et al. (2008).

Careful readers of the original paper noticed signs of analytic flexibility (as discussed in Chapters 3 and 6), however. For example, the key statistical model was fit to dataset created by post-hoc consolidation of experiments, which yielded $p = .025$ on the key interaction (Redick et al. 2013). When data were disaggregated, it was clear that the measures and effects had differed in each of the different sub-experiments (Figure 9.15).

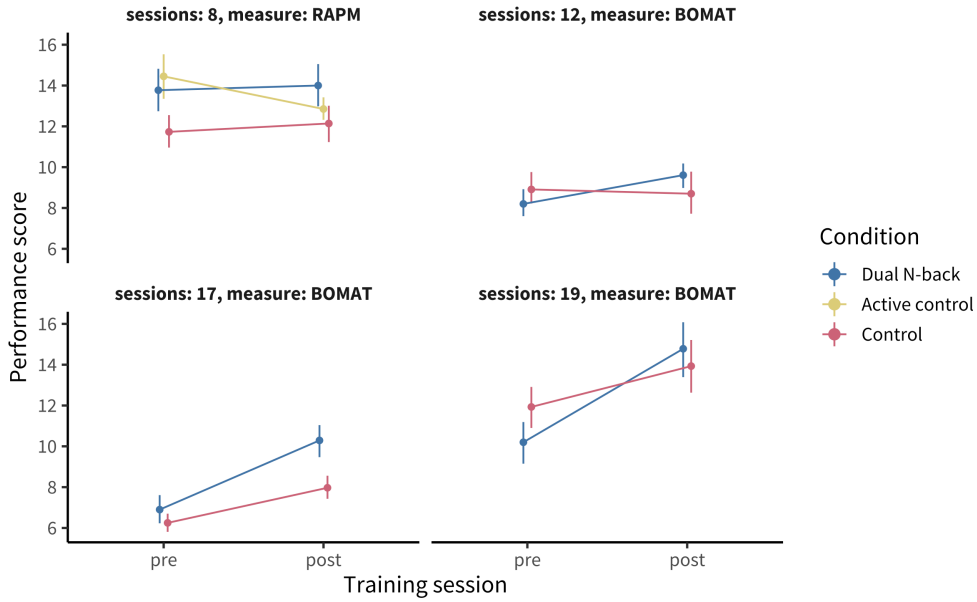


Figure 9.15: The four sub-experiments of Jaeggi et al. (2008), now disaggregated. Panels show 8-, 12-, 17-, and 19-session studies. Note the different measures: RAPM = Raven's Advanced Progressive Matrices; BOMAT = Bochumer Matrizentest. Based on Redick et al. (2013).

Several replications by the same group addressed some of these issues, but still failed to show convincing evidence of transfer. In particular, there was no comparison to an **active control group** in which participants did some kind of alternative activity for the same amount of time (Simons et al. 2016). Such a comparison is critical because a comparison to a **passive control group** (a group that does no intervention) confounds participants' general effort and involvement in the study with the specific training being used. Successful transfer compared to passive control could be the result of participants' involvement, expectations, or motivation rather than brain training per se.

A careful replication of the training study (N=74) with an active control group and a wide range of outcome measures failed to find any transfer effects from working-memory training (Redick et al. 2013). A meta-analysis of 23 studies concluded that their findings cast doubt on working memory training for increasing cognitive functioning (Melby-Lervåg and Hulme 2013). In one convincing test of the cognitive transfer theory, a BBC show ("Bang Goes The Theory") encouraged its listeners to participate in a six week online brain training study. More than 11,000 listeners completed the pre- and post-tests and at least two training sessions. Neither focused training of planning and reasoning nor broader training on memory, attention and mathematics led to transfer to untrained tasks.

Placebo effects are one plausible explanation for some positive findings in the brain training literature. Foroughi et al. (2016) recruited participants to participate via two different advertisements. The first advertised that "numerous studies have shown working memory training can increase fluid intelligence" ("placebo treatment" group) while the second simply offered experimental credits (control group). After a single training session, the placebo treatment group showed significant improvements to their matrix reasoning abilities. Participants in the placebo treatment group realized gains from training out of proportion with any they could have realized through training. Further, those participants who responded to the placebo treatment ad tended to endorse statements about the malleability of intelligence, suggesting that they might have been especially likely to self-select into the intervention.

Summarizing the voluminous literature on brain training, Simons et al. (2016) wrote: "Despite marketing claims from brain-training companies of 'proven benefits'… we find the evidence of benefits from cognitive brain training to be 'inadequate.' "

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 non-verbally in much the same way that Clever Hans learned (Rosnow and Rosenthal 1997).

Table 9.2: Magnitudes of expectancy effects. Based on Rosenthal (1994).

Domain	d	r	Example of type of study
Laboratory interviews	0.14	.07	Effects of sensory restriction on reports of hallucinatory experiences
Reaction time	0.17	.08	Latency of word associations to certain stimulus words
Learning and ability	0.54	.26	IQ test scores, verbal conditioning (learning)
Person perception	0.55	.27	Perception of other people's success
Inkblot tests	0.84	.39	Ratio of animal to human Rorschach responses
Everyday situations	0.88	.40	Symbol learning, athletic performance
Psychophysical judgments	1.05	.46	Ability to discriminate tones
Animal learning	1.73	.65	Learning in mazes and Skinner boxes
Weighted mean	0.70	.33	
Unweighted mean	0.74	.35	
Median	0.70	.33	

In medical research, the gold standard is an experimental design where neither patients nor experimenters know which condition the patients are in.[16] 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.

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

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.

9.2.1 External validity of manipulations

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).[17]

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

For testing the effect of money on happiness, our manipulation was to give participants $1000. 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 might have to do more experiments to figure out 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 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*, they would also have to be tracking the beliefs of depictions of non-human, 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.[18] 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, Brady, and Crum 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**).

[18] 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!

9.3 Summary: Experimental design

In this chapter, we started by examining some common experimental designs that allow us to measure effects associated with one or more manipulations. Our advice, in brief, was: "keep it simple!" The failure mode of many experiments is that they contain too many manipulations, and these manipulations are measured with too little precision.

Start with just a single manipulation, and measure it carefully. Ideally this measurement should be done via a within-participants design unless the manipulation is completely incompatible with this design. And if this design can incorporate a dose-response manipulation, it is more likely to provide a basis for quantitative theorizing.

How do you ensure that your manipulation is valid? A careful experimenter needs to consider possible confounds and ensure that these are controlled or randomized. They must also consider other artifacts including placebo, demand, and expectancy effects. Finally, they must begin thinking about the relation of their manipulation to the broader theoretical construct whose causal role they hope to test.

Discussion questions

1. Choose a classic study in your area of psychology. Analyze the design choices: how many factors were manipulated? How many measures were taken? Did it use a within-participants or between-participants design? Were measures repeated? Can you justify these choices with respect to trade-offs (e.g., carry-over effects, fatigue,

etc.)?

2. Consider the same study. Design an alternative version that varies one of these design parameters (e.g., drops a manipulation or measure, changes within- to between-participants, etc.). What are the pros and cons of this change? Do you think your design improves on the original?

Readings

– Much of this material is covered in more depth in the classic text on research methods: Rosenthal, R. & Rosnow, R. L. 2008. *Essentials of Behavioral Research: Methods and Data Analysis*. Third Edition. New York: McGraw-Hill. http://dx.doi.org/10.34944/dspace/66.

References

Baddeley, Alan D, Neil Thomson, and Mary Buchanan. 1975. "Word Length and the Structure of Short-Term Memory." *Journal of Verbal Learning and Verbal Behavior* 14 (6): 575–89.

Benedetti, Fabrizio. 2020. *Placebo Effects*. Oxford University Press.

Boyce, Veronica, Maya Mathur, and Michael C Frank. 2023. "Eleven Years of Student Replication Projects Provide Evidence on the Correlates of Replicability in Psychology." *Royal Society Open Science* 10 (11): 231240.

Brennan, Wendy M, Elinor W Ames, and Ronald W Moore. 1966. "Age Differences in Infants' Attention to Patterns of Different Complexities." *Science* 151 (3708): 354–56.

Clark, Herbert H. 1973. "The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research." *Journal of Verbal Learning and Verbal Behavior* 12 (4): 335–59.

Cunningham, Scott. 2021. *Causal Inference*. Yale University Press.

El Kaddouri, Rachida, Lara Bardi, Diana De Bremaeker, Marcel Brass, and Roeljan Wiersema. 2020. "Measuring Spontaneous Mentalizing with a Ball Detection Task: Putting the Attention-Check Hypothesis by Phillips and Colleagues (2015) to the Test." *PSYCHOLOGICAL RESEARCH-PSYCHOLOGISCHE FORSCHUNG* 84 (6): 1749–57.

Foroughi, Cyrus K, Samuel S Monfort, Martin Paczynski, Patrick E McKnight, and PM Greenwood. 2016. "Placebo Effects in Cognitive Training." *Proceedings of the National Academy of Sciences* 113 (27): 7470–74.

Gelman, Andrew. 2017. "Poisoning the Well with a Within-Person Design? What's the Risk?" 2017. https://statmodeling.stat.columbia.edu/2017/11/25/poisoning-well-within-person-design-whats-risk/.

Greenwald, Anthony G. 1976. "Within-Subjects Designs: To Use or Not to Use?" *Psychological Bulletin* 83 (2): 314.

Jaeggi, Susanne M, Martin Buschkuehl, John Jonides, and Walter J Perrig. 2008. "Improving Fluid Intelligence with Training on Working Memory." *Proceedings of the National Academy of Sciences* 105 (19): 6829–33.

Kane, Michael T. 1992. "An Argument-Based Approach to Validity." *Psychological Bulletin* 112 (3): 527.

Kidd, Celeste, Steven T Piantadosi, and Richard N Aslin. 2012. "The Goldilocks Effect: Human Infants Allocate Attention to Visual Sequences That Are Neither Too Simple nor Too Complex." *PloS One* 7 (5): e36399.

Kovács, Ágnes Melinda, Ernő Téglás, and Ansgar Denis Endress. 2010. "The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults." *Science* 330 (6012): 1830–34.

Lakens, Daniel. 2016. "Why Within-Subject Designs Require Fewer Participants Than Between-Subject Designs." 2016. https://daniellakens.blogspot.com/2016/11/why-within-subject-designs-require-less.html.

Lovatt, Peter, Steve E Avons, and Jackie Masterson. 2000. "The Word-Length Effect and Disyllabic Words." *The Quarterly Journal of Experimental Psychology: Section A* 53 (1): 1–22.

McClelland, Gary H, and Charles M Judd. 1993. "Statistical Difficulties of Detecting Interactions and Moderator Effects." *Psychological Bulletin* 114 (2): 376.

Melby-Lervåg, Monica, and Charles Hulme. 2013. "Is Working Memory Training Effective? A Meta-Analytic Review." *Developmental Psychology* 49 (2): 270.

Myung, Jay I, and Mark A Pitt. 2009. "Optimal Experimental Design for Model Discrimination." *Psychological Review* 116 (3): 499.

Orne, Martin T. 1962. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications." *American Psychologist* 17 (11): 776.

Phillips, Jonathan, Desmond C Ong, Andrew DR Surtees, Yijing Xin, Samantha Williams, Rebecca Saxe, and Michael C Frank. 2015. "A Second Look at Automatic Theory of Mind: Reconsidering Kovács, téglás, and Endress (2010)." *Psychological Science* 26 (9): 1353–67.

Redick, Thomas S, Zach Shipstead, Tyler L Harrison, Kenny L Hicks, David E Fried, David Z Hambrick, Michael J Kane, and Randall W Engle. 2013. "No Evidence of Intelligence Improvement After Working Memory Training: A Randomized, Placebo-Controlled Study." *Journal of Experimental Psychology: General* 142 (2): 359.

Rosenthal, Robert. 1994. "Interpersonal Expectancy Effects: A 30-Year Perspective." *Current Directions in Psychological Science* 3 (6): 176–79.

Rosnow, Ralph, and Robert Rosenthal. 1997. *People Studying People: Artifacts and Ethics in Behavioral Research*. WH Freeman.

Simons, Daniel J, Walter R Boot, Neil Charness, Susan E Gathercole, Christopher F Chabris, David Z Hambrick, and Elizabeth AL Stine-Morrow. 2016. "Do 'Brain-Training' Programs Work?" *Psychological Science in the Public Interest* 17 (3): 103–86.

Walton, Gregory M, Shannon T Brady, and AJ Crum. 2020. "The Social-Belonging Intervention." *Handbook of Wise Interventions: How Social Psychology Can Help People Change*, 36–62.

Walton, Gregory M, and Geoffrey L Cohen. 2011. "A Brief Social-Belonging Intervention Improves Academic and Health Outcomes of Minority Students." *Science* 331 (6023): 1447–51.

Westfall, Jacob, Charles M Judd, and David A Kenny. 2015. "Replicating Studies in Which Samples of Participants Respond to Samples of Stimuli." *Perspectives on Psychological Science* 10 (3): 390–99.

Young, Liane, Fiery Cushman, Marc Hauser, and Rebecca Saxe. 2007. "The Neural Basis of the Interaction Between Theory of Mind and Moral Judgment." *Proceedings of the National Academy of Sciences* 104 (20): 8235–40.

10 Sampling

Learning goals

- Discuss sampling theory and stratified sampling
- Reason about the limitations of different samples, especially convenience samples
- Consider sampling biases and how they affect your inferences
- Learn how to choose and justify an appropriate sample size for your experiment

As we keep reminding you, experiments are designed to yield measurements of a causal effect. But a causal effect of what, and for whom? These are questions that are often given surprisingly little air time in our papers. Titles in our top journals read "Daxy thinking promotes fribbles," "Doing fonzy improves smoodling," or "Blicket practice produces more foozles than smonkers."[1] Each of these uses **generic language** to state a claim that is implied to be generally true (DeJesus et al. 2019),[2] but for each of these, we could reasonably ask "for whom?". Is it everyone? Or a particular set of people? These are questions about our key theme, GENERALIZABILITY.

Let's focus on smoodling. We wouldn't let the authors get away with a fully universal version of their claim: "Doing [*any*] fonzy improves smoodling [*for everyone*]." The non-generic version states a generalization that goes way beyond the evidence we actually have. But it seems that we are often OK with authors *implying* (with generic language) that their findings generalize broadly. Imagine for a second what the completely specific version of one of these titles might look like: "Reading one particular selection of fonzy for fifteen minutes in the lab improved 36 college students' smoodling scores on a questionnaire." This paper sounds pretty narrow in its applicability!

We've already run into generalizability in our treatment of statistical estimation and inference. When we estimated a particular quantity (say, the effect of fonzy), we did so in our own sample. But we then used inferential tools to reason about how the estimate in this **sample** related to the parameter in the **population** as a whole. How do we link up these *statistical* tools for generalization to the *scientific* questions we have about the generalizability of our findings? That's the question of this chapter.

[1] Titles changed to protect the original authors. These researchers might very well have said more specific things in the text of their paper.

[2] Generic language is a fascinating linguistic phenomenon. When we say things like "mosquitoes transmit malaria," we don't mean that *all* mosquitoes do it, only something like "it's a valid and diagnostic generalization about mosquitoes in contrast to other relevant insects or other creatures that they are spreaders of malaria" (Tessler and Goodman 2019).

A key set of decisions in experiment planning is what population to sample from and how to sample. We'll start by talking about the basics of **sampling theory**: different ways of sampling and the generalizations they do and don't license. The second section of the chapter will then deal with **sampling biases** that can compromise our effect estimates. A final set of key decisions is about **sample size** planning. In the third part of the chapter we'll address this issue, starting with classic **power analysis** but then introducing several other ways that an experimenter can plan and justify their sample size.

10.1 Sampling theory

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.[3] Sampling strategies are split into two categories. **Probability sampling** strategies are those in which each member of the population has some known, pre-specified 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."

[3] 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.

Case study

Is everyone bad at describing smells?

Since Darwin, scientists have assumed that smell is a vestigial sense in humans – one that we don't even bother to encode in language. In English we don't even have consistent words for odors. We can say something is "stinky," "fragrant", or maybe "musty," but beyond these, most of our words for smells are about the *source* of the smell, not the qualities of it. Bananas, roses, and skunks all have distinctive smells, but we don't have any vocabulary for naming what is common or uncommon about them. And when we make up ad-hoc vocabulary, it's typically quite inconsistent (Majid and Burenhult 2014). The same situation applies across many languages.

So, would it be a good generalization about human beings – all people – that olfaction as a sense is de-emphasized relative to, say, vision? This inference has a classic sample-to-population structure. Within several samples of participants using widely-spoken languages, we observe limited and inconsistent vocabulary for smells, as well as poor discrimination. We use these samples to license an inference to the population – in this case, the entire human population.

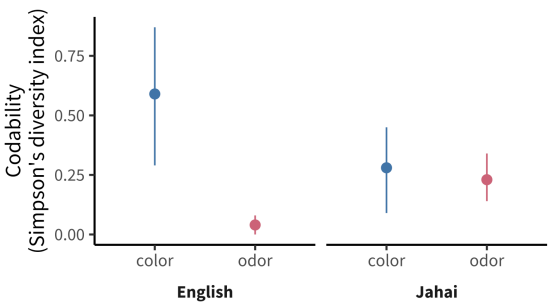


Figure 10.1: Data from Majid and Burenhult (2014) on the consistency of color and odor naming in English and Jahai speakers. Higher values indicate more consistent descriptions. Error bars show standard deviation.

But these inferences about the universal lack of olfactory vocabulary are likely based on choosing non-representative samples! Multiple hunter-gatherer groups appear to have large vocabularies for consistent smell description. For example, the Jahai, a hunter-gatherer group on the Malay Peninsula, have a vocabulary that includes at least twelve words for distinct odors, for example /cŋ□s/, which names odors with a "stinging smell" like gasoline, smoke, or bat droppings. When Jahai speakers are asked to name odors, they produce shorter and much more consistent descriptions than English speakers – in fact, their smell descriptions were as consistent as their color descriptions (Figure 10.1). Further studies implicate the hunter-gatherer lifestyle as a factor: while several hunter-gatherer groups show good odor naming, nearby horticulturalist groups don't (Majid and Kruspe 2018).

Generalizations about humans are tricky. If you want to estimate the average odor naming ability, you could take a random sample of humans and evaluate their odor naming. Most of the individuals in the sample would likely speak English, Mandarin, Hindi, or Spanish. Almost certainly, none of them would speak Jahai, which is spoken by only a little more than a thousand people and is listed as Threatened by Ethnologue (https://www.ethnologue.com/language/jhi). Your estimate of low odor naming stability might be a good guess for the *majority* of the world's population, but would tell you little about the Jahai.

On the other hand, it's more complicated to jump from a statistical generalization about average ability to a richer claim, like "humans have low olfactory naming ability." Such claims about universal aspects of the human experience require much more care and much stronger evidence (Piantadosi and Gibson 2014). From a sampling perspective, human behavior and cognition show immense and complex **heterogeneity** – variability of individuals and variability across clusters. Put simply, if we want to know what people in general are like, we have to think carefully about which people we include in our studies.

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 get overwhelming. How will you find their names – what if they aren't in the phone book? How will you contact them – what if they don't have email? 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?

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 resemble the population in their representation of one or several 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.[4]

[4] 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 stu-

Depth

Representative samples and stratified sampling

Stratified sampling is a cool method that can help you get more precise estimates of your experimental effect, if you think it varies across some grouping in your sample. Imagine you're interested in a particular measure in a population – say, attitudes towards tea drinking across US adults – but you think that this measure will vary with one or more characteristics such as whether the adults are frequent, infrequent, or non-coffee drinkers. Even worse, your measure might be more variable within one group: perhaps most frequent and infrequent coffee drinkers feel

one.

OK about tea, but as a group non-coffee drinkers tend to hate it (most don't drink any caffeinated beverages).

A simple random sample from this heterogeneous population *will* yield statistical estimates that converge asymptotically to the correct population average for tea-drinking attitudes. But it will do so more slowly than ideal because any given sample may over- or under-sample non-drinkers just by chance. In a small sample, if you happen to get too many non-coffee drinkers, your estimate of attitudes will be biased downward; if you happen to get too few, you will be biased upwards. All of this will come out in the wash eventually, but any individual sample (especially a small one) will be noisier than ideal.

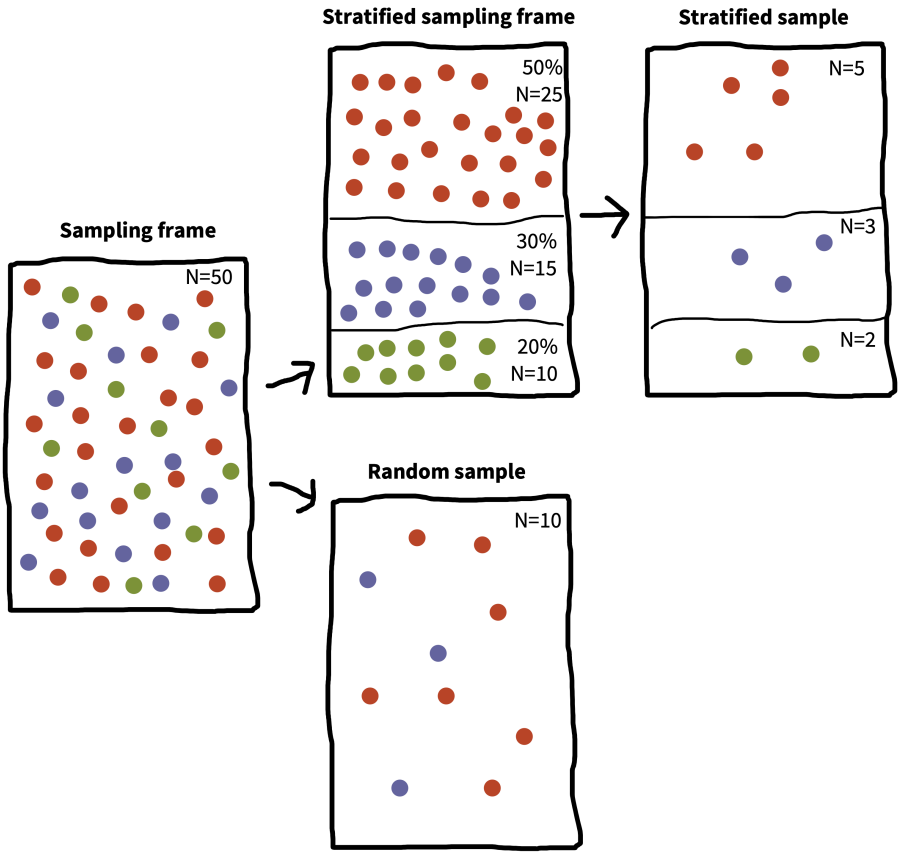


Figure 10.2: Illustration of stratified sampling. The left panel shows the sampling frame. The upper frames show the sampling frame stratified by a participant characteristic and a stratified sample. The lower frame shows a simple random sample, which happens to omit one group completely by chance.

But, if you know the proportion of frequent, infrequent, or non-coffee drinkers in the population, you can perform stratified sampling within those subpopulations to ensure that your sample is representative along this dimension (Neyman 1992). This situation is pictured in Figure 10.2, which shows how a particular sampling frame can be broken up into groups for stratified sampling. The result is a sample that matches the population proportions on a particular characteristic. In contrast, a simple random sample can over- or under-sample the subgroups by chance.

Stratified sampling can lead to substantial gains in the precision of your estimate. These gains are most prominent when either the groups differ a lot in their mean or when they differ a lot in their variance. There are several

important refinements of stratified sampling in case you think these methods are important for your problem. In particular, **optimal sampling** can help you figure out how to over-sample groups with higher variance. On the other hand, if the characteristic on which you stratify participants doesn't relate to your outcome at all, then estimates from stratified sampling converge just as fast as random sampling (though it's a bit more of a pain to implement).

Figure 10.3 shows a simulation of the scenario in Figure 10.2, in which each coffee preference group has a different tea attitude mean, and the smallest group has the biggest variance. Although the numbers here are invented, it's clear that estimation error is much smaller in the stratified group and estimation error declines much more quickly as samples get larger.

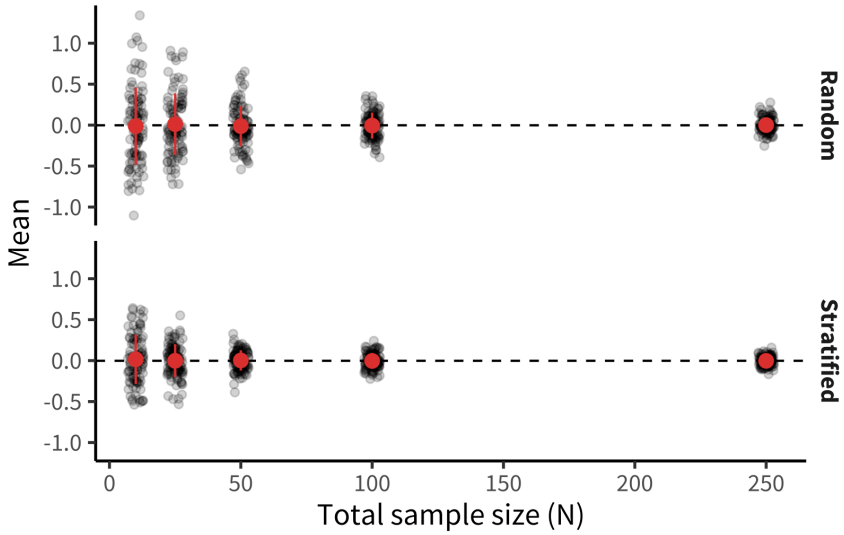


Figure 10.3: Simulation showing the potential benefits of stratification. Each dot is an estimated mean for a sample of a particular size, sampled randomly or with stratification. Red points show the mean and standard deviation of sample estimates.

Stratification is everywhere, and it's useful even in convenience samples. For example, researchers who are interested in development typically stratify their samples across ages (e.g., recruiting equal numbers of two- and three-year-olds for a study of preschoolers). You can estimate developmental change in a pure random sample, but you are guaranteed good coverage of the range of interest when you stratify.

If you have an outcome that you think varies with a particular characteristic, it's not a bad idea to consider stratification. But don't go overboard – you can drive yourself to distraction finding the last left-handed non-binary coffee drinker to complete your sample. Focus on stratifying when you know the measure varies with the characteristic of interest.

10.2 Convenience samples, generalizability, and the WEIRD problem

Now let's go back to the question of generalizability. How generalizable are the experimental effect estimates that we obtain in experiments that are conducted only with convenience samples? We'll start by laying out the worst version of the problem of generalizability in experimental

psychology. We'll then try to pull back from the brink and discuss some reasons why we might not want to be in despair despite some of the generalizability issues that plague the psychology literature.

10.2.1 The worst version of the problem

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, quoted in Rosenthal and Rosnow 1984). 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, those local and national contexts are also unrepresentative of the broad range of human experiences. Henrich, Heine, and Norenzayan (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.[5]

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, Judd, and Kenny 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.

[5] 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.

> 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

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 non-representative 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, Powell, and Sumner 2018).

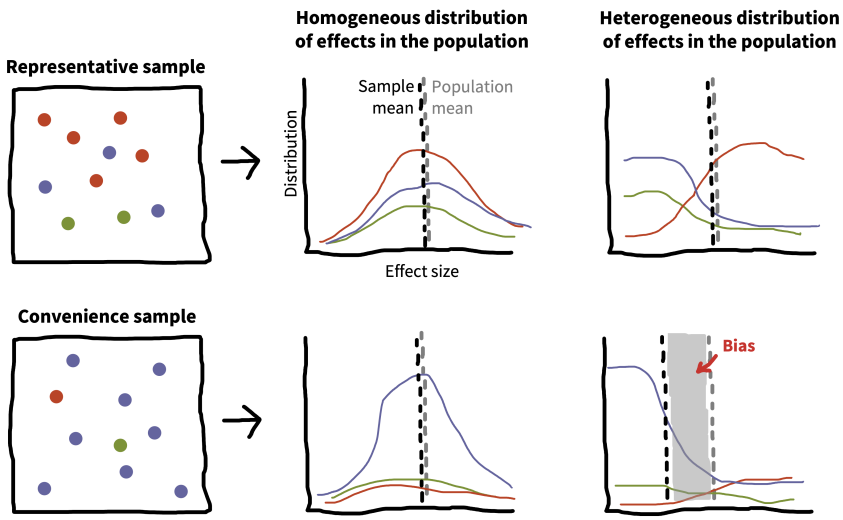


Figure 10.4: Illustration of the interaction of heterogeneity and convenience samples. Colors indicate arbitrary population subgroups. Left hand panels show sample composition. Individual plots show the distribution of effect sizes in each subgroup.

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 over-represented in the sample. The convenience sample doesn't cause problems if the size of your effect is homogeneous in the population – as with the case of smoodling or Stroop. The trouble comes when you have an effect that is heterogeneous. Because one group is over-represented, 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 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.[6] As naïve 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, Heine, and Norenzayan 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, Leeper, and Mullinix (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!

[6] 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 (2007)'s "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.

10.3 Biases in the sampling process

In fields like econometrics or epidemiology that use observational methods to estimate causal effects, reasoning about **sampling biases** is a critical part of estimating generalizable effects. If your sample does not represent the population of interest, then your effect estimates will be biased.[7] In the kind of experimental work we are discussing many of these issues are addressed by random assignment, including the first issue we treat: **collider bias**. Not so for the second one, **attrition bias**, which is an issue even in randomized experiments.

[7] There is a deep literature on correcting these biases using causal inference frameworks. These techniques are well outside of the scope of this book, but if you're interested, you might look at some of the textbooks we recommended earlier, e.g. Cunningham (2021).

10.3.1 Collider bias

Imagine you want to measure the association between money and happiness through a (non-experimental) survey. As we discussed in Chapter 1, there are plenty of causal processes that could lead to this association. Figure 10.5 shows several of these scenarios. Money could truly cause happiness (1); happiness could cause you to make more money (2); or some third factor – say having lots of friends – could cause people to be happier *and* richer (3).

But we can also create spurious associations if we are careless in our sampling. One prominent problem that we can induce is called **collider bias**. Suppose we recruited our sample from the clients of a social services agency. Unfortunately, both of our variables might affect presence in a social service agency (Figure 10.5, 4): people might be interacting with the agency for financial or benefits assistance, or else for psychological services (perhaps due to depression).

Being in a social services sample is called a **collider** variable because the two causal arrows *collide* into it (they both point to it). If we look just within the social services sample, we might see a *negative* association between wealth and happiness – on average the people coming for financial assistance would have less wealth and more happiness than the people coming for psychological services. The take-home here is that in observational research, you need to think carefully about the causal structure of your sampling process (Rohrer 2018)!

If you are doing experimental research, you are mostly protected from this kind of bias: Random assignment still "works" even in sub-selected samples. If you run a money intervention within a social-services population using random assignment, you can still make an unbiased estimate of the effect of money on happiness. But that estimate will only be valid *for members of that sub-selected population*.

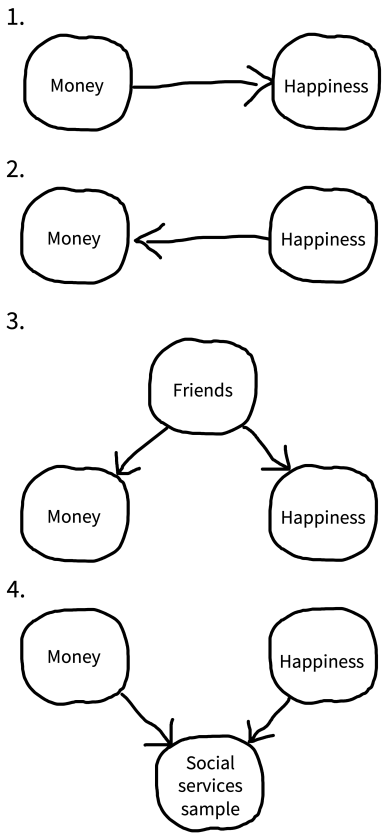


Figure 10.5: Four reasons why money and happiness can be correlated in a particular sample: 1. causal relationship, 2. reverse causality, 3. confounding with friendship, and 4. collider bias. For this last scenario, we have to assume that our measurement is *conditioned* on being in this sample, meaning we only look at the association of money and happiness within the social services sample.

10.3.2 Attrition bias

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 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.

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, Aronson, and Bankhead 2018)! Imagine 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.[8]

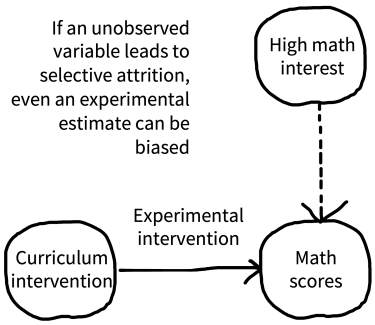


Figure 10.6: Selective attrition can lead to a bias even in the presence of random assignment. Dashed line indicates a causal relationship that is unobserved by the researcher.

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 4 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 counter-intuitive 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

[8] 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.

or threats to the generalizability of your findings.[9]

10.4 Sample size planning

Now that you have spent some time considering your sample and what population it represents, how many people will your sample contain? Continuing to collect data until you observe a $p < .05$ in an inferential test is a good way to get a false positive. This practice, known as "optional stopping," is a good example of a practice that invalidates p-values, much like the cases of analytic flexibility discussed in Chapter 3 and Chapter 6.

Decisions about when to stop collecting data should not be data-dependent. Instead you should transparently declaring your data collection **stopping rule** in your study preregistration (see Chapter 11). This step will reassure readers that there is no risk of bias from optional stopping. The simplest stopping rule is "I'll collect data until I get to a target N" – all that's needed in this case is a value for N.

But how do you decide N? It's going to be dependent on the effect that you want to measure, and how it varies in the population. Smaller effects will require larger sample sizes. Classically, N was computed using **power analysis**, which can provide a sample size for which you have a good chance of rejecting the null hypothesis (given a particular expected effect size). We'll introduce this computation below.

Classical power analysis is not the only way to plan your sample size. There are a number of other useful strategies, some of which rely on the same kinds of computations as power analysis (Table 10.1). Each of these can provide a valid justification for a particular sample size, but they are useful in different situations.

Table 10.1: Types of data collection stopping rules.

Method	Stopping Rule	Example
Power analysis	Stop at N for known probability of rejecting the null given known effect size	Randomized trial with strong expectations about effect size
Resource constraint	Stop collecting data after a certain amount of time or after a certain amount of resources are used	Time-limited field work
Smallest effect size of interest	Stop at N for known probability of rejecting the null for effects greater than some minimum	Measurement of a theoretically important effect with unknown magnitude

[9] 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.

Method	Stopping Rule	Example
Precision analysis	Stop at N that provides some known degree of precision in measure	Experimental measurement to compare with predictions of cognitive models
Sequential analysis	Stop when a known inferential criterion is reached	Intervention trial designed to accept or reject null with maximal efficiency

10.4.1 Power analysis

Let's start by reviewing the null-hypothesis significance testing paradigm that we introduced in Chapter 6. Recall that we introduced the Neyman-Pearson decision-theoretic view of testing in Chapter 6, shown again in Figure 10.7. The idea was that we've got some null hypothesis H_0 and some alternative H_1 – something like "no effect" and "yes, there is some effect with known size"– and we want to use data to decide which state we're in. α is our criterion for rejecting the null, conventionally set to $\alpha = .05$.

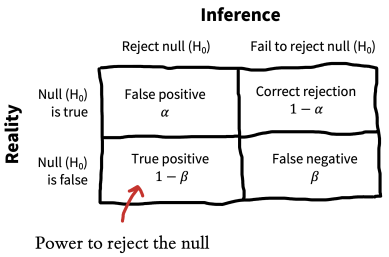


Figure 10.7: Standard decision matrix for NHST.

But what if H_0 is actually false and the alternative H_1 is true? Not all experiments are equally well set up to reject the null in those cases. Imagine doing an experiment with $N = 3$. In that case, we'd almost always fail to reject the null, even if it were false. Our sample would almost certainly be too small to rule out sampling variation as the source of our observed data.

Let's try to quantify our willingness to *miss* the effect – the false negative rate. We'll denote this probability with β. If β is the probability of missing an effect (failing to reject the null when it's really false), then $1-\beta$ is the probability that we *correctly reject the null when it is false*. That's what we call the **statistical power** of the experiment.

We can only compute power if we know the effect size for the alternative hypothesis. If the alternative hypothesis is a small effect, then the probability of rejecting the null will typically be low (unless the sample size is very large). In contrast, if the alternative hypothesis is a large effect, then the probability of rejecting the null will be higher.

The same dynamic holds with sample size: the same effect size will be easier to detect with a larger sample size than with a small one. Figure 10.8 shows how this relationship works. A large sample size creates a tighter null distribution (right side) by reducing sampling error. A tighter null distribution means you can reject the null more of the time based on the variation in a true effect. If your sample size is too

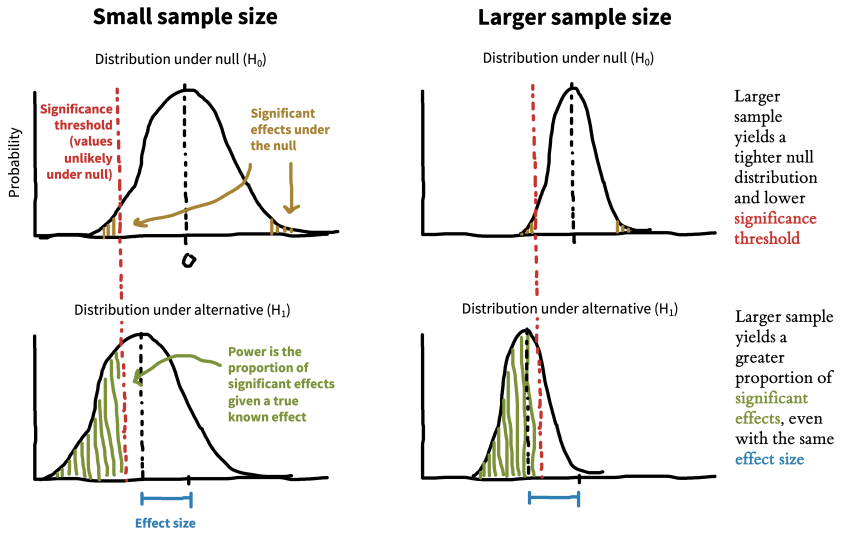


Figure 10.8: Illustration of how larger sample sizes lead to greater power.

small to detect your effect much of the time, we call this being **underpowered**.[10]

Classical power analysis involves computing the sample size N that's necessary in order to achieve some level of power, given α and a known effect size.[11] The mathematics of the relationship between α, β, N, and effect size have been worked out for a variety of different statistical tests (Cohen 2013) and codified in software like G*Power (Faul et al. 2007) and the `pwr` package for R (Champely et al. 2017). For other cases (including mixed effects models), you may have to conduct a simulation in which you generate many simulated experimental runs under known assumptions and compute how many of these lead to a significant effect; luckily, R packages exist for this purpose as well, including the `simr` package (Green and MacLeod 2016).

[10] You can also refer to a design as **overpowered**, though we object slightly to this characterization, since the value of large datasets is typically not just to reject the null but also to measure an effect with high precision and to investigate how it is moderated by other characteristics of the sample.

[11] Our focus here is on giving you a conceptual introduction to power analysis, but we refer you to Cohen (1992) for a more detailed introduction.

10.4.2 Power analysis in practice

Let's do a power analysis for our hypothetical money and happiness experiment. Imagine the experiment is a simple two group design in which participants from a convenience population are randomly assigned either to receive $1000 and some advice on saving money (experimental condition) vs. just receiving the advice and no money (control condition). We then follow up a month later and collect self-reported happiness ratings. How many people should we have in our study in order to be able to reject the null? The answer to this question depends on our desired values of α and β as well as our expected effect size for the intervention.

For α we will just set a conventional significance threshold of $\alpha = .05$. But what should be our desired level of power? The usual standard in the social sciences is to aim for power above 80% (e.g., $\beta < .20$); this gives you 4 out of 5 chances to observe a significant effect. But just like $\alpha = .05$, this is a conventional value that is perhaps a little bit too loose for modern standards – a strong test of a particular effect should probably have 90% or 95% power.[12]

[12] Really, researchers interested in using power analysis in their work should give some thought to what sort of chance of a false negative they are willing to accept. In exploratory research perhaps a higher chance of missing an effect is reasonable; in contrast, in confirmatory research it might make sense to aim for a higher level of power.

These choices are relatively easy, compared to the fundamental issue: our power analysis requires some expectation about our effect size. This is the **first fundamental problem of power analysis**: if you knew the effect size, you might not need to do the experiment!

So how are you supposed to get an estimate of effect size? Here are a few possibilities:

- **Meta-analysis**. If there is a good meta-analysis of the effect that you are trying to measure (or something closely related), then you are in luck. A strong meta-analysis will have not only a precise effect size estimate but also some diagnostics detecting and correcting potential publication bias in the literature (see Chapter 16). While these diagnostics are imperfect, they still can give you a sense for whether you can use the meta-analytic effect size estimate as the basis for a power analysis.
- **Specific prior study**. A more complicated scenario is when you have only one or a handful of prior studies that you would like to use as a guide. The trouble is that any individual effect in the literature is likely to be inflated by publication and other selective reporting biases (see Chapter 3). Thus, using this estimate likely means your study will be under-powered – you might not get as lucky as a previous study did!
- **Pilot testing**. Many people (including us) at some point learned that one way to do a power analysis is to conduct a pilot study, estimate the effect size from the pilot, and then use this effect estimate for power analysis in the main study. We don't recommend this practice. The trouble is that your pilot study will have a small sample size, leading to a very imprecise estimate of effect size (Browne 1995). If you over-estimate the effect size, your main study will be very under-powered. If you under-estimate, the opposite will be true. Using a pilot for power analysis is a recipe for problems.
- **General expectations about an effect of interest**. In our view, perhaps the best way you can use power analysis (in the absence of a really strong meta-analysis, at least) is to start with a general idea

about the size of effect you expect and would like to be able to detect. It is totally reasonable to say, "I don't know how big my effect is going to be, but let's see what my power would be if it were *medium-sized* (say $d = .5$), since that's the kind of thing we're hoping for with our money intervention." This kind of power analysis can help you set your expectations about what range of effects you might be able to detect with a given sample size.

For our money study, using our general expectation of a medium size effect, we can compute power for $d = .5$. In this case, we'll simply use the two-sample t-test introduced in Chapter 6, for which 80% power at $\alpha = .05$ and $d = .5$ is achieved by having $N = 64$ in each group.

CODE

Classic power analysis in R is quite simple using the `pwr` package. The package offers a set of test-specific functions like `pwr.t.test()`. For each, you supply three of the four parameters specifying effect size (`d`), number of observations (`n`), significance level (`sig.level`), and power (`power`); the function computes the fourth. For classic power analysis, we leave out `n`:

```
pwr.t.test(d = .5,
           power = .8,
           sig.level = .05,
           type = "two.sample",
           alternative = "two.sided")
```

But it is also possible to use this same function to compute the power achieved at a combination of n and d, for example.

There's a second issue, however. The **second fundamental problem of power analysis** is that the real effect size for an experiment may be zero. And in that case, *no* sample size will let you correctly reject the null. Going back to our discussion in Chapter 6, the null hypothesis significance testing framework is just not set up to let you *accept* the null hypothesis. If you are interested in a bi-directional approach to hypothesis testing in which you can accept *and* reject the null, you may need to consider Bayes Factor or **equivalence testing** approaches (Lakens, Scheel, and Isager 2018), which don't fit the assumptions of classical power analysis.

10.4.3 Alternative approaches to sample size planning

Let's now consider some alternatives to classic power analysis that can still yield reasonable sample size justifications.

1. **Resource constraint**. In some cases, there are fundamental resource constraints that limit data collection. For example, if you are doing fieldwork, sometimes the right stopping criterion for data collection is "when the field visit is over," since every additional datapoint is valuable. When pre-specified, these kinds of sample size justifications can be quite reasonable, although they do not preclude being under-powered to test a particular hypothesis.

2. **Smallest effect size of interest (SESOI)**. SESOI analysis is a variant on power analysis that includes some resource constraint planning. Instead of trying to intuit how big your target effect is, you instead choose a level below which you might not be interested in detecting the effect. This choice can be informed by theory (what is predicted), applied concerns (what sort of effect might be useful in a particular context), or resource constraints (how expensive or time-consuming it might be to run an experiment). In practice, SESOI analysis simply a classic power analysis with a particular small effect as the target.

3. **Precision-based sample planning**. As we discussed in Chapter 6, the goal of research is not always to reject the null hypothesis! Sometimes – we'd argue that it should be most of the time – the goal is to estimate a particular causal effect of interest with a high level of precision, since these estimates are a prerequisite for building theories. If what you want is an estimate with known precision (say, a confidence interval of a particular width), you can compute the sample size necessary to achieve that precision (Bland 2009; Rothman and Greenland 2018).[13]

4. **Sequential analysis**. Your stopping rule need not be a hard cutoff at a specific N. Instead, it's possible to plan a **sequential analysis** using either frequentist or Bayesian methods, in which you plan to stop collecting data once a particular inferential threshold is reached. For the frequentist version, the key thing that keeps sequential analysis from being p-hacking is that you pre-specify particular values of N at which you will conduct tests and then correct your p-values for having tested multiple times (Lakens 2014). For Bayesian sequential analysis, you can actually compute a running Bayes factor as you collect data and stop when you reach a pre-specified level of evidence (Schönbrodt et al. 2017). This latter alternative has the advantage of allowing you to collect evidence *for* the null as well as against it.[14]

In sum, there are many different ways of justifying your sample size or your stopping rule. The most important things are 1) to pre-specify

[13] In our experience, this kind of planning is most useful when you are attempting to gather measurements with sufficient precision to compare between computational models. Since the models can make quantitative predictions that differ by some known amount, then it's clear how tight your confidence intervals need to be.

[14] Another interesting variant is sequential parameter estimation, in which you collect data until a desired level of precision is achieved (Kelley, Darku, and Chattopadhyay 2018); this approach combines some of the benefits of both precision-based analysis and sequential analysis.

your strategy and 2) to give a clear justification for your choice. Table 10.2 gives an example sample size justification that draws on several different concepts discussed here, using classical power computations as one part of the justification. A reviewer could easily follow the logic of this discussion and form their own conclusion about whether this study had an adequate sample size and whether it should have been conducted given the researchers' constraints.

Table 10.2: Example sample size justification, referencing elements of SESOI, resource-limitation, and power-based reasoning.

Element	Justification Text
Background	We did not have strong prior information about the likely effect size, so we could not compute a classical power analysis.
Smallest effect of interest	Because of our interest in meaningful factors affecting word learning, we were interested in effect sizes as small as d=.5.
Resource limitation	We were also limited by our ability to collect data only at our on-campus preschool.
Power computation	We calculated that based on our maximal possible sample size of N=120 (60 per group), we would achieve at least 80% power to reject the null for effects as small as d = .52.

Depth

Sample sizes for replication studies

Setting the sample size for a replication study has been a persistent issue in the meta-science literature. Naïvely speaking, it seems like you should be able to compute the effect size for the original study and then simply use that as the basis for a classical power analysis.

This naïve approach has several flaws, however. First, the effect size from the original published paper is likely an overestimate of the true effect size due to publication bias (Nosek et al. 2021). Second, the power analysis will only yield the sample size at which the replication will have a particular chance of rejecting the null at some criterion. But it's quite possible that the original experiment could be $p < .05$, the replication could be $p > .05$, *and* 3) the original experiment and the replication results are not significantly different from each other. So a statistically significant replication of the original effect size is not necessarily what you want to aim for.

Faced with these issues, a replication sample size can be planned in several other ways. First, replicators can use standard strategies above such as SESOI or resource-based planning to rule out large effects, either with high probability or within a known amount of time or money. If the SESOI is high or limited resources are allocated, these strategies can produce an inconclusive result, however. A conclusive answer can require a very substantial commitment of resources.

Second, Simonsohn (2015) recommends the "small telescopes" approach. The idea is not to test whether there *is* an effect, but rather where there is an effect *large enough that the original study could have detected it*. The analogy is to astronomy. If a birdwatcher points their binoculars at the sky and claims to have discovered a new planet, we want to ask not just whether there is a planet at that location, but also whether there is any possibility that they could

have seen it using binoculars – if not, perhaps they are right but for the wrong reasons! Simonsohn shows that, if a replicator collects 2.5 times as large a sample as the original, they have 80% power to detect any effect that was reasonably detectable by the original. This simple rule of thumb provides one good starting place for conservative replication studies.

Finally, replicators can make use of sequential Bayesian analysis, in which they attempt to gather substantial evidence relative to the support for H_1 *or* H_0. Sequential bayes is an appealing option because it allows for efficient collection of data that reflects whether an effect is likely to be present in a particular sample, especially in the face of the sometimes prohibitively large samples necessary for SESOI or "small telescopes" analyses.

10.5 Chapter summary: Sampling

Your goal as an experimenter is to estimate a causal effect. But the effect for whom? This chapter has tried to help you think about how you generalize from your experimental sample to some target population. It's very rare to be conducting an experiment based on a probability sample in which every member of the population has an equal chance of being selected. In the case that you are using a convenience sample, you will need to consider how bias introduced by the sample could relate to the effect estimate you observed. Do you think this effect is likely to be very heterogeneous in the population? Are there theories that suggest that it might be larger or smaller for the convenience sample you recruited?

Questions about generalizability and sampling depend on the precise construct you are studying, and there is no mechanistic procedure for answering them. Instead, you simply have to ask yourself: how does my sampling procedure qualify the inference I want to make based on my data? Being transparent about your reasoning can be very helpful – both to you and to readers of your work who want to contextualize the generality of your findings.

Discussion questions

1. We want to understand human cognition generally, but do you think it is a more efficient research strategy to start by studying certain features of cognition (perception, for example) in WEIRD convenience populations and then later check our generalizations in non-WEIRD groups? What are the arguments against this efficiency-based strategy?
2. One alternative position regarding sampling is that the most influential experiments aren't generalizations of some number to a population; they are demonstration experiments that show that some particular effect is possible under some circumstances (think Milgram's conformity studies, see Chapter 4). On this argument, the specifics of population sampling are often secondary. Do you think this position makes sense?
3. One line of argument says that we can't ever make generalizations about the human mind because so much of the

historical human population is simply inaccessible to us (we can't do experiments on ancient Greek psychology). In other words, sampling from a particular population is *also* sampling a particular moment in time. How should we qualify our research interpretations to deal with this issue?

Readings

- The original polemic article on the WEIRD problem: Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The WEIRDest people in the world? *Behavioral and Brain Sciences,* 33, 61–83.
- A very accessible introduction to power analysis from its originator: Cohen, J. (1992) A power primer. *Psychological Bulletin,* 112, 155–9.
- A thoughtful and in-depth discussion of generalizability issues: Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences,* 45, 1–37.

Part IV

Execution

References

Bland, John Martin. 2009. "The Tyranny of Power: Is There a Better Way to Calculate Sample Size?" *British Medical Journal* 339 (October): b3985.

Browne, Richard H. 1995. "On the Use of a Pilot Sample for Sample Size Determination." *Statistics in Medicine* 14 (17): 1933–40.

Champely, Stephane, Claus Ekstrom, Peter Dalgaard, Jeffrey Gill, Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic, and Helios De Rosario. 2017. "Pwr: Basic Functions for Power Analysis."

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112 (1): 155.

———. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Coppock, Alexander, Thomas J Leeper, and Kevin J Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates Across Samples." *Proceedings of the National Academy of Sciences* 115 (49): 12441–46.

Cunningham, Scott. 2021. *Causal Inference*. Yale University Press.

DeJesus, Jasmine M, Maureen A Callanan, Graciela Solis, and Susan A Gelman. 2019. "Generic Language in Scientific Communication." *Proceedings of the National Academy of Sciences* 116 (37): 18370–77.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G* Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods* 39 (2): 175–91.

Green, Peter, and Catriona J MacLeod. 2016. "SIMR: An r Package for Power Analysis of Generalized Linear Mixed Models by Simulation." *Methods in Ecology and Evolution* 7 (4): 493–98.

Hedge, Craig, Georgina Powell, and Petroc Sumner. 2018. "The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences." *Behavior Research Methods* 50 (3): 1166–86.

Henrich, Joseph, Steven J Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2-3): 61–83.

Kelley, Ken, Francis Bilson Darku, and Bhargab Chattopadhyay. 2018. "Accuracy in Parameter Estimation for a General Class of Effect Sizes: A Sequential Approach." *Psychological Methods* 23 (2): 226.

Klein, Olivier, Tom E Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsonne, Wolf Vanpaemel, and Michael C Frank. 2018. "A Practical Guide for Transparency in Psychological Science."

Lakens, Daniël. 2014. "Performing High-Powered Studies Efficiently with Sequential Analyses." *European Journal of Social Psychology* 44 (7): 701–10.

Lakens, Daniël, Anne M. Scheel, and Peder M. Isager. 2018. "Equivalence Testing for Psychological Research: A Tutorial." *Advances in Methods and Practices in Psychological Science* 1 (2): 259–69. https://doi.org/10.1177/2515245918770963.

Little, Roderick JA, and Donald B Rubin. 2019. *Statistical Analysis with Missing Data*. Vol. 793. John Wiley & Sons.

Majid, Asifa, and Niclas Burenhult. 2014. "Odors Are Expressible in Language, as Long as You Speak the Right Language." *Cognition* 130 (2): 266–70.

Majid, Asifa, and Nicole Kruspe. 2018. "Hunter-Gatherer Olfaction Is Special." *Current Biology* 28 (3): 409–13.

Markus, Hazel R, and Shinobu Kitayama. 1991. "Culture and the Self: Implications for Cognition, Emotion, and Motivation." *Psychological Review* 98 (2): 224.

Neyman, Jerzy. 1992. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." In *Breakthroughs in Statistics*, 123–50. Springer.

Nosek, Brian A, Tom E Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S Corker, Anna Dreber Almenberg, Fiona Fidler, et al. 2021. "Replicability, Robustness, and Reproducibility in Psychological Science." *Annual Review of Psychology*.

Nunan, David, Jeffrey Aronson, and Clare Bankhead. 2018. "Catalogue of Bias: Attrition Bias." *Evidence Based Medicine* 23 (1): 21–22.

Piantadosi, Steven T, and Edward Gibson. 2014. "Quantitative Standards for Absolute Linguistic Universals." *Cognitive Science* 38 (4): 736–56.

Rohrer, Julia M. 2018. "Thinking Clearly about Correlations and Causation: Graphical Causal Models for Observational Data." *Advances in Methods and Practices in Psychological Science* 1 (1): 27–42.

Rosenthal, Robert, and Ralph L Rosnow. 1984. *Essentials of Behavioral Research: Methods and Data Analysis*. New York: McGraw-Hill.

Rothman, Kenneth J, and Sander Greenland. 2018. "Planning Study Size Based on Precision Rather Than Power." *Epidemiology* 29 (5): 599–603.

Schönbrodt, Felix D, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. 2017. "Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences." *Psychol. Methods* 22 (2): 322–39.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66.

Simonsohn, Uri. 2015. "Small Telescopes: Detectability and the Evaluation of Replication Results." *Psychol. Sci.* 26 (5): 559–69.

Syed, Moin, and U Kathawalla. 2020. "Cultural Psychology, Diversity, and Representation in Open Science." In *Cultural Methods in Psychology: Describing and Transforming Cultures*, edited by Kate C McLean, 427–54. Oxford University Press.

Tessler, Michael Henry, and Noah D Goodman. 2019. "The Language of Generalization." *Psychological Review* 126 (3): 395.

Tsai, Jeanne L. 2007. "Ideal Affect: Cultural Causes and Behavioral Consequences." *Perspectives on Psychological Science* 2 (3): 242–59.

Westfall, Jacob, Charles M Judd, and David A Kenny. 2015. "Replicating Studies in Which Samples of Participants Respond to Samples of Stimuli." *Perspectives on Psychological Science* 10 (3): 390–99.

Yarkoni, Tal. 2020. "The Generalizability Crisis." *Behavioral and Brain Sciences* 45: 1–37.

Yeager, David S, Paul Hanselman, Gregory M Walton, Jared S Murray, Robert Crosnoe, Chandra Muller, Elizabeth Tipton, et al. 2019. "A National Experiment Reveals Where a Growth Mindset Improves Achievement." *Nature* 573 (7774): 364–69.

Zhou, Haotian, and Ayelet Fishbach. 2016. "The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (yet False) Research Conclusions." *Journal of Personality and Social Psychology* 111 (4): 493.

11 Preregistration

Learning goals

- Recognize the dangers of researcher degrees of freedom
- Understand the differences between exploratory and confirmatory modes of research
- Articulate how preregistration can reduce risk of bias and increase transparency

> When not planned beforehand, data analysis can approximate a projective technique, such as the Rorschach, because the investigator can project on the data his own expectancies, desires, or biases and can pull out of the data almost any "finding" he may desire.
>
> — Theodore X. Barber (1976)

> The first principle is that you must not fool yourself–and you are the easiest person to fool… After you've not fooled yourself, it's easy not to fool other scientists. You just have to be honest in a conventional way after that.
>
> — Richard Feynman (1974)

The last section of the book focused on planning a study – in particular, making decisions around measurement, design, and sampling. In this next section, we turn to the nuts and bolts of executing a study. We start with preregistration (this chapter), before discussing the logistics of data collection (Chapter 12) and project management (Chapter 13). These chapters touch on the themes of *transparency* and *bias reduction* through decisions about how to document and organize your data collection.

Let's start with simply documenting choices about design and analysis. Although there are plenty of *incorrect* ways to design and analyse experiments, there is no single *correct* way. In fact, most research decisions have many justifiable choices – sometimes called "researcher degrees of freedom". For example, will you stop data collection after 20, 200, or 2000 participants? Will you remove outlier values and how will you define them? Will you conduct subgroup analyses to see whether the results are affected by sex, or age, or some other factor?

Consider a simplified, hypothetical case where you have to make five analysis decisions and there are five justifiable choices for each decision — this alone would result in 3125 (5^5) unique ways to analyze the data! If you were to make these decisions **post hoc** (after observing the data) then there's a danger your decisions will be influenced by the outcome of the analysis ("data-dependent decision making") and skew towards choices that generate outcomes more aligned with your personal preferences. Now think back to the last time you read a research paper. Of all the possible ways that the data could have been analyzed, how do you know that the researchers did not just select the approach that generated results most favourable to their pet hypothesis?

In this chapter, we will find out why flexibility in the design, analysis, reporting, and interpretation of experiments, combined with data-dependent decision-making, can introduce bias, and lead to scientists fooling themselves and each other. We will also learn about **preregistration**, the process of writing down and registering your research decisions before you observe the data. Preregistration intersects with two of our themes: it can be used to REDUCE BIAS in our data analysis, and it can provide the TRANSPARENCY that other scientists need to properly evaluate and interpret our results (Hardwicke and Wagenmakers 2022).

CASE STUDY

Undisclosed analytic flexibility?

Educational apps for children are a huge market, but relatively few randomized trials have been done to see whether or when they produce educational gains. Filling this important gap, Berkowitz et al. (2015) reported a high-quality field experiment of a free educational app, "Bedtime Math at Home," with participants randomly assigned to either math or reading conditions over the course of a full school year. Critically, along with random assignment, the study also included standardized measures of math and reading achievement. These measures allowed the authors to compute effects in grade-level equivalents, a meaningful unit from a policy perspective.

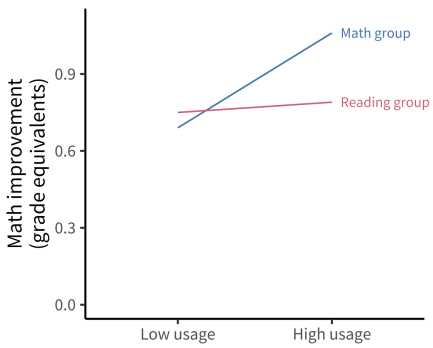


Figure 11.1: Model fits reported in Figure 1 of Berkowitz et al. (2015). Estimated years of math achievement gained over the school year across groups, as a function of app usage level.

The key result is shown in Figure 11.1. Families who used the math app frequently showed greater gains in math than the control group. Although this finding appeared striking, the figure didn't directly visualize the primary causal effect of interest, namely the size of the effect of study condition on math scores. Instead the data were presented as estimated effects for specific levels of app usage.

Because the authors made their data openly available, it was possible for Frank (2016) to do a simple analysis to examine the causal effect of interest. When not splitting the data by usage and adjusting by covariates, there was no significant main effect of the intervention on math performance Figure 11.2. Since this analysis was not favorable to the primary intervention – and because it was not reported in the paper – it could have been the case that the authors had analyzed the data several ways and chosen to present an analysis that was more favorable to their hypotheses of interest.

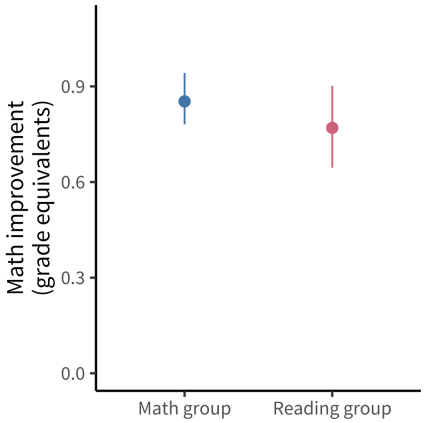


Figure 11.2: Estimated years of math achievement gained over the school year across groups in the Berkowitz et al. (2016) math app trial. Error bars show bootstrapped 95% confidence intervals. Based on Frank (2016).

As is true for many papers prior to the rise of preregistration, it's not possible to know definitively whether the reported analysis in Berkowitz et al. (2015) was influenced by the authors' desired result. As we'll see below, such data-dependent analyses can lead to substantial bias in reported effects. This uncertainty about a paper's analytic strategy can be avoided by the use of preregistration. In this case, preregistration would have convinced readers that the analyses decisions were not influenced by the data, thereby increasing the value of this otherwise high-quality study.

11.1 Lost in a garden of forking paths

One way to visualize researcher degrees of freedom is as a vast decision tree or "garden of forking paths" (Figure 11.3). Each node represents a decision point and each branch represents a justifiable choice. Each unique pathway through the garden terminates in an individual research outcome.

Because scientific observations typically consist of both noise (random variation unique to this sample) and signal (regularities that will reoccur in other samples), some of these pathways will inevitably lead to results that are misleading (e.g., inflated effect sizes, exaggerated evidence, or false positives). The more potential paths in the garden that you might explore, the higher the chance of encountering misleading results.

Statisticians refer to this issue as a **multiplicity** (multiple comparisons) problem. As we talked about in Chapter 6, multiplicity can be addressed to some extent with statistical countermeasures, like the Bonferroni correction; however, these adjustment methods need to account for every path that you *could have* taken (Gelman and Loken 2014; de Groot 1956/2014). When you navigate the garden of forking paths while working with the data, it is easy to forget – or even be unaware of – every path that you could have taken, so these methods can no longer be used effectively.

The signal-to-noise ratio is worse in particular situations (as common in psychology) with small effect sizes, high variation, and large measurement errors (Ioannidis 2005). Researcher degrees of freedom may be constrained to some extent by strong theory (Oberauer and Lewandowsky 2019), community methodological norms, or replication studies, though these constraints may be more implicit than explicit, and can still leave plenty of room for flexible decision-making.

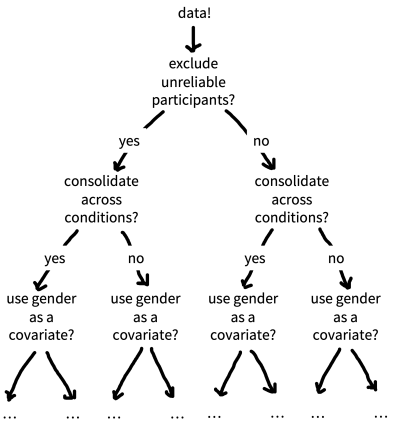


Figure 11.3: The garden of forking paths: many justifiable but different analytic choices are possible for an individual dataset.

11.1.1 Data-dependent analysis

When a researcher navigates the garden of forking paths during data analysis, their choices might be influenced by the data (**data-dependent** decision making) which can introduce bias. If a researcher is seeking a particular kind of result (see Depth box below), then they are more likely to follow the branches that steer them in that direction.

You could think of this a bit like playing a game of "hot (🔥) or cold (❄)" where 🔥 indicates that the choice will move the researcher closer to a desirable overall result and ❄ indicates that the choice will move them further away. Each time the researcher reaches a decision point, they try one of the branches and get feedback on how that choice affects the results. If the feedback is 🔥 then they take that branch. If the answer is ❄, they try a different branch. If they reach the end of a complete pathway, and the result is ❄, maybe they even retrace their steps and try some different branches earlier in the pathway. This strategy creates a risk of bias because it systematically skews results towards researcher's preferences (Hardwicke and Wagenmakers 2022).[1]

[1] We say "risk of bias" rather than just "bias" because in most scientific contexts, we do not have a known ground truth to compare the results to. So in any specific situation, we do not know the extent to which data-dependent decisions have actually biased the results.

Depth

Only human: Cognitive biases and skewed incentives

There's a storybook image of the scientist as an objective, rational, and dispassionate arbiter of truth (Veldkamp et al. 2017). But in reality, scientists are only human: they have egos, career ambitions, and rent to pay! So even if we do want to live up to the storybook image, its important to acknowledge that our decisions and behavior are also influenced by a range of cognitive biases and external incentives that can steer us away from that goal. Let's first look at some relevant cognitive biases that might lead scientists astray:

- **Confirmation bias**: Preferentially seeking out, recalling, or evaluating information in a manner that reinforces one's existing beliefs (Nickerson 1998).
- **Hindsight bias**: Believing that past events were always more likely to occur relative to our actual belief in their likelihood before they happened ("I knew it all along!") (Slovic and Fischhoff 1977).
- **Motivated reasoning**: Rationalizing prior decisions so they are framed in a favorable light, even if they were irrational (Kunda 1990).
- **Apophenia**: Detecting seemingly meaningful patterns in noise (Gilovich, Vallone, and Tversky 1985).

To make matters worse, the incentive structure of the scientific ecosystem often adds additional motivation to get things wrong. The allocation of funding, awards, and publication prestige is often based on the nature of research results rather than research quality (Smaldino and McElreath 2016; Nosek, Spies, and Motyl 2012). For example, many academic journals, especially those that are widely considered to be the most prestigious, appear to have a preference for novel, positive, and statistically significant results over incremental, negative, or null results (Bakker, Dijk, and Wicherts 2012). There is also pressure to write articles with concise, coherent, and compelling narratives (Giner-Sorolla 2012). This set of forces incentivizes scientists to be "impressive" over being right and encourages questionable research practices. The process of iteratively p-hacking and HARKing one's way to a "beautiful" scientific paper has been dubbed "The Chrysalis Effect" (O'Boyle, Banks, and Gonzalez-Mulé 2017), illustrated in Figure 11.4.

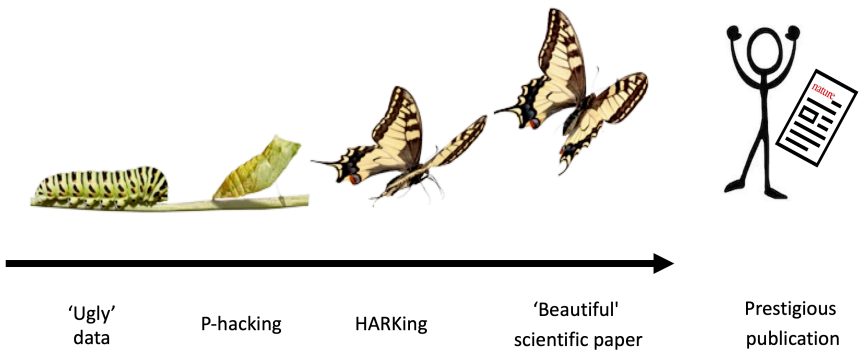


Figure 11.4: The Chrysalis Effect, when ugly truth becomes a beautiful fiction.

In sum, scientists' human flaws – and the scientific ecosystem's flawed incentives – highlight the need for transparency and intellectual humility when reporting the findings of our research (Hoekstra and Vazire 2021).

In the most egregious cases, a researcher may try multiple pathways until they obtain a desirable result and then **selectively report** that result, neglecting to mention that they have tried several other analysis strategies (also known as p-hacking, a practice we've discussed throughout the book).[2] You may remember an example of this practice in Chapter 3, where participants apparently became younger when they listened to "When I'm 64" by The Beatles. Another example of how damaging the garden of forking paths can be comes from the "discovery" of brain activity in a dead Atlantic Salmon! Researchers deliberately exploited flexibility in the fMRI analysis pipeline and avoided multiple comparisons corrections, allowing them to find brain activity where there was only dead fish Figure 11.5.

[2] "If you torture the data long enough, it will confess" (Good 1972).

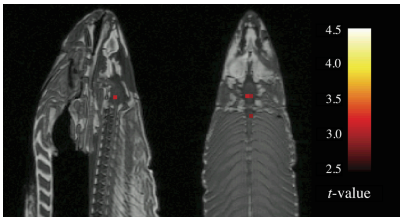


Figure 11.5: By deliberately exploiting analytic flexibility in the processing pipeline of fMRI data, Bennett, Miller, and Wolford (2009) were able to identify 'brain activity' in a dead Atlantic Salmon. From Bennett, Miller, and Wolford (2009) (licensed under CC BY).

11.1.1 Hypothesizing after results are known

In addition to degrees of freedom in experimental design and analysis, there is additional flexibility in how researchers *interpret* research results. As we discussed in Chapter 2, theories can accommodate even conflicting results in many different ways – for example, by positing auxiliary hypotheses that explain why a particular datapoint is special.

The practice of selecting or developing your hypothesis after observing the data has been called "Hypothesizing After the Results are Known", or "HARKing" (Kerr 1998). HARKing is potentially problematic because it expands the garden of forking paths and helps to justify the use of various additional design and analysis decisions (Figure 11.6). For example, you may come up with an explanation for why an intervention is effective in men but not in women in order to justify a post-hoc subgroup analysis based on sex (see Case Study. The extent to which HARKing is problematic is contested (for discussion see Hardwicke and Wagenmakers 2022). But at the very least it's important to be honest about whether hypotheses were developed before or after observing the data.

But hang on a minute! Isn't it a good thing to seek out interesting results if they are there in the data? Shouldn't we "let the data speak"? The answer is yes! But it's crucial to understand the distinction between **exploratory** and **confirmatory** modes of research.[3] Confirmation involves making research decisions *before* you've seen the data whereas exploration involves making research decisions *after* you've seen data.

The key things to remember about exploratory research are that you need to (1) be aware of the increased risk of bias arising from data-dependent decision making and calibrate your confidence in the results accordingly; (2) be honest with other researchers about your analysis

[3] In practice, an individual study may contain both exploratory and confirmatory aspects which is why we describe them as different "modes."

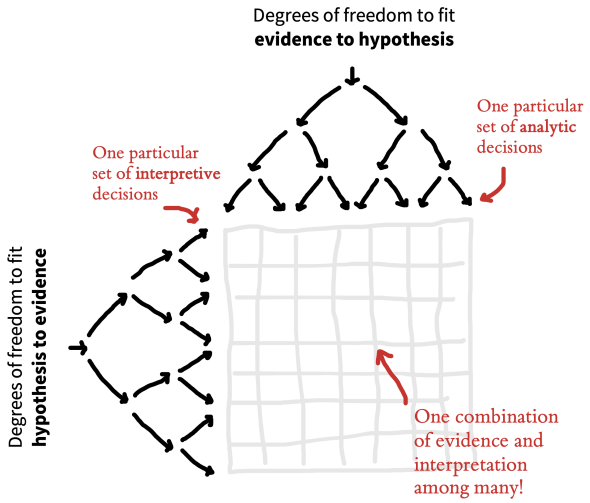


Figure 11.6: A grid of individual research results. The horizontal axis provides a simplified illustration of the many justifiable design and analysis choices that a scientist can use to generate the evidence. The vertical axis illustrates that there are often several potential hypotheses which could be constructed or selected when interpreting the evidence. An unconstrained scientist can simultaneously fit evidence to hypotheses and fit hypotheses to evidence in order to obtain their preferred overall result.

strategy so they are also aware of the risk of bias and can calibrate *their* confidence in the outcomes accordingly. In the next section, we will learn about how preregistration helps us to make this important distinction between exploratory and confirmation research.

11.2 Reducing risk of bias, increasing transparency, and calibrating confidence with preregistration

You can counter the problem of researcher degrees of freedom and data-dependent decision-making by making research decisions before you have seen the data – like planning your route through the garden of forking paths before you start your journey (Wagenmakers et al. 2012; Hardwicke and Wagenmakers 2022). If you stick to the planned route, then you have eliminated the possibility that your decisions were influenced by the data.

Preregistration is the process of declaring your research decisions in a public registry before you analyze (and often before you collect) the data. Preregistration ensures that your research decisions are data-independent, which reduces risk of bias arising from the issues described above. Preregistration also transparently conveys to others what you planned, helping them to determine the risk of bias and calibrate their confidence in the research results. In other words, preregistration can dissuade researchers from engaging in questionable research practices like p-hacking and HARKing, because they can be held accountable to their original plan, while also providing the context needed to properly evaluate and interpret research.

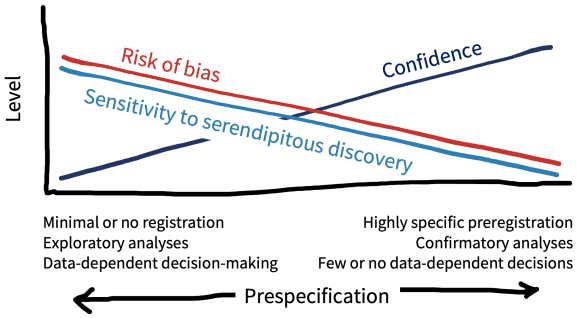


Figure 11.7: Preregistration clarifies where research activities fall on the continuum of prespecification. When the preregistration provides little constraint over researcher degrees of freedom (i.e., more exploratory research), decisions are more likely to be data-dependent, and consequently there is a higher risk of bias. When preregistration provides strong constraint over researcher degrees of freedom (i.e., more confirmatory research), decisions are less likely to be data-dependent, and consequently there is a lower risk of bias. Exploratory research activities are more sensitive to serendipitous discovery, but also have a higher risk of bias relative to confirmatory research activities. Preregistration transparently communicates where particular results are located along the continuum, helping readers to appropriately calibrate their confidence.

Preregistration does not require that you specify all research decisions in advance, only that you are transparent about what was planned, and what was not planned. This transparency helps to make a distinction between which aspects of the research were exploratory and which were confirmatory (Figure 11.7). All else being equal, we should have more confidence in confirmatory results, because there is a lower risk of bias. Exploratory results have a higher risk of bias, but they are also more **sensitive** to serendipitous (unexpected discoveries. So the confirmatory mode is best suited to testing hypotheses and the exploratory mode is best suited to generating them. Therefore, exploratory and confirmatory research are both valuable activities – it is just important to differentiate them (Tukey 1980)! Preregistration offers the best of both worlds by clearly separating one from the other.

In addition to the benefits described above, preregistration may improve the quality of research by encouraging closer attention to study planning. We've found that the process of writing a preregistration really helps facilitate communication between collaborators, and can catch addressable problems before time and resources are wasted on a poorly designed study. Detailed advanced planning can also create opportunities for useful community feedback, particularly in the context of Registered Reports (see Depth box below), where dedicated peer reviewers will evaluate your study before it has even begun.

Depth

Preregistration and friends: A toolbox to address researcher degrees of freedom

Several useful tools can be used to complement or extend preregistration. In general, we would recommend that these tool are combined with preregistration, rather than used as a replacement because preregistration provides transparency about the research and planning process (Hardwicke and Wagenmakers 2022). The first two of these are discussed in more detail in the last section of Chapter 7.

Robustness checks. Robustness checks (also called "sensitivity analyses") assess how different decision choices in the garden of forking paths affect the eventual pattern of results. This technique is particularly helpful when you have to choose between several justifiable analytic choices, neither of which seem superior to the other, or which have complementary strengths and weaknesses. For example, you might run the analysis three times using three different methods for handling missing data. Robust results should not vary substantially across the three different choices.

Multiverse analyses. Recently, some researchers have started running large-scale robustness checks called "multiverse" (Steegen et al. 2016) or "specification curve" (Simonsohn, Simmons, and Nelson 2020) analyses. We discussed these a bit in Chapter 7. Some have argued that these large-scale robustness checks make preregistration redundant; after all, why prespecify a single path if you can explore them all (Rubin 2020; Oberauer and Lewandowsky 2019)? But interpreting the results of a multiverse analysis are not straightforward; for example, it seems unlikely that all of the decision choices are equally justifiable (Giudice and Gangestad 2021). Furthermore, if multiverse analyses are not preregistered, then they introduce researcher degrees of freedom, and create an opportunity for selective reporting, which increases risk of bias.

Held-out sample. One option to benefit from both exploratory and confirmatory research modes is to split your data into **training** and **test** samples. (The test sample is commonly called the "held out" because it is "held out" from the exploratory process.) You can generate hypotheses in an exploratory mode in the training sample and use that as the basis to preregister confirmatory analyses in the hold-out sample. A notable disadvantage of this strategy is that splitting the data reduces statistical power, but in cases where data are plentiful – including in much of machine learning – this technique is the gold standard.

Masked analysis (traditionally called "blind analysis"). Sometimes problems, such as missing data, attrition, or randomization failure that you did not anticipate in your preregistered plan can arise during data collection. How do you diagnose and address these issues without increasing risk of bias through data-dependent analysis? One option is masked analysis, which disguises key aspects of the data related to the results (for example, by shuffling condition labels or adding noise) while still allowing some degree of data inspection (Dutilh, Sarafoglou, and Wagenmakers 2019). After diagnosing a problem, you can adjust your preregistered plan without increasing risk of bias, because your decisions have not been influenced by the results.

Standard Operating Procedures. Community norms, perhaps at the level of your research field or lab, can act as a natural constraint on researcher degrees of freedom. For example, there may be a generally accepted approach for handling outliers in your community. You can make these constraints explicit by writing them down in a Standard Operating Procedures document – a bit like a living meta-preregistration (Lin and Green 2016).

Open lab notebooks. Maintaining a lab notebook can be a useful way to keep a record of your decisions as a research project unfolds. Preregistration is bit like taking a snapshot of your lab notebook at the start of the project, when all you have written down is your research plan. Making your lab notebook publicly available is a great way to transparently document your research and departures from the preregistered plan.

Figure 11.8: Registered Reports (from https://www.cos.io/initiatives/registered-reports, licensed under CC BY 4.0).

Registered Reports. Registered Reports are a type of article format that embeds preregistration directly into the publication pipeline , Figure 11.8. The idea is that you submit your preregistered protocol to a journal and it is peer reviewed, before you've even started your study. If the study is approved, the journal agrees to publish it, regardless of the results. This is a radical departure from traditional publication models where peer reviewers and journals evaluate your study *after* its been completed and the results are known. Because the study is accepted for publication independently of the results, Registered Reports can offer the benefits of preregistration with additional protection against publication bias. They also provide a great opportunity to obtain feedback on your study design while you can still change it!

11.3 How to preregister

High-stakes studies such as medical trials must be preregistered (Dickersin and Rennie 2012). In 2005, a large international consortium of medical journals decided that they would not publish unregistered trials. The discipline of economics also has strong norms about study registration (see e.g. https://www.socialscienceregistry.org). But preregistration is pretty new to psychology (Nosek et al. 2018), and there's still no standard way of doing it – you're already at the cutting edge!

We recommend using the Open Science Framework (OSF) as your registry. OSF is one of the most popular registries in psychology and you can do lots of other useful things on the platform to make your research transparent, like sharing data, materials, analysis scripts, and preprints. On OSF it is possible to "register" any file you have uploaded. When you register a file, it creates a time-stamped, read-only copy, with a dedicated link. You can add this link to articles reporting your research.

Table 11.1: Preregistration template outline.

	Question
1	Data collection. Have any data been collected for this study already?
2	Hypothesis. What's the main question being asked or hypothesis being tested in this study?
3	Dependent variable. Describe the key dependent variable(s) specifying how they will be measured.
4	Conditions. How many and which conditions will participants be assigned to?
5	Analyses Specify exactly which analyses you will conduct to examine the main question/hypothesis.
6	Outliers and Exclusions. Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
7	Sample Size. How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
8	Other. Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?

One approach to preregistration is to write a protocol document that specifies the study rationale, aims or hypotheses, methods, and analysis plan, and register that document.[4] OSF also has a collection of dedicated preregistration templates that you can use if you prefer. An outline of such a template is shown in Table 11.1. These templates are often tailored to the needs of particular types of research. For example, there are templates for general quantitative psychology research ("PRP-QUANT," Bosnjak et al. 2022), cognitive modelling (Crüwell and Evans 2021), and secondary data analysis (Akker et al. 2019). The OSF interface may change, but currently this guide[5] provides a set of steps to create a preregistration.

Once you've preregistered your plan, you just go off and run the study and report the results, right? Well hopefully… but things might not turn out to be that straightforward. It's quite common to forget to include something in your plan or to have to depart from the plan due to something unexpected. Preregistration can actually be pretty hard in practice (Nosek et al. 2019)!

Don't worry though - remember that a key goal of preregistration is transparency to enable others to evaluate and interpret research results. If you decide to depart from your original plan and conduct data-dependent analyses, then this decision may increase the risk of bias. But if you communicate this decision transparently to your readers, they can appropriately calibrate their confidence in the results. You may even be able to run both the planned and unplanned analyses

[4] You can think of a study protocol a bit like a research paper without a results and discussion section (here's an example from one of our own studies: https://osf.io/2cnkq/).

[5] https://help.osf.io/hc/en-us/articles/360019738834-Create-a-Preregistration

as a robustness check (see Depth box) to evaluate the extent to which this particular choice impacts the results.

When you report your study, it is important to distinguish between what was planned and what was not. If you ran a lot of data-dependent analyses, then it might be worth having separate exploratory and confirmatory results sections. On the other hand, if you mainly stuck to your original plan, with only minor departures, then you could include a table (perhaps in an appendix) that outlines these changes (for example, see Supplementary Information A of this article[6]).

[6] https://doi.org/10.31222/osf.io/wt5ny

11.4 Chapter summary: Preregistration

We've advocated here for preregistering your study plan. This practice helps to reduce the risk of bias caused by data-dependent analysis (the "garden of forking paths" that we described) and transparently communicate the risk of bias to other scientists. Importantly, preregistration is a "plan, not a prison[7]": in most cases preregistered, confirmatory analyses coexist with exploratory analyses. Both are an important part of good research – the key is to disclose which is which!

[7] https://www.cos.io/blog/preregistration-plan-not-prison

Discussion questions

1. P-hack your way to scientific glory! To get a feel for how data-dependent analyses might work in practice, have a play around with this app: https://projects.fivethirtyeight.com/p-hacking/. Do you think preregistration would affect your confidence in claims made about this dataset?
2. Preregister your next experiment! The best way to get started with preregistration is to have a go with your next study. Head over to https://osf.io/registries/osf/new and register your study protocol or complete one of the templates. What aspects of preregistration did you find most difficult and what benefits did it bring?

Readings

– Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*, 2600–2606. https://doi.org/10.1073/pnas.1708274114.

– Hardwicke, T. E., & Wagenmakers, E.-J. (2023). Reducing bias, increasing transparency, and calibrating confidence with preregistration. *Nature Human Behaviour*. 7, 15-26, https://doi.org/10.1038/s41562-022-01497-2.

References

Akker, Olmo van den, Sara J. Weston, Lorne Campbell, William J. Chopik, Rodica I. Damian, Pamela Davis-Kean, Andrew Hall, et al. 2019. "Preregistration of Secondary Data Analysis: A Template and Tutorial." PsyArXiv. https://psyarxiv.com/hvfmr/.

Bakker, Marjan, Annette van Dijk, and Jelte M. Wicherts. 2012. "The Rules of the Game Called Psychological Science." *Perspectives on Psychological Science* 7 (6): 543–54. https://doi.org/10.1177/1745691612459060.

Barber, Theodore Xenophon. 1976. *Pitfalls in Human Research: Ten Pivotal Points*. Pergamon General Psychology Series ; v. 67. New York: Pergamon Press.

Bennett, CM, MB Miller, and GL Wolford. 2009. "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument for Multiple Comparisons Correction." *NeuroImage*, Organization for Human Brain Mapping 2009 Annual Meeting, 47 (July): S125. https://doi.org/10.1016/S1053-8119(09)71202-9.

Berkowitz, Talia, Marjorie W. Schaeffer, Erin A. Maloney, Lori Peterson, Courtney Gregor, Susan C. Levine, and Sian L. Beilock. 2015. "Math at Home Adds up to Achievement in School." *Science* 350 (6257): 196–98. https://doi.org/10.1126/science.aac7427.

Berkowitz, Talia, Marjorie W Schaeffer, Christopher S Rozek, Erin A Maloney, Susan C Levine, and Sian L Beilock. 2016. "Response to Comment on 'Math at Home Adds up to Achievement in School'." *Science* 351 (6278): 1161–61.

Bosnjak, Michael, Christian Fiebach, David Thomas Mellor, Stefanie Mueller, Daryl O'Connor, Fred Oswald, and Rose Sokol-Chang. 2022. "A Template for Preregistration of Quantitative Research in Psychology: Report of the Joint Psychological Societies Preregistration Task Force." *American Psychologist* 77 (4): 602–15. https://doi.org/10.1037/amp0000879.

Chambers, Chris, and Loukia Tzavella. 2020. "Registered Reports: Past, Present and Future." MetaArXiv. https://doi.org/10.31222/osf.io/43298.

Crüwell, Sophia, and Nathan J. Evans. 2021. "Preregistration in Diverse Contexts: A Preregistration Template for the Application of Cognitive Models." *Royal Society Open Science* 8 (10): 210155. https://doi.org/10.1098/rsos.210155.

de Groot, A. D. 1956/2014. "The Meaning of 'Significance' for Different Types of Research." Translated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier A. Kievit, Marjan Bakker, Angélique O. J. Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas. *Acta Psychologica* 148 (1956/2014): 188–94. https://doi.org/10.1016/j.actpsy.2014.02.001.

Dickersin, Kay, and Drummond Rennie. 2012. "The Evolution of Trial Registries and Their Use to Assess the Clinical Trial Enterprise." *JAMA* 307 (17): 1861–64. https://doi.org/10.1001/jama.2012.4230.

Dutilh, Gilles, Alexandra Sarafoglou, and Eric-Jan Wagenmakers. 2019. "Flexible yet Fair: Blinding Analyses in Experimental Psychology." *Synthese*, August. https://doi.org/https://doi.org/10.1007/s11229-019-02456-7.

Feynman, Richard P. 1974. "Cargo Cult Science." http://calteches.library.caltech.edu/51/2/CargoCult.pdf.

Frank, Michael C. 2016. "Comment on 'Math at Home Adds up to Achievement in School'." *Science*.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102 (6): 460–65. https://doi.org/10.1511/2014.111.460.

Gilovich, Thomas, Robert Vallone, and Amos Tversky. 1985. "The Hot Hand in Basketball: On the Misperception of Random Sequences." *Cognitive Psychology* 17 (3): 295–314. https://doi.org/10.1016/0010-0285(85)90010-6.

Giner-Sorolla, Roger. 2012. "Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science." *Perspectives on Psychological Science* 7 (6): 562–71. https://doi.org/10.1177/1745691612457576.

Giudice, M Del, and SW Gangestad. 2021. "A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions." *Advances in Methods and Practices in Psychological Science* 4 (1): 1–15. https://doi.org/https://doi.org/10.1177/2515245920954925.

Good, I. J. 1972. "Statistics and Today's Problems." *The American Statistician* 26 (3): 11–19. https://doi.org/10.1080/00031305.1972.10478922.

Hardwicke, Tom E, and Eric-Jan Wagenmakers. 2022. "Reducing Bias, Increasing Transparency, and Calibrating Confidence with Preregistration." *Nature Human Behaviour*. https://doi.org/10.31222/osf.io/d7bcu.

Hoekstra, Rink, and Simine Vazire. 2021. "Aspiring to Greater Intellectual Humility in Science." *Nature Human Behaviour* 5 (12): 1602–7.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Kerr, Norbert L. 1998. "HARKing: Hypothesizing After the Results Are Known." *Personality & Social Psychology Review (Lawrence Erlbaum Associates)* 2 (3): 196. https://doi.org/10.1207/s15327957pspr0203_4.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108 (3): 480–98. https://doi.org/10.1037/0033-2909.108.3.480.

Lin, Winston, and Donald P. Green. 2016. "Standard Operating Procedures: A Safety Net for Pre-Analysis Plans." *PS: Political Science & Politics* 49 (03): 495–500. https://doi.org/10.1017/S1049096516000810.

Nickerson, Raymond S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2 (2): 175–220. https://doi.org/10.1037/1089-2680.2.2.175.

Nosek, Brian A, Emorie D. Beck, Lorne Campbell, Jessica K. Flake, Tom E. Hardwicke, David T. Mellor, Anna E. van 't Veer, and Simine Vazire. 2019. "Preregistration Is Hard, and Worthwhile." *Trends in Cognitive Sciences* 23 (10): 815–18. https://doi.org/10.1016/j.tics.2019.07.009.

Nosek, Brian A, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115 (11): 2600–2606. https://doi.org/10.1073/pnas.1708274114.

Nosek, Brian A, Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability." *Perspectives on Psychological Science* 7 (6): 615–31. https://doi.org/10.1177/1745691612459058.

O'Boyle, Ernest Hugh, George Christopher Banks, and Erik Gonzalez-Mulé. 2017. "The Chrysalis Effect: How Ugly Initial Results Metamorphosize into Beautiful Articles." *Journal of Management* 43 (2): 376–99. https://doi.org/10.1177/0149206314527133.

Oberauer, Klaus, and Stephan Lewandowsky. 2019. "Addressing the Theory Crisis in Psychology." *Psychonomic Bulletin & Review* 26 (5): 1596–1618.

Rubin, Mark. 2020. "Does Preregistration Improve the Credibility of Research Findings?" *The Quantitative Methods for Psychology* 16 (4): 15. https://doi.org/10.20982/tqmp.16.4.p376.

Simonsohn, Uri, Joseph P Simmons, and Leif D Nelson. 2020. "Specification Curve Analysis." *Nature Human Behaviour*, July, 1–7. https://doi.org/10.1038/s41562-020-0912-z.

Slovic, Paul, and Baruch Fischhoff. 1977. "On the Psychology of Experimental Surprises." *Journal of Experimental Psychology: Human Perception and Performance* 3 (4): 544–51. https://doi.org/10.1037/0096-1523.3.4.544.

Smaldino, Paul E, and Richard McElreath. 2016. "The Natural Selection of Bad Science." *Royal Society Open Science* 3 (9): 160384. https://doi.org/10.1098/rsos.160384.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11 (5): 702–12. https://doi.org/10.1177/1745691616658637.

Tukey, John W. 1980. "We Need Both Exploratory and Confirmatory." *The American Statistician* 34 (1): 23–25. https://doi.org/10.2307/2682991.

Veldkamp, Coosje L. S., Chris H. J. Hartgerink, Marcel A. L. M. van Assen, and Jelte M. Wicherts. 2017. "Who Believes in the Storybook Image of the Scientist?" *Accountability in Research* 24 (3): 127–51. https://doi.org/10.1080/08989621.2016.1268922.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit. 2012. "An Agenda for Purely Confirmatory Research." *Perspectives on Psychological Science* 7 (6): 632–38. https://doi.org/10.1177/1745691612463078.

12 Data collection

Learning goals

- Outline key features of informed consent and participant debriefing
- Identify additional protections necessary for working with vulnerable populations
- Review best practices for online and in-person data collection
- Implement data integrity checks, manipulation checks, and pilot testing

You have selected your measure and manipulation and planned your sample. Your preregistration is set. Now it's time to think about the nuts and bolts of collecting data. Though the details may vary between contexts, this chapter will describe some general best practices for data collection.[1] We organize our discussion of these practices around two perspectives: the participant and the researcher.

The first section takes the perspective of a participant. We begin by reviewing the importance of informed consent. A key principle of running experiments with human participants is that we respect their autonomy, which includes their right to understand the study and choose whether to take part. When we neglect the impact of our research on the people we study, we not only violate regulations governing research, we also create distrust that undermines the moral basis of scientific research.

In the second section, we begin to shift perspectives, discussing the choice of online vs. in-person data collection and some of the advantages of online data collection for TRANSPARENCY. We consider how to optimize the experimental experience for participants in both settings. We then end by taking the experimenter's perspective more fully, discussing how we can maximize data quality (contributing to MEASUREMENT PRECISION) using pilot testing, manipulation checks, and attention checks, while still being cognizant of both changes to the participant's experience and the integrity of statistical inferences (both contributing to BIAS REDUCTION).

[1] The metaphor of "collection" implies to some researchers that the data exist independent of the researcher's own perspective and actions, so they reject it in favor of the term "data generation." Unfortunately, this alternative label doesn't distinguish generating data via interactions with participants on the one hand and generating data from scratch via statistical simulations on the other. We worry that "data generation" sounds too much like the kinds of fraudulent data generation that we talked about in Chapter 4, so we have opted to keep the more conventional "data collection" label.

Case study

The rise of online data collection

Since the rise of experimental psychology laboratories in university settings during the period after World War 2 (Benjamin 2000), experiments have typically been conducted by recruiting participants from what has been referred to as the "subject pool." This term denotes a group of people who can be recruited for experiments, typically students from introductory psychology courses (Sieber and Saks 1989) who are required to complete a certain number of experiments as part of their course work. The ready availability of this convenient population inevitably led to the massive over-representation of undergraduates in published psychology research, undermining its generalizability (Sears 1986; Henrich, Heine, and Norenzayan 2010).

Yet over the last couple of decades, there has been a revolution in data collection. Instead of focusing on university undergraduates, increasingly, researchers recruit individuals from crowdsourcing websites like Amazon Mechanical Turk and Prolific Academic. Crowdsourcing services were originally designed to recruit and pay workers for ad-hoc business tasks like retyping receipts, but they have also become marketplaces to connect researchers with research participants who are willing to complete surveys and experimental tasks for small payments (Litman, Robinson, and Abberbock 2017). As of 2015, more than a third of studies in top social and personality psychology journals were conducted on crowdsourcing platforms (another third were still conducted with college undergraduates) and this proportion is likely continuing to grow (Anderson et al. 2019).

Initially, many researchers worried that crowdsourced data from online convenience samples would lead to a decrease in data quality. However, several studies suggest that data quality from online convenience samples is typically comparable to in-lab convenience samples (Mason and Suri 2012; Buhrmester, Kwang, and Gosling 2011). In one particularly compelling demonstration, a set of online experiments were used to replicate a group of classic phenomena in cognitive psychology, with clear successes on every experiment except those requiring sub-50 millisecond stimulus presentation (Crump, McDonnell, and Gureckis 2013). Further, as we discuss below, researchers have developed a suite of tools to ensure that online participants understand and comply with the instructions in complex experimental tasks.

Since these initial successes, however, attention has moved away from the validity of online experiments to the ethical challenges of engaging with crowdworkers. In 2020, nearly 130,000 people completed MTurk studies (Moss et al. 2020). Of those, an estimated 70% identified as White, 56% identified as women, and 48% had an annual household income below $50,000. A sampling of crowd work determined that the average wage earned was just $2.00 per hour, and less than 5% of workers were paid at least the federal minimum wage (Hara et al. 2018). Further, many experimenters routinely withheld payment from workers based on their performance in experiments. These practices clearly violate ethical guidelines for research with human participants, but are often overlooked by institutional review boards who may be unfamiliar with online recruitment platforms or consider that platforms are offering a "service" rather than simply being alternative routes for paying individuals.

With greater attention to the conditions of workers (e.g., Salehi et al. 2015), best practices for online research have progressed considerably. As we describe below, working with online populations requires attention to both standard ethical issues of consent and compensation, as well as new issues around the "user experience" of participating in research. The availability of online convenience samples can be transformative for the pace of research, for example by enabling large studies to be run in a single day rather than over many months. But online participants are vulnerable in different ways than university convenience samples, and we must take care to ensure that research online is conducted ethically.

12.1 Informed consent and debriefing

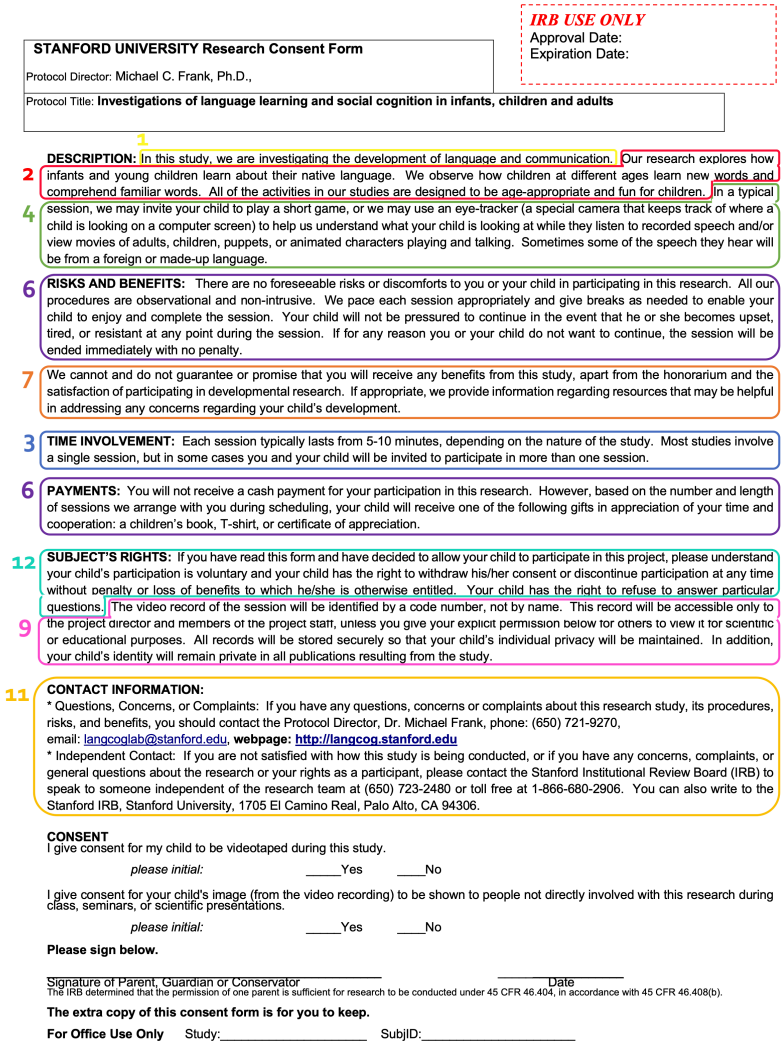
As we discussed in Chapter 4, experimenters must respect the autonomy of their participants: they must be informed about the risks and benefits of participation before they agree to participate. Researchers must also discuss and contextualize the research by debriefing participants after they have completed the study. Here we look at the nuts and bolts of each of these processes, ending with guidance on the special protections that are required to protect the autonomy of especially vulnerable populations.

12.1.1 Getting consent

Experimental participants must give consent. In most regulatory frameworks, there are clear guidelines about what the process of giving consent should look like. Typically participants are expected to read and sign a **consent form**: a document that explains the goals of the research and its procedures, describes potential risks and benefits, and asks for participants' explicit consent to participate voluntarily. Table 12.1 gives the full list of consent form requirements from the US Office for Human Research Protections, and Figure 12.1 shows how these individual requirements are reflected in a real consent form used in our research.

Table 12.1: US Office of Human Research Protections requirements for a consent form (edited for length).

	Requirement
1	A statement that the study involves research
2	An explanation of the purposes of the research
3	The expected duration of the subject's participation
4	A description of the procedures to be followed
5	Identification of any procedures which are experimental
6	A description of any reasonably foreseeable risks or discomforts to the subject
7	A description of any benefits to the subject or to others which may reasonably be expected from the research
8	A disclosure of appropriate alternative procedures or courses of treatment, if any, that might be advantageous to the subject
9	A statement describing the extent, if any, to which confidentiality of records identifying the subject will be maintained
10	For research involving more than minimal risk, an explanation as to whether any compensation or medical treatments are available if injury occurs
11	An explanation of whom to contact for answers to pertinent questions about the research and research subjects' rights
12	A statement that participation is voluntary, refusal to participate will involve no penalty, and that subject may discontinue participation at any time without penalty



STANFORD UNIVERSITY Research Consent Form

Protocol Director: Michael C. Frank, Ph.D.,

Protocol Title: **Investigations of language learning and social cognition in infants, children and adults**

IRB USE ONLY
Approval Date:
Expiration Date:

DESCRIPTION: In this study, we are investigating the development of language and communication. Our research explores how infants and young children learn about their native language. We observe how children at different ages learn new words and comprehend familiar words. All of the activities in our studies are designed to be age-appropriate and fun for children. In a typical session, we may invite your child to play a short game, or we may use an eye-tracker (a special camera that keeps track of where a child is looking on a computer screen) to help us understand what your child is looking at while they listen to recorded speech and/or view movies of adults, children, puppets, or animated characters playing and talking. Sometimes some of the speech they hear will be from a foreign or made-up language.

RISKS AND BENEFITS: There are no foreseeable risks or discomforts to you or your child in participating in this research. All our procedures are observational and non-intrusive. We pace each session appropriately and give breaks as needed to enable your child to enjoy and complete the session. Your child will not be pressured to continue in the event that he or she becomes upset, tired, or resistant at any point during the session. If for any reason you or your child do not want to continue, the session will be ended immediately with no penalty.

We cannot and do not guarantee or promise that you will receive any benefits from this study, apart from the honorarium and the satisfaction of participating in developmental research. If appropriate, we provide information regarding resources that may be helpful in addressing any concerns regarding your child's development.

TIME INVOLVEMENT: Each session typically lasts from 5-10 minutes, depending on the nature of the study. Most studies involve a single session, but in some cases you and your child will be invited to participate in more than one session.

PAYMENTS: You will not receive a cash payment for your participation in this research. However, based on the number and length of sessions we arrange with you during scheduling, your child will receive one of the following gifts in appreciation of your time and cooperation: a children's book, T-shirt, or certificate of appreciation.

SUBJECT'S RIGHTS: If you have read this form and have decided to allow your child to participate in this project, please understand your child's participation is voluntary and your child has the right to withdraw his/her consent or discontinue participation at any time without penalty or loss of benefits to which he/she is otherwise entitled. Your child has the right to refuse to answer particular questions. The video record of the session will be identified by a code number, not by name. This record will be accessible only to the project director and members of the project staff, unless you give your explicit permission below for others to view it for scientific or educational purposes. All records will be stored securely so that your child's individual privacy will be maintained. In addition, your child's identity will remain private in all publications resulting from the study.

CONTACT INFORMATION:
* Questions, Concerns, or Complaints: If you have any questions, concerns or complaints about this research study, its procedures, risks, and benefits, you should contact the Protocol Director, Dr. Michael Frank, phone: (650) 721-9270, email: langcoglab@stanford.edu, **webpage: http://langcog.stanford.edu**
* Independent Contact: If you are not satisfied with how this study is being conducted, or if you have any concerns, complaints, or general questions about the research or your rights as a participant, please contact the Stanford Institutional Review Board (IRB) to speak to someone independent of the research team at (650) 723-2480 or toll free at 1-866-680-2906. You can also write to the Stanford IRB, Stanford University, 1705 El Camino Real, Palo Alto, CA 94306.

CONSENT
I give consent for my child to be videotaped during this study.
please initial: _____Yes ____No

I give consent for your child's image (from the video recording) to be shown to people not directly involved with this research during class, seminars, or scientific presentations.
please initial: _____Yes ____No

Please sign below.

_____ _____
Signature of Parent, Guardian or Conservator Date

The IRB determined that the permission of one parent is sufficient for research to be conducted under 45 CFR 46.404, in accordance with 45 CFR 46.408(b).

The extra copy of this consent form is for you to keep.

For Office Use Only Study:_____ SubjID:_____

Figure 12.1: Consent form annotated to show how specific text fulfills the requirements in Table 12.1. Categories 5, 8, and 10 were not required for this minimal risk psychology experiment.

These are just samples. Since ethics regulation is almost always managed at the institutional level, your local ethics board will often provide guidance on the specific information you should include in the consent form and they will almost always need to approve the form before you are allowed to begin recruiting participants.

When providing consent information, researchers should focus on what someone might think or feel as a result of participating in the study. Are there any physical or emotional risks associated? What should someone know about the study that may give them pause about agreeing to participate in the first place? Our advice is to center the *participant* in the consent process rather than the research question. Information about specific research goals can typically be provided during debriefing.[2]

[2] Some experimenters worry that informing participants about the study that they are about to participate in may influence their behavior in the study via so-called "demand characteristics", discussed in Chapter 9. But the goal of a consent form is not to explain the specific psychological construct being manipulated. Instead, a consent form typically focuses on the experience of being in the study (for example, that a participant would be asked to provide quick verbal responses to pictures). This sort of general explanation should not create demand characteristics.

If there are specific pieces of information that about study goals or procedures that *must* be withheld from participants during consent, **deception** of participants may be warranted. Deception can be approved by ethics boards as long as it poses little risk and is effectively addressed via more extensive debriefing. But an experimental protocol that includes deception will likely undergo greater scrutiny during ethics review, as it must be justified by a specific experimental need.

During the consent process, researchers should explain to participants what will be done with their data. Requirement 9 in Table 12.1 asks for a statement about data confidentiality, but such a statement is a mere minimum. Some modern consent forms explicitly describe different uses of the data and ask for consent for each. For example, the form in Figure 12.1 asks permission for showing recordings as part of presentations.[3]

[3] Some ethics boards will ask for consent for sharing even anonymized data files. As we discuss in Chapter 13, fully anonymized data can often be shared without explicit consent. You may still choose to ask participants' permission, but this practice may lead to an awkward situation, for example, a dataset with heterogeneous sharing permissions such that most but not all data can be shared publicly. Norms around anonymized data sharing are shifting, so it's worth having a conversation with your ethics board about how they interpret your particular regulatory obligations.

12.1.2 Prerequisites of consent

In order to give consent, participants must have the cognitive capacity to make decisions (competence), understand what they are being asked to do (comprehension), and know that they have the right to withdraw consent at any time (voluntariness) (Kadam 2017).

Typically, we assume competence for adult volunteers in our experiments, but if we are working with children or other vulnerable populations (see below), we may need to consider whether they are legally competent to provide consent. Participants who cannot consent on their own should still be informed about participation in an experiment and, if possible, you should still obtain their **assent** (informal agreement) to participate. When a person has no legal ability to consent, you must obtain consent from their legal guardian. But if they do not assent, you should also respect their decision not to participate – even if you previously obtained consent from their guardian.

The second prerequisite is comprehension. It is good practice to discuss consent forms verbally with participants, especially if the study is involved and takes place in person. If the study is online, ensure that participants know how to contact you if they have questions about the study. The consent form itself must be readable for a broad audience, meaning care should be taken to use accessible language and clear formatting. Consider giving participants a copy of the consent form in advance so they can read at their own pace, think of any questions they might have, and decide how to proceed without any chance of feeling coerced (Young, Hooker, and Freeberg 1990).

Finally, participants must understand that their involvement is voluntary, meaning that they are under no obligation to be involved in a study and always have the right to withdraw at any time. Experimenters should not only state that participation is voluntary, they should also pay attention to other features of the study environment that might lead to **structural coercion** (Fisher 2013). For example, high levels of compensation can make it difficult for lower-income participants to withdraw from research. Similarly, factors like race, gender, and social class can lead participants to feel discomfort around discontinuing a study. It is incumbent on experimenters to provide a comfortable study environment and to avoid such coercive factors wherever possible.

12.1.3 Debriefing participants

Once a study is completed, researchers should always debrief participants. A debriefing is composed of sevearl parts: (1) gratitude, (2) discussion of goals, (3) explanation of deception (if relevant), and (4) questions and clarification (Allen 2017). Together these serve to contextualize the experience for the participant and to mitigate any potential harms from the study.

1. **Gratitude**. Thank participants for their contribution! Sometimes thanks is enough (for a short experiment), but many studies also include monetary compensation or course credit. Compensation should be commensurate with the amount of time and effort required for participation. Compensation structures vary widely from place to place; typically local ethics boards will have specific guidelines.

2. **Discussion of goals**. Researchers should share the purpose of the research with participants in, aiming for a short and accessible statement that avoids technical jargon. Sharing goals is especially important when some aspect of the study appears evaluative – participants will often be interested in knowing how well they performed against their peers. For example, a parent whose child completed a word-recognition task may request information about their child's performance. It can assuage parents' worries to highlight that the goals of the study are about measuring a particular experimental effect, not about individual evaluation and ranking.[4]

3. **Explanation of deception**. Researchers must reveal any deception during debriefing, regardless of how minor the deception seems to the researcher. This component of the debriefing process can be thought of as "dehoaxing" because it is meant to illuminate

[4] At the study's conclusion, you might also consider sharing any findings with participants – many participants appreciate learning about research findings that they contributed to, even months or years after participation.

any aspects of the study that were previously misleading or inaccurate (Holmes 1976). The goal is both to reveal the true intent of the study and to alleviate any potential anxiety associated with the deception. Experimenters should make clear both where in the study the deception occurred and why the deception was necessary for the study's success.

4. **Questions and clarification**. Finally, researchers should answer any questions or address any concerns raised by participants. Many researchers use this opportunity to ask participants about their own ideas about the study goals. This practice not only illuminates aspects of the study design that may have been unclear to or hidden from participants, it also begins a discussion where both researchers and participants can communicate about this joint experience. This step is also helpful in identifying negative emotions or feelings resulting from the study (Allen 2017). When participants do express negative emotions, researchers are responsible for sharing resources participants can use to help them.[5]

[5] In the case that participants report substantial concerns or negative reactions to an experiment – during debriefing or otherwise – researchers will typically have an obligation to report these to their ethics board.

12.1.4 Special considerations for vulnerable populations

Regardless of who is participating in research, investigators have an obligation to protect the rights and well-being of all participants. Some populations are considered especially **vulnerable** because of their decreased agency – either in general or in the face of potentially coercive situations. Research with these populations receives additional regulatory oversight. In this section, we will consider several vulnerable populations.

Children. Children are some of the most commonly used vulnerable populations in research because the study of development can contribute both to children's welfare and to our understanding of the human mind. In the US, children under the age of 18 may only participate in research with written consent from a parent or guardian. Unless they are pre-verbal, children should additionally be asked for their assent. The risks associated with a research study focusing on children also must be no greater than minimal unless participants may receive some direct benefit from participating or participating in the study may improve a disorder or condition the participant was formally diagnosed with.

People with disabilities. There are thousands of disabilities that affect cognition, development, motor ability, communication, and decision-making with varying degrees of interference, so it is first important to

remember that considerations for this population will be just as diverse as its members. No laws preclude people with disabilities from participating in research. However, those with cognitive disabilities who are unable to make their own decisions may only participant with written consent from a legal guardian and with their individual assent (if applicable). Those retaining full cognitive capacity but who have other disabilities that make it challenging to participate normally in the study should receive appropriate assistance to access information about the study, including the risks and benefits of participation.

Incarcerated populations. Nearly 2.1 million people are incarcerated in the United States alone (Gramlich 2021). Due to early (and repugnant) use of prisoners as a convenience population that could not provide consent, the use of prisoners in research has been a key focus of protective efforts. The US Office for Human Research Protections (OHRP) supports their involvement in research under very limited circumstances – typically when the research specifically focuses on issues relevant to incarcerated populations (Office for Human Research Protections 2003). When researchers propose to study incarcerated individuals, the local ethics board must reconfigure to include at least one active prisoner (or someone who can speak from a prisoner's perspective) and ensure that less than half of the board has any affiliation to the prison system, public or private. Importantly, researchers must not suggest or promise that participation will have any bearing on an individual's prison sentence or parole eligibility, and compensation must be otherwise commensurate with their contribution.

Low-income populations. Participants with fewer resources may be more persuaded to participate by monetary incentives, creating a potentially coercive situation. Researchers should consult with their local ethics board to conform to local standards for non-coercive payment.

Indigenous populations. There is a long and negative history of the involvement of indigenous populations in research without their consent. In the case that research requires the participation of indigenous individuals – because of potential benefits to their communities, rather than due to convenience – then community leadership must be involved to discuss the appropriateness of the research as well as how the consent process should be structured (Fitzpatrick et al. 2016).

Crowdworkers. Ethics boards do not usually consider crowdworkers on platforms like Amazon Mechanical Turk to be a specific vulnerable population, but many of the same concerns about diminished autonomy and greater need for protection still arise (see Depth Box below). Without platform or ethics board standards, it is up to individual experimenters to commit to fair pay, which should ideally match or exceed

the applicable minimum wage (e.g., the US federal minimum wage). Further, in the context of reputation management systems like those of Amazon Mechanical Turk, participants can be penalized for withdrawing from an experiment – once they have their work "rejected" by an experimenter, it can be harder for them to find new jobs, causing serious long-term harm to their ability to earn on the platform.

12.2 Designing the "research experience"

For the majority of psychology experiments, the biggest factor that governs whether a participant has a positive or negative experience of an experiment is not its risk profile, since for many psychology experiments the quantifiable risk to participants is minimal.[6] Instead, it is the participants' experience. Did they feel welcome? Did they understand the instructions? Did the software work as designed? Was their compensation clearly described and promptly delivered? These aspects of "user experience" are critical both for ensuring that participants have a good experience in the study (an ethical imperative) and for gathering good data. An experiment that leaves participants unhappy typically doesn't satisfy either the ethical or the scientific goals of research. In this section, we'll discuss how to optimize the research experience for both in-person and online experiments, as well as providing some guidance on how to decide between these two administration contexts.

[6] There are of course exceptions, including research with more sensitive content. Even in these cases, however, attention to the participant's experience can be important for ensuring good scientific outcomes.

12.2.1 Ensuring good experiences for in-lab participants

A participant's experience begins even before they arrive at the lab. Negative experiences with the recruitment process (e.g., unclear consent forms, poor communication, complicated scheduling) or transit to the lab (e.g., difficulty navigating or finding parking) can lead to frustrated participants with a negative view of your research. Anything you can do to make these experiences smoother and more predicable – prompt communication, well-tested directions, reserved parking slots, etc. – will make your participants happier and increase the quality of your data.[7]

Once a participant enters the lab, every aspect of the interaction with the experimenter can have an effect on their measured behavior (Gass and Seiter 2018)! For example, a likable and authoritative experimenter who clearly describes the benefits of participation is following general principles for persuasion (Cialdini and Goldstein 2004). This interaction should lead to better compliance with experimental instructions, and hence better data, than an interaction with an unclear or indifferent experimenter.

[7] For some reason, the Stanford Psychology Department building is notoriously difficult to navigate. This seemingly minor issue has resulted in a substantial number of late, frustrated, and flustered participants over the years.

Any interaction with participants must be scripted and standardized so that all participants have as similar an experience as possible. A lack of standardization can result in differential treatment for participants with different characteristics, which could result in data with greater variability or even specific sociodemographic biases. An experimenter that was kinder and more welcoming to one demographic group would be acting unethically, and they also might find a very different result than they intended.

Even more importantly, experimenters who interact with participants should ideally be unaware of the experimental condition each participant is assigned to. This practice is often called "blinding" or "masking". Otherwise it is easy for experimenter knowledge to result in small differences in interaction across conditions, which in turn can influence participants' behavior, resulting in experimenter expectancy effects (see Chapter 9)! Even if the experimenter must know a participant's condition assignment – as is sometimes the case – this information should be revealed at the last possible moment to avoid contamination of other aspects of the experimental session.[8]

[8] In some experiments, an experimenter delivers a manipulation and hence it cannot be masked from them. In such cases, it's common to have two experimenters such that one delivers the manipulation and another (masked to condition) collects the measurements. This situation often comes up with studies of infancy, since stimuli are often delivered via an in-person puppet show; at a minimum, behavior should be coded by someone other than the puppeteer.

12.2.2 Ensuring good experiences for online participants

The design challenges for online experiments are very different than for in-lab experiments. As the experimental procedure is delivered through a web browser, experimenter variability and potential expectancy effects are almost completely eliminated. On the other hand, some online participants do many hours of online tasks a day and many are multitasking in other windows or on other devices. It can be much harder to induce interest and engagement in your research when your manipulation is one of dozens the participant has experienced that day and when your interactions are mediated by a small window on a computer screen.

When creating an online experimental experience, we consider four issues: (1) design, (2) communication, (3) payment policies, and (4) effective consent and debriefing.[9]

[9] For extensive further guidance on this topic, see Litman and Robinson (2020).

Basic UX design. Good experiment design online is a subset of good web **user experience** (UX) design more generally. If your web experiment is unpleasant to interact with, participants will likely become confused and frustrated. They will either drop out or provide data that are lower quality. A good interface should be clean and well-tested and should offer clear places where the participant must type or click to interact. If a participant presses a key at an appropriate time, the experiment should offer a response – otherwise the participant will likely press

it again. If the participant is uncertain how many trials are left, they may be more likely to drop out of the experiment so it is also helpful to provide an indication of their progress. And if they are performing a speeded paradigm, they should receive practice trials to ensure that they understand the experiment prior to beginning the critical blocks of trials.

Communication. Many online studies involve almost no direct contact with participants. When participants do communicate with you it is very important to be responsive and polite (as it is with in-lab participants, of course). Unlike the typical undergraduate participant, the work that a crowdworker is doing for your study may be part of how they earn their livelihood, and a small issue in the study for you may feel very important for them. For that reason, rapid resolution of issues with studies – typically through appropriate compensation – is very important. Crowdworkers often track the reputation of specific labs and experimenters [sometimes through forums or specialized software; Irani and Silberman (2013)]. A quick and generous response to an issue will ensure that future crowdworkers do not avoid your studies.

Payment policies. Unclear or punitive payment policies can have a major impact on crowdworkers. We strongly recommend *always* paying workers if they complete your experiment, regardless of result. This policy is comparable to standard payment policies for in-lab work. We assume good faith in our participants: if someone comes to the lab, they are paid for the experiment, even if it turns out that they did not perform correctly. The major counterargument to this policy is that some online marketplaces have a population of workers who are looking to cheat by being non-compliant with the experiment (e.g., entering gibberish or even using scripts or artificial intelligence tools to progress quickly through studies). Our recommendation is to address this issue through the thoughtful use of "check" trials (see below) – not through punitive non-payment. The easiest way for a participant to complete your experiment should be by complying with your instructions.

Table 12.2: Sample online consent statement from our course.

By answering the following questions, you are participating in a study being performed by cognitive scientists in the Stanford Department of Psychology. If you have questions about this research, please contact us at stanfordpsych251@gmail.com. You must be at least 18 years old to participate. Your participation in this research is voluntary. You may decline to answer any or all of the following questions. You may decline further participation, at any time, without adverse consequences. Your anonymity is assured; the researchers who have requested your participation will not receive any personal information about you.

Consent and debriefing. Because online studies are typically fully automated, participants do not have a chance to interact with researchers around consent and debriefing. Further, engagement with long consent forms may be minimal. In our work we have typically relied on short consent statements such as the one from our class that is shown in Table 12.2. Similarly, debriefing often occurs through a set of pages that summarize all components of the debriefing process (participation gratitude, discussion of goals, explanation of deception if relevant, and questions and clarification). Because these interactions are so short, it is especially important to include contact information prominently so that participants can follow up.

12.2.3 When to collect data online?

Online data collection is increasingly ubiquitous in the behavioral sciences. Further, the web browser – alongside survey software like Qualtrics or packages like jsPsych (De Leeuw 2015) – can be a major aid to transparency in sharing experimental materials. Replication and reuse of experimental materials is vastly simpler if readers and reviewers can click a link and share the same experience as a participant in your experiment. By and large, well-designed studies yield data that are as reliable as in-lab data (Buhrmester, Kwang, and Gosling 2011; Mason and Suri 2012; Crump, McDonnell, and Gureckis 2013).

Still, online data collection is not right for every experiment. Studies that have substantial deception or induce negative emotions may require an experimenter present to alleviate ethical concerns or provide debriefing. Beyond ethical issues, we discuss four broader concerns to consider when deciding whether to conduct data collection online: (1) population availability, (2) the availability of particular measures, (3) the feasibility of particular manipulations, and (4) the length of experiments.

Population. Not every target population can be tested online. Indeed, initially, convenience samples from Amazon Mechanical Turk were the only group easily available for online studies. More recently, new tools have emerged to allow pre-screening of crowd participants, including sites like Cloud Research and Prolific (Eyal et al. 2021; Peer et al. 2021).[10] And it may initially have seemed implausible that children could be recruited online, but during the COVID-19 pandemic a substantial amount of developmental data collection moved online, with many studies yielding comparable results to in-lab studies (e.g., Chuey et al. 2021).[11] Finally, new, non-US crowdsourcing platforms continue to grow in popularity, leading to greater global diversity in the available online populations.

[10] These tools still have significant weaknesses for accessing socio-demographically diverse populations within and outside the US, however – screening tools can remove participants, but if the underlying population does not contain many participants from a particular demographic, it can be hard to gather large enough samples. For an example of using crowdsourcing and social media sites to gather diverse participants, see DeMayo et al. (2021).

[11] Sites like LookIt (https://lookit.mit.edu) now offer sophisticated platforms for hosting studies for children and families (Scott and Schulz 2017).

Online measures. Not all measures are available online, though more and more are. Although online data collection was initially restricted to the use of survey measures – including ratings and text responses – measurement options have rapidly expanded. The widespread use of libraries like jsPsych (De Leeuw 2015) has meant that millisecond accuracy in capturing response times is now possible within web-browsers; thus, most reaction time tasks are quite feasible (Crump, McDonnell, and Gureckis 2013). The capture of sound and video is possible with modern browser frameworks (Scott and Schulz 2017). Further, even measures like mouse- and eye-tracking are beginning to become available (Maldonado, Dunbar, and Chemla 2019; Slim and Hartsuiker 2023). In general, almost any variable that can be measured in the lab without specialized apparatus can also be collected online. On the other hand, studies that measure a broader range of physiological variables (e.g., heart rate or skin conductance) or a larger range of physical behaviors (e.g., walking speed or pose) are still likely difficult to implement online.

Online manipulations. Online experiments are limited to the set of manipulations that can be created within a browser window – but this restriction excludes many different manipulations that involve real-time social interactions with a human being.[12] Synchronous chat sessions can be a useful substitute (Hawkins, Frank, and Goodman 2020), but these focus the experiment on the content of what is said and exclude the broader set of non-verbal cues available to participants in a live interaction (e.g., gaze, race, appearance, accent, etc.). Creative experimenters can circumvent these limitations by using pictures, videos, and other methods. But more broadly, an experimenter interested in implementing a particular manipulation online should ask how compelling the online implementation is compared with an in-lab implementation. If the intention is to induce some psychological state – say stress, fear, or disgust – experimenters must trade off the greater ease of recruitment and larger scale of online studies with the more compelling experience they may be able to offer in a controlled lab context.

The length of online studies. One last concern is about attention and focus in online studies. Early guidance around online studies tended to focus on making studies short and easy, with the rationale that crowdsourcing workers were used to short jobs. Our sense is that this guidance no longer holds. Increasingly, researchers are deploying long and complex batteries of tasks to relatively good effect (e.g., Enkavi et al. 2019) and conducting repeated longitudinal sampling protocols (discussed in depth in Litman and Robinson 2020). Rather than relying on hard and fast rules about study length, a better approach for online testing is to ensure that participants' experience is as smooth and compelling as pos-

[12] So-called "moderated" experiments – in which the experimental session is administered through a synchronous video chat have been used widely in online experiments for children but these designs are less common in experiments with adults because they are expensive and time-consuming to administer (Chuey et al. 2021).

sible. Under these conditions, if an experiment is viable in the lab, it is likely viable online.

Online testing tools continue to grow and change but they are already mature enough that using them should be part of most behavioral researchers' basic toolkit.[13]

[13] It is of course import to keep in mind that if a person works part- or full-time on a crowdsourcing platform, they are not a representative sample of the broader national population. Unfortunately, similar caveats hold true for in-person convenience samples (see Chapter 10). Ultimately, researchers must reason about what their generalization goal is and whether that goal is consistent with the samples they can access (online or otherwise).

12.3 Ensuring high quality data

In the final section of this chapter, we review some key data collection practices that can help researchers collect high quality data while respecting our ethical obligations to participants. By "high quality," here we especially mean datasets that are uncontaminated by responses generated by misunderstanding of instructions, fatigue, incomprehension, or intentional neglect of the experimental task.

We'll begin by discussing the issue of pilot testing; we recommend a systematic procedure for piloting that can maximize the chance of collecting high quality data. Next, we'll discuss the practice of checking participants' comprehension and attention and what such checks should and shouldn't be used for. Finally, we'll discuss the importance of maintaining consistent data collection records.

12.3.1 Conduct effective pilot studies

A **pilot study** is a small study conducted before you collect your main sample. The goal is to ensure smooth and successful data collection by first checking if your experimental procedures and data collection workflow are working correctly. Pilot studies are also an opportunity to get feedback from participants about their experience of the experimental task, for example, is it too easy, too difficult, or too boring.

Because pilot studies usually involve a small number of participants, they are not a reliable indicator of the study results, such as the expected effect size or statistical significance (as we discussed in Chapter 10). *Don't* use pilots to check if your effect is present or to estimate an effect size for power analysis. What pilots *can* do is tell you about whether your experimental procedure is viable. For example, pilots studies can reveal:

- if your code crashes under certain circumstances
- if your instructions confuse a substantial portion of participants
- if you have a very high dropout rate
- if your data collection procedure fails to log variables of interest
- if participants are disgruntled by the end of the experiment

We recommend that all experimenters perform – at the very minimum – two pilot studies before they launch a new experiment.[14]

The first pilot, which we call your **non-naïve participant pilot**, can make use of participants who know the goals of the experiment and understand the experimental manipulation – this could be a friend, collaborator, colleague, or family member.[15] The goal of this pilot study is to ensure that your experiment is comprehensible, that participants can complete it, and that the data are logged appropriately. You must *analyze* the data from the non-naive pilot, at least to the point of checking that the relevant data about each trial is logged.

The second pilot, your **naïve participant pilot**, should consist of a test of a small set of participants recruited via the channel you plan to use for your main study. The number of participants you should pilot depends on the cost of the experiment in time, money, and opportunity as well as its novelty. A brand new paradigm is likely more prone to error than a tried and tested paradigm. For a short online survey-style experiment, a pilot of 10–20 people is reasonable. A more time-consuming laboratory study might require piloting just two or three people.[16]

The goal of the naïve pilot study is to understand properties of the participant experience. Were participants confused? Did they withdraw before the study finished? Even a small number of pilots can tell you that your dropout rate is likely too high: for example, if 5 of 10 pilot participants withdraw you likely need to reconsider aspects of your design. It's critical for your naïve participant pilot that you debrief more extensively with your participants. This debriefing often takes the form of an interview questionnaire after the study is over. "What did you think the study was about?" and "is there any way we could improve the experience of being in the study?" can be helpful questions. Often this debriefing is more effective if it is interactive, so even if you are running an online study you may want to find some way to chat with your participants.

Piloting – especially piloting with naïve participants to optimize the participant experience – is typically an iterative process. We frequently launch an experiment for a naive pilot, then recognize from the data or from participant feedback that the experience can be improved. We make tweaks and pilot again. Be careful not to over-fit to small differences in pilot data, however. Piloting should be more like workshopping a manuscript to remove typos than doing statistical analysis. If someone has trouble understanding a particular sentence – whether in your manuscript or in your experiment instructions – you should edit to make it clearer!

[14] We mean especially when deploying a new experimental paradigm or when collecting data from a new population. Once you have run many studies with a similar procedure and similar sample, extensive piloting is less important. Any time you change something, it's always good to run one or two pilots, though, just to check that you didn't inadvertently mess up your experiment.

[15] In a pinch you can even run yourself through the experiment a bunch of times (though this isn't preferable because you're likely to miss a lot of aspects of the experience that you are habituated to, especially if you've been debugging the experiment already).

[16] In the case of especially expensive experiments, it can be a dilemma whether to run a larger pilot to identify difficulties since such a pilot will be costly. In these cases, one possibility is to plan to include the pilot participants in the main dataset if no major procedural changes are required. In this case, it is helpful to preregister a contingent testing strategy to avoid introducing data-dependent bias (see Chapter 11). For example, in a planned sample of 100 participants, you could preregister running 20 as a pilot sample with the stipulation that you will look only at their dropout rate – and not at any condition differences. Then the preregistration can state that, if the dropout rate is lower than 25%, you will collect the next 80 participants and analyze the whole dataset, including the initial pilot, but if dropout rate is higher than 25%, you will discard the pilot sample and make changes. This kind of strategy can help you split the difference between cautious piloting and conservation of rare or costly data.

Accident report

Data logging much?

When Mike was in graduate school, his lab got a contract to test a very large group of participants in a battery of experiments, bringing them into the lab over the course of a series of intense bursts of participant testing. He got the opportunity to add an experiment to the battery, allowing him to test a much larger sample than resources would otherwise allow. He quickly coded up a new experiment as part of a series of ongoing studies and began deploying it, coming to the lab every weekend for several months to help move participants through the testing protocol. Eagerly opening up the data file to reap the reward of this hard work, he found that the condition variable was missing from the data files. Although the experimental manipulation had been deployed properly, there was no record of which condition each participant had been run in, and so the data were essentially worthless. Had he run a quick pilot (even with non-naive participants) and attempted to analyze the data, this error would have been detected, and many hours of participant and experimenter effort would not have been lost.

12.3.1 Measure participant compliance

You've constructed your experiment and piloted it. You are almost ready to go – but there is one more family of tricks for helping to achieve high quality data: integrating measures of participant compliance into your paradigm. Collecting data on compliance (whether participants followed the experimental procedures as expected) can help you quantify whether participants understood your task, engaged with your manipulation, and paid attention to the full experimental experience. These measures in turn can be used both to modify your experimental paradigm and to exclude specific participants that were especially non-compliant (Hauser, Ellsworth, and Gonzalez 2018; Ejelöv and Luke 2020).

Below we discuss four types of compliance checks: (1) passive measures, (2) comprehension checks, (3) manipulation checks, and (4) attention checks. Passive measures and comprehension checks are very helpful for enhancing data quality. Manipulation checks also often have a role to play. In contrast, we typically caution in the use of attention checks.

1. **Passive measures of compliance.** Even if you do not ask participants anything extra in an experiment, it is often possible to tell if they have engaged with the experimental procedure simply by how long it takes them to complete the experiment. If you see participants with completion times substantially above or below the median, there is a good chance that they are either multi-tasking or rushing through the experiment without engaging.[17] Passive measures cost little to implement and should be inserted whenever possible in experiments.[18]

[17] Measurements of per-page or per-element completion times can be even more specific since they can, for example, identify participants that simply did not read an assigned passage.

[18] One variation that we endorse in certain cases is to force participants to engage with particular pages for a certain amount of time through the use of timers. Though, beware, this kind of feature can lead to an adversarial relationship with participants – in the face of this kind of coercion, many will opt to pull out their phone and multi-task until the timer runs down.

2. **Comprehension checks.** For tasks with complex instructions or experimental materials (say a passage that must be understood for a judgment to be made about it), it can be very helpful to get a signal that participants have understood what they have read or viewed. Comprehension checks, which ask about the content of the experimental instructions or materials, are often included for this purpose. For the comprehension of instructions, the best kinds of questions simply query the knowledge necessary to succeed in the experiment, for example, "what are you supposed to do when you see a red circle flash on the screen?" In many platforms, it is possible to make participants reread the instructions again until they can answer these correctly. This kind of repetition is nice because it corrects participants' misconceptions rather than allowing them to continue in the experiment when they do not understand.[19]

3. **Manipulation checks.** If your experiment involves more than a very transient manipulation – for example, if you plan to induce some state in participants or have them learn some content – then you can include a measure in your experiment that confirms that your manipulation succeeded (Ejelöv and Luke 2020). This measure is known as a manipulation check because it measures some prerequisite difference between conditions that is not the key causal effect of interest but is causally prerequisite to this effect. For example, if you want to see if anger affects moral judgment, then it makes sense to measure whether participants in your anger induction condition rate themselves as angrier than participants in your control condition. Manipulation checks are useful in the interpretation of experimental findings because they can decouple the failure of a manipulation from the failure of a manipulation to affect your specific measure of interest.[20]

4. **Attention checks.** A final type of compliance check is a check that participants are paying attention to the experiment at all. One simple technique is to add questions that have a known and fairly obvious right answer (e.g., "what's the capital of the United States."). These trials can catch participants that are simply ignoring all text and "mashing buttons", but they will not find participants who are mildly inattentive. Sometimes experimenters also use trickier compliance checks, such as putting an instruction for participants to click a particular answer deep within a question text that otherwise would have a different answer (Oppenheimer, Meyvis, and Davidenko 2009) (Figure 12.2). Such compliance checks decrease so-called "satisficing" behavior, in which participants read as quickly as

[19] If you are querying comprehension of experimental materials rather than instructions, you may not want to re-expose participants to the same passage again in order to avoid confounding a participants' initial comprehension and the amount of exposure that they receive.

[20] Hauser, Ellsworth, and Gonzalez (2018) worry that manipulation checks can themselves change the effect of a manipulation – this worry strikes us as sensible, especially for some types of manipulations like emotion inductions. Their recommendation is to test the efficacy of the manipulation in a separate study, rather than trying to nest the manipulation check within the main study.

they can get away with (doing only the minimum. On the other hand, participants may see such trials as indications that the experimenter is trying to trick them, and adopt a more adversarial stance towards the experiment, which may result in less compliance with other aspects of the design (unless they are at the end of the experiment, Hauser, Ellsworth, and Gonzalez 2018). If you choose to include attention checks like these, be aware that you are likely reducing variability in your sample – trading off representativeness for compliance.

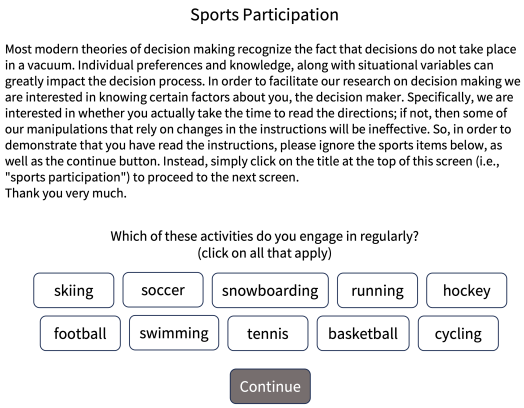


Figure 12.2: An attention check trial based on Oppenheimer, Meyvis, and Davidenko (2009). These trials can decrease variability in participant attention, but at the cost of selecting a subsample of participants, so they should be used cautiously.

Data from all of these types of checks are used in many different – often inconsistent – ways in the literature. We recommend that you:

1. Use passive measures and comprehension checks as pre-registered exclusion criteria to eliminate a (hopefully small) group of participants who might be non-compliant with your experiment.
2. Check that exclusions are low and that they are uniform across conditions. If exclusion rates are high, your design may have deeper issues. If exclusions are asymmetric across conditions, you may be compromising your randomization by creating a situation in which (on average) different kinds of participants are included in one condition compared with the other. Both of these situations substantially compromise any estimate of the causal effect of interest.

Accident report

Does data quality vary throughout the semester?

Every lab that collects empirical data repeatedly using the same population builds up lore about how that population varies in different contexts. Many researchers who conducted experiments with college undergraduates were taught never to run their studies at the end of the semester. Exhausted and stressed students would likely yield low-quality data, or so the argument went. Until the rise of multi-lab collaborative projects like ManyLabs (see Chapter 3), such beliefs were almost impossible to test.

ManyLabs 3 aimed specifically to evaluate data quality variation across the academic calendar (Ebersole et al. 2016). With 2,696 participants at 20 sites, the study conducted replications of 13 previously published findings. Although only six of these findings showed strong evidence of replicating across sites, none of the six effects was substantially moderated by being collected later in the semester. The biggest effect they observed was a change in the Stroop effect from $d = .89$ during the beginning and middle of the semester to $d = .92$ at the end. There was some evidence that participants *reported* being less attentive at the end of the semester, but this trend wasn't accompanied by a moderation of experimental effects.

Researchers are subject to the same cognitive illusions and biases as any human. One of these biases is the search to find meaning in the random fluctuations they sometimes observe in their experiments. The intuitions formed through this process can be helpful prompts for generating hypotheses – but beware of adopting them into your "standard operating procedures" without further examination. Labs that avoided data collection during the end of the semester might have sacrificed 10–20% of their data collection capacity for no reason!

3. Deploy manipulation checks if you are concerned about whether your manipulation effectively induces a difference between groups. Analyze the manipulation check separately from the dependent variable to test whether the manipulation was causally effective (Ejelöv and Luke 2020).
4. Make sure that your attention checks are not confounded in any way with condition – remember our cautionary tale from Chapter 9, in which an attention check that was different across conditions actually created an experimental effect.
5. *Do not* include any of these checks in your analytic models as a covariate, as including this information in your analysis compromises the causal inference from randomization and introduces bias in your analysis (Montgomery, Nyhan, and Torres 2018).[21]

Used appropriately, compliance checks can provide both a useful set of exclusion criteria and a powerful tool for diagnosing potential issues with your experiment during data analysis and correcting them down the road.

[21] Including this information means you are "conditioning on a post-treatment variable," as we described in Chapter 7. In medicine, analysts distinguish "intent-to-treat" analysis, where you analyze data from everyone you gave a drug, and "as treated" analysis, where you analyze data depending on how much of the drug people actually took. In general, intent-to-treat gives you the generalizable causal estimate. In our current situation, if you include compliance as a covariate, you are essential doing an "as treated" analysis and your estimate can be biased as a result. Although there is occasional need for such analyses, in general you probably want to avoid them.

12.3.1 Keep consistent data collection records

As an experimentalist, one of the worst feelings is to come back to your data directory and see a group of data files, `run1.csv`, `run2.csv`, `run3.csv` and not know what experimental protocol was run for each. Was `run1` the pilot? Maybe a little bit of personal archaeology with timestamps and version history can tell you the answer, but there is no guarantee.[22]

[22] We'll have a lot to say about this issue in Chapter 13.

	A	B	C	D	E	F	G
1	DOT	RA	SID	DOB	Gender	Consent	Source
2	12/14/12	ak, fp	ASD_01	9/19/98	m	1	fp
3	12/17/12	ak, fp	ASD_02	6/17/90	f	0	fp
4	12/18/12	ak, fp	ASD_03	8/15/90	f	1	fp
5	12/20/12	mf, fp	ASD_04	9/21/08	m	1	fp
6	1/21/13	mf, fp	ASD_05	8/31/07	m	1	fp
7	1/29/13	ak, ca	ASD_06	8/30/10	f	1	ah
8	1/31/13	ak, fp	ASD_07	10/26/05	m	1	fp

Figure 12.3: Part of a run sheet for a developmental study.

As well as collecting the actual data in whatever form they take (e.g., paper surveys, videos, or files on a computer), it is important to log **metadata** – data about your data – including relevant information like the date of data collection, the sample that was collected, the experiment version, the research assistants who were present, etc. The relevant meta-data will vary substantially from study to study – the important part is that you keep detailed records. Figure 12.3 and Figure 12.4 give two examples from our own research. The key feature is that they provide some persistent metadata about how the experiments were conducted.

12.4 Chapter summary: Data collection

In this chapter, we took the perspective of both the participant and the researcher. Our goal was to discuss how to achieve a good research outcome for both. On the side of the participant, we highlighted the responsibility of the experimenter to ensure a robust consent and debriefing process. We also discussed the importance of a good experimental experience in the lab and online – ensuring that the experiment is not only conducted ethically but is also pleasant to participate in. Finally, we discussed how to address some concerns about data quality from the researcher perspective, recommending both the extensive use of non-naive and naive pilot participants and the use of comprehension and manipulation checks.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Added a simple familiarization slide substitute that presents Bob and
shows that the experiment is about a person talking to you. Before
that, the familiarization slide was simply skipped.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
----------------------------
November 18 2013
50 subjects | Betting | No familiarization | Friend
var participant_response_type = 1;
var participant_feature_count = 1;
var linguistic_framing = 0;
var question_type = 0;
----------------------------
November 18 2013
50 subjects | Likert | No familiarization | Friend
var participant_response_type = 2;
var participant_feature_count = 1;
var linguistic_framing = 0;
var question_type = 2;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
The experiment now asked the subjects the referent of Bobs statement
at the bottom of the page. The previous experiments always had the
input field just below the stimuli or, in the case of 3fc hoovering
over the images did highlighted possible ones.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
----------------------------
November 30 2013 ~ 7 pm:
50 subjects | 3 forced choice condition | No familiarization | Friend
var participant_response_type = 0;
var participant_feature_count = 1;
var linguistic_framing = 0;
var question_type = 0;
```

Figure 12.4: Excerpt of a log for an iterative run of online experiments.

Discussion questions

1. "Citizen science" is a movement to have a broader base of individuals participate in research because they are interested in discoveries and want to help. In practice, citizen science projects in psychology like Project Implicit (https://implicit.harvard.edu/implicit/), Children Helping Science (https://lookit.mit.edu), and TheMusicLab.org (https://themusiclab.org) have all succeeded by offering participants a compelling experience. Check one of these out, participate in a study, and make a list the features that make it fun and easy to contribute data.
2. Be a Turker! Sign up for an account as an Amazon Mechanical Turk or Prolific Academic worker and complete a couple of tasks. How did you feel about browsing the list of tasks looking for work? What features of tasks attracted your interest? How hard was it to figure out how to participate in each task? And how long did it take to get paid?

Readings

– An introduction to online research: Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. Perspectives on Psychological Science, 13(2), 149-154. https://doi.org/10.1177/1745691617706516.

References

Allen, Michael. 2017. "Debriefing of Participants." In *The SAGE Encyclopedia of Communication Research Methods*. Vol. 1–4. Thousand Oaks, CA: Sage Publications.

Anderson, Craig A, Johnie J Allen, Courtney Plante, Adele Quigley-McBride, Alison Lovett, and Jeffrey N Rokkum. 2019. "The MTurkification of Social and Personality Psychology." *Personality and Social Psychology Bulletin* 45 (6): 842–50.

Benjamin, Ludy T. 2000. "The Psychology Laboratory at the Turn of the 20th Century." *American Psychologist* 55 (3): 318.

Buhrmester, Michael, Tracy Kwang, and Samuel D Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data?" *Perspectives on Psychological Science* 6 (1): 3–5.

Chuey, Aaron, Mika Asaba, Sophie Bridgers, Brandon Carrillo, Griffin Dietz, Teresa Garcia, Julia A Leonard, et al. 2021. "Moderated Online Data-Collection for Developmental Research: Methods and Replications." *Frontiers in Psychology*, 4968.

Cialdini, Robert B, and Noah J Goldstein. 2004. "Social Influence: Compliance and Conformity." *Annual Review of Psychology* 55 (1): 591–621.

Crump, Matthew J C, John V McDonnell, and Todd M Gureckis. 2013. "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research." *PLoS One* 8 (3): e57410.

De Leeuw, Joshua R. 2015. "jsPsych: A JavaScript Library for Creating Behavioral Experiments in a Web Browser." *Behavior Research Methods* 47 (1): 1–12.

DeMayo, Benjamin, Danielle Kellier, Mika Braginsky, Christina Bergmann, Cielke Hendriks, Caroline F Rowland, Michael Frank, and Virginia Marchman. 2021. "Web-CDI: A System for Online Administration of the MacArthur-Bates Communicative Development Inventories." *Language Development Research*.

Ebersole, Charles R, Olivia E Atherton, Aimee L Belanger, Hayley M Skulborstad, Jill M Allen, Jonathan B Banks, Erica Baranski, et al. 2016. "Many Labs 3: Evaluating Participant Pool Quality Across the Academic Semester via Replication." *J. Exp. Soc. Psychol.* 67 (November): 68–82.

Ejelöv, Emma, and Timothy J Luke. 2020. "'Rarely Safe to Assume': Evaluating the Use and Interpretation of Manipulation Checks in Experimental Social Psychology." *Journal of Experimental Social Psychology* 87: 103937.

Enkavi, A Zeynep, Ian W Eisenberg, Patrick G Bissett, Gina L Mazza, David P MacKinnon, Lisa A Marsch, and Russell A Poldrack. 2019. "Large-Scale Analysis of Test–Retest Reliabilities of Self-Regulation Measures." *Proceedings of the National Academy of Sciences* 116 (12): 5472–77.

Eyal, Peer, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. "Data Quality of Platforms and Panels for Online Behavioral Research." *Behavior Research Methods*, 1–20.

Fisher, Jill A. 2013. "Expanding the Frame of" Voluntariness" in Informed Consent: Structural Coercion and the Power of Social and Economic Context." *Kennedy Institute of Ethics Journal* 23 (4): 355–79.

Fitzpatrick, Emily FM, Alexandra LC Martiniuk, Heather D'Antoine, June Oscar, Maureen Carter, and Elizabeth J Elliott. 2016. "Seeking Consent for Research with Indigenous Communities: A Systematic Review." *BMC Medical Ethics* 17 (1): 1–18.

Gass, Robert H, and John S Seiter. 2018. *Persuasion: Social Influence and Compliance Gaining*. Routledge.

Gramlich, John. 2021. "America's Incarceration Rate Falls to Lowest Level Since 1995." https://www.pewresearch.org/fact-tank/2021/08/16/americas-incarceration-rate-lowest-since-1995/.

Hara, Kotaro, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. "A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.

Hauser, David J, Phoebe C Ellsworth, and Richard Gonzalez. 2018. "Are Manipulation Checks Necessary?" *Frontiers in Psychology* 9: 998.

Hawkins, Robert D, Michael C Frank, and Noah D Goodman. 2020. "Characterizing the Dynamics of Learning in Repeated Reference Games." *Cognitive Science* 44 (6): e12845.

Henrich, Joseph, Steven J Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2-3): 61–83.

Holmes, David S. 1976. "Debriefing After Psychological Experiments: I. Effectiveness of Postdeception Dehoaxing." *American Psychologist* 31 (12): 858.

Irani, Lilly C, and M Six Silberman. 2013. "Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 611–20.

Kadam, Rashmi Ashish. 2017. "Informed Consent Process: A Step Further Towards Making It Meaningful!" *Perspectives in Clinical Research* 8 (3): 107.

Litman, Leib, and Jonathan Robinson. 2020. *Conducting Online Research on Amazon Mechanical Turk and Beyond*. Sage Publications.

Litman, Leib, Jonathan Robinson, and Tzvi Abberbock. 2017. "TurkPrime. Com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences." *Behavior Research Methods* 49 (2): 433–42.

Maldonado, Mora, Ewan Dunbar, and Emmanuel Chemla. 2019. "Mouse Tracking as a Window into Decision Making." *Behavior Research Methods* 51 (3): 1085–1101.

Mason, Winter, and Siddharth Suri. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods* 44 (1): 1–23.

Montgomery, Jacob M, Brendan Nyhan, and Michelle Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62 (3): 760–75.

Moss, Aaron J, Cheskie Rosenzweig, Jonathan Robinson, and Leib Litman. 2020. "Demographic Stability on Mechanical Turk Despite COVID-19." *Trends in Cognitive Sciences* 24 (9): 678–80.

Office for Human Research Protections. 2003. "Prisoner Involvement in Research." https://www.hhs.gov/ohrp/regulations-and-policy/guidance/prisoner-research-ohrp-guidance-2003/index.html.

Oppenheimer, Daniel M, Tom Meyvis, and Nicolas Davidenko. 2009. "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology* 45 (4): 867–72.

Peer, Eyal, David M Rothschild, Zak Evernden, Andrew Gordon, and Ekaterina Damer. 2021. "MTurk, Prolific or Panels? Choosing the Right Audience for Online Research." *Social Science Research Network*. https://doi.org/https://doi.org/10.2139/SSRN.3765448.

Salehi, Niloufar, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, and Kristy Milland. 2015. "We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers." In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1621–30.

Scott, Kimberly, and Laura Schulz. 2017. "Lookit (Part 1): A New Online Platform for Developmental Research." *Open Mind* 1 (1): 4–14.

Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51 (3): 515.

Sieber, Joan E, and Michael J Saks. 1989. "A Census of Subject Pool Characteristics and Policies." *American Psychologist* 44 (7): 1053.

Slim, Mieke Sarah, and Robert J Hartsuiker. 2023. "Moving Visual World Experiments Online? A Web-Based Replication of Dijkgraaf, Hartsuiker, and Duyck (2017) Using PCIbex and WebGazer. Js." *Behavior Research Methods* 55 (7): 3786–3804.

Young, Daniel R, Donald T Hooker, and Fred E Freeberg. 1990. "Informed Consent Documents: Increasing Comprehension by Reducing Reading Level." *IRB: Ethics & Human Research* 12 (3): 1–5.

13 PROJECT MANAGEMENT

LEARNING GOALS

- Manage your research projects efficiently and transparently
- Develop strategies for data organization
- Optimize sharing of research products, like data and analysis code, by ensuring they are Findable, Accessible, Interoperable, Reusable (FAIR)
- Discuss potential ethical constraints on sharing research products

> Your closest collaborator is you six months ago, but you don't reply to emails.
>
> — Karl Broman (2016)

Have you ever returned to an old project folder to find a chaotic mess of files with names like `analysis-FINAL`, `analysis-FINAL-COPY`, and `analysis-FINAL-COPY-v2`? Which file is actually the final version!? Or perhaps you've spent hours searching for a data file to send to your advisor, only to realize with horror that it was *only* stored on your old laptop – the one that experienced a catastrophic hard drive failure when you spilled coffee all over it one sleepy Sunday morning. These experiences may make you sympathetic to Karl Broman's quip above. Good project management practices not only make it easier to share your research with others, they also make for a more efficient and less error prone workflow that will avoid giving your future self a headache. This chapter is about the process of managing all of the products of your research workflow – methodological protocols, materials[1], data, and analysis scripts. We focus especially on managing projects in ways that maximize their value to you and to the broader research community by aligning with open science practices (maximizing TRANSPARENCY).

When we talk about research products, we typically think of articles in academic journals, which have been scientists' main method of communication since the scientific revolution in the 1600s.[2] But articles only provide written summaries of research; they are not the original research products. In recent years, there have been widespread calls for increased sharing of research products, such as materials, data, and analysis code (Munafò et al. 2017). When shared appropriately, these

[1] We use the term "materials" here to cover a range of things another researcher might need in order to repeat your study, for example, stimuli, survey instruments, and code for computer-based experiments.

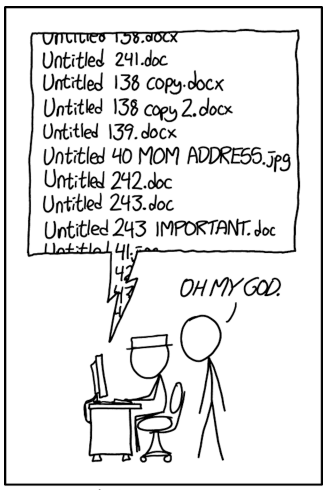


Figure 13.1: Poor file management creates chaos! "Documents" by xkcd (https://xkcd.com/1459, licensed under https://xkcd.com/license.html).

[2] The world's oldest scientific journal is the *Philosophical Transactions of the Royal Society*, first published in 1665.

other products can be as valuable as a summary article: Shared stimulus materials can be reused for new studies in creative ways; shared analysis scripts can allow for reproduction of reported results and become templates for new analyses; and shared data can enable new analyses or meta-analyses. Indeed, many funding agencies, and some journals, now require that research products be shared publicly, except when there are justified ethical or legal constraints, such as with sensitive medical data (Nosek et al. 2015).

Data sharing, in particular, has been the focus of intense interest. Sharing data is associated with benefits in terms of error detection (Hardwicke et al. 2021), creative re-use that generates new discoveries (Voytek 2016), increased citations (Piwowar and Vision 2013), and detection of fraud (Simonsohn 2013). According to surveys, researchers are usually willing to share data in principle (Houtkoop et al. 2018), but unfortunately, in practice, they often do not, even if you directly ask them (Hardwicke and Ioannidis 2018)! Often authors simply do not respond, but when they do, they frequently report that data have been lost because they were stored on a misplaced or damaged computer or drive, or team members with access to the data are no longer contactable (Tenopir et al. 2020).

As we have discussed in Chapter 3, even when data are shared, they are not always formatted in a way that they can be easily understood and re-used by other researchers, or even the original authors! This issue highlights the critical role of **metadata**: information that documents the data (and other products) that you share, including README files, **codebooks** that document datasets themselves, licenses that provide legal restrictions on reuse, etc. We will discuss best-practices for metadata throughout the chapter.

Sound project management practices and sharing of research projects are mutually reinforcing goals that bring benefits for both yourself, the broader research community, and scientific progress. One particularly important benefit of good project management practices is that they enable reproducibility. As we discussed in Chapter 3, computational reproducibility involves being able to trace the provenance of any reported analytic result in a research report back to its original source. That means being able to recreate the entire analytic chain from data collection to data files, though analytic specifications to the research results reported in text, tables, and figures. If data collection is documented appropriately, and if data are stored, organized, and shared, then the provenance of a particular result is relatively easy to verify. But once this chain is broken it can be hard to reconstruct , Figure 13.2. That's

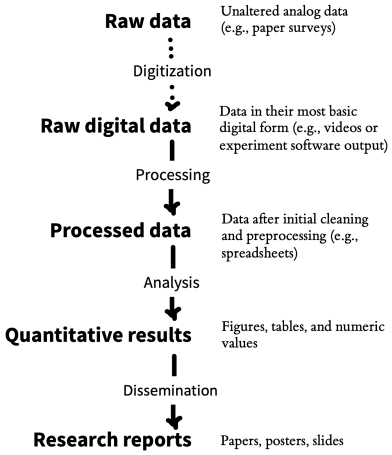


Figure 13.2: Illustration of the analytic chain from raw data through to research report.

why it's critical to build good project management practices into your research workflow right from the start.

In this chapter, you will learn how to manage your research project both efficiently and transparently.[3] Working towards these goals can create a virtuous cycle: if you organize your research products well, they are easier to share later, and if you assume that you will be sharing, you will be motivated to organize your work better! We begin by discussing some important principles of project management, including folder structure, file naming, organization, and version control. Then we zoom in specifically on data and discuss best practices for data sharing. We end by discussing the question of what research products to share and some of the potential ethical issues that might limit your ability to share in certain circumstances.

[3] This chapter – especially the last section – draws heavily on Klein et al. (2018), an article on research transparency that several of us contributed to.

Case study

ManyBabies, ManySpreadsheetFormats!

The ManyBabies project is an example of "Big Team Science" in psychology. A group of developmental psychology researchers (including some of us) were worried about many of the issues of reproducibility, replicability, and experimental methods that we've been discussing throughout this book, so they set up a large-scale collaboration to replicate key effects in developmental science. The first of these studies was ManyBabies 1 (The ManyBabies Consortium et al. 2020), a study of infants' preference for baby-talk (also known as "infant directed speech").

The core team expected a handful of labs to contribute, but after a year-long data collection period, they ended up receiving data from 69 labs around the world! The outpouring of interest signaled a lot of enthusiasm from the community for this kind of collaborative science. Unfortunately, it also made for a tremendous data management headache. All kinds of complications and hilarity ensued as the idiosyncratic data formatting preferences of the various labs were reorganised to fit into a single standardized analysis pipeline (Byers-Heinlein et al. 2020).

All of the specific formatting changes that individual labs made were reasonable – altering column names for clarity, combining templates into a single Excel file, changing units (e.g., from seconds to milliseconds) – but together they created a very challenging **data validation** problem for the core analysis team, requiring many dozens of hours of coding and hand-checking. The data checking was critical: an error in one lab's data was flagged during validation and led to the painful decision to drop those data from the final dataset. In future ManyBabies projects, the group has committed to using shared data validation software (https://manybabies.org/validator/) to ensure that data files uploaded by individual labs conform to a shared standard.

13.1 Principles of project management

A lot of project management problems can be avoided by following a very simple file organisation system.[4] For those researchers that "grew up" managing their files locally on their own computers and emailing colleagues versions of data files and manuscripts with names like `manuscript-FINAL-JS-rev1.xlsx`, a few aspects of this system may

[4] We're going to talk in this chapter about managing research products, which is one important part of project management. We won't talk about some other aspects of managing projects such as calendaring, managing tasks, or project communications. These are all important, they are just a bit out of scope for a book on doing experiments!

seem disconcerting. However, with a little practice, this new way of working will start to feel intuitive and have substantial benefits.

Here are the principles:

1. There should be exactly one definitive copy of each document in the project, with its name denoting what it is. For example, `fifo_manuscript.Rmd` or `fifo_manuscript.docx` is the write-up of the "fifo" project as a journal manuscript.
2. The location of each document should be within a folder which serves to uniquely identify the document's function within the project. For example, `/analysis/experiment1/eye_tracking_preprocessing.Rmd` is clearly the file that performs pre-processing for the analysis of eye-tracking data from Experiment 1.
3. The full project should be accessible to all collaborators via the cloud, either using a version control platform (e.g., <github.com>) or another cloud storage provider (e.g., Dropbox, Google Drive).
4. The revision history of all text- and text-based documents (minimally, data, analysis code, and manuscript files) should be archived automatically. Automatic versioning is the key feature of all version control systems and is often included by cloud storage providers.

Keeping these principles in mind, we discuss best practices for project organization, version control, and file naming.

13.1.1 Organizing your project

To the greatest extent possible, all files related to a project should be stored in the same project folder (with appropriate sub-folders), and on the same storage provider. There are cases where this is impractical due to the limitations of different software packages. For example, in many cases a team will manage its data and analysis code via github but decide to write collaboratively using google docs, overleaf, or another collaborative platform. (It can also be hard to ask all collaborators to use a version control system they are unfamiliar with.) In that case, the final paper should still be linked in some way to the project repository.[5]

Figure 13.3 shows an example project stored on the Open Science Framework. The top level folder contains sub-folders for analyses, materials, raw and processed data (kept separately). It also contains the paper manuscript, and, critically, a README file in a text format that describes the project. A README is a great way to document any

[5] The biggest issue that comes up in using a split workflow like this is the need to ensure reproducible written products, a process we cover in Chapter 14.

other metadata that the authors would like to be associated with the research products, for example a license, explained below.

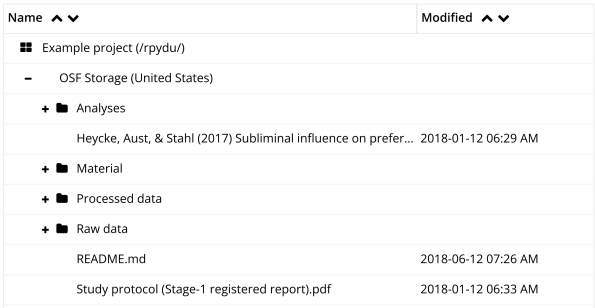


Figure 13.3: Sample top level folder structure for a project. From Klein et al. (2018). Original visible on the Open Science Framework (https://osf.io/xf6ug).

There are many reasonable ways to organize the sub-folders of a research project, but the broad categories of materials, data, analysis, and writing are typically present.[6] In some projects – such as those involving multiple experiments or complex data types – you may have to adopt a more complex structure. In many of our projects, it's not uncommon to find paths like `/data/raw_data/exp1/demographics`. The key principle is to create a hierarchical structure in which subfolders uniquely identify the part of the broader space of research products that are found inside them – that is, `/data/raw_data/exp1` contains all the raw data from Experiment 1, and `/data/raw_data/exp1/demographics` contains all the raw *demographics* data from that particular experiment.

[6] We like the scheme followed by Project TIER (https://www.projecttier.org), which provides very clear guidance about file structure and naming conventions. TIER is primarily designed for a copy-and-paste workflow, which is slightly different from the "dynamic documents" workflow that we primarily advocate for (e.g., using R Markdown or Quarto as in Appendix C).

13.1.2 Versioning

Probably everyone who has ever collaborated electronically has experienced the frustration of editing a document, only to find out that you are editing the wrong version – perhaps some of the problems you are working on have already been corrected, or perhaps the section you are adding has already been written by someone else. A second common source of frustration comes when you take a wrong turn in a project, perhaps by reorganizing a manuscript in a way that doesn't work or refactoring code in a way that turns out to be short-sighted.

These two problems are solved by modern version control systems. Here we focus on the use of **git**, which is the most widely used version control system. Git is a great general solution for version control, but many people – including several of us – don't love it for collaborative manuscript writing. We'll introduce git and its principles here, while noting that online collaboration tools like Google Docs and Overleaf[7] can be easier for writing prose (as opposed to code); we cover this topic in a bit more depth in Chapter 14.

[7] Overleaf is actually supported by git on the backend!

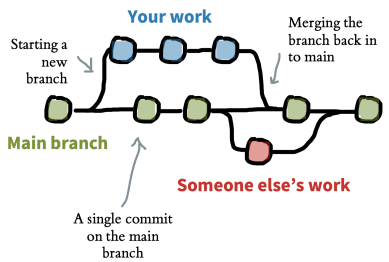


Figure 13.4: Visualisation of Git version control showing a series of commits (circles) on three different branches: the main branch (green) and two others (blue and red). Branches can be created and then merged back into the main branch.

Git is a tool for creating and managing projects, which are called **repositories**. A Git repository is a directory whose revision history is tracked via a series of **commits** – snapshots of the state of the project. These commits can form a tree with different **branches**, as when two contributors to the project are working on two different parts simultaneously (Figure 13.4). These branches can later be **merged** either automatically or via manual intervention in the case of conflicting changes.

Commonly, Git repositories are hosted by an online service like Github[8] to facilitate collaboration. With this workflow. a user makes changes to a local version of the repository on their own computer and **pushes** those changes to the online repository. Another user can then **pull** those changes from the online repository to their own local version. The online "origin" copy is always the definitive copy of the project and a record is kept of all changes. Appendix B provides a practical introduction to Git and Github, and there are a variety of good tutorials available online and in print (Blischak, Davenport, and Wilson 2016).

[8] https://github.com

Collaboration using version control tools is designed to solve many of the problems we've been discussing:

- A remotely hosted Git repository is a cloud-based backup of your work, meaning it is less vulnerable to accidental erasure.[9]
- By virtue of having versioning history, you have access to previous drafts in case you find you have been following a blind alley and want to roll back your changes.
- By creating new branches, you can create another, parallel history for your project, so that you can try out major changes or additions without disturbing the main branch in the process.
- A project's commit history is labeled with each commit's author and date, facilitating record keeping and collaboration.
- Automatic merging can allow synchronous editing of different parts of a manuscript or codebase.[10]

[9] In 48BC, Julius Caesar accidentally burned down part of the Great Library of Alexandria where the sole copies of many valuable ancient works were stored. To this day, many scientists have apparently retained the habit of storing single copies of important information in vulnerable locations. Even in the age of cloud computing, hard drive failure is a surprisingly common source of problems!

[10] Version control isn't magic, and if you and a collaborator edit the same paragraph or function, you will likely have to merge your changes by hand. But Git will at least show you where the conflict is!

Organizing a project repository for collaboration and hosting on a remote platform is an important first step towards sharing! Many of our projects (like this book) are actually **born open**: we do all of our work on a publicly hosted repository for everyone to see (Rouder 2015). This philosophy of "working in the open" encourages good organization practices from the beginning. It can feel uncomfortable at first, but this discomfort soon vanishes as you realize that basically no one is looking at your in-progress project.

One concern that many people raise about sharing in-progress research openly is the possibility of "scooping" – that is, other researchers getting an idea or even data from the repository and writing a paper before

you do. We have two responses to this concern. First, the empirical frequency of this sort of scooping is difficult to determine, but likely very low – we don't know of any documented cases. Mostly, the problem is getting people to care about your experiment at all, not people caring so much that they would publish using your data or materials! In Gary King's words (King and Shieber 2013), "The thing that matters the least is being scooped. The thing that matters the most is being ignored." On the other hand, if you are in an area of research that you perceive to be competitive, or where there is some significant risk of this kind of shenanigans, it's very easy to keep part, or all, of a repository, private among your collaborators until you are ready to share more widely. All of the benefits we described still accrue. For an appropriately organized and hosted project, often the only steps required to share materials, data, and code are 1) to make the hosted repository public and 2) to link it to an archival storage platform like the Open Science Framework.

13.1.3 File names

As Phil Karlton reportedly said[11], "There are only two hard things in Computer Science: cache invalidation and naming things." What's true for computer science is true for research in general.[12] Naming files is hard! Some very organized people survive on systems like `INFO-r1-draft-2020-07-13-js.docx` - meaning, "the INFO project revision 1 draft of July 13th, 2020, with edits by JS." But this kind of system needs a lot of rules and discipline, and it requires everyone in a project to buy in completely.

On the other hand, if you are naming a file in a hierarchically organized version control repository, the naming problem gets dramatically easier. All of a sudden, you have a context in which names make sense. `data.csv` is a terrible name for a data file on its own. But the name is actually perfectly informative – in the context of a project repository with a README that states that there is only a single experiment, a repository structure such that the file lives in a folder called `raw_data`, and a commit history that indicates the file's commit date and author.

As this example shows, naming is hard *out of context*. So here's our rule: name a file with what it contains. Don't use the name to convey the context of who edited it, when, or where it should go in a project. That is metadata that the platform should take care of.[13]

[11] https://www.karlton.org/2017/12/naming-things-hard/

[12] We won't talk about cache invalidation; that's a more technical problem in computer science that is beyond the scope of this book.

[13] The platform won't take care of it if you email it to a collaborator – precisely why you should share access to the full *platform*, not just the out-of-context file!

13.2 Data Management

We've just discussed how to manage projects in general; in this section we zoom in on datasets specifically. Data are often the most valuable research product because they represent the evidence generated by our research. We maximize the value of the evidence when other scientists can reuse it for independent verification or generation of novel discoveries. Yet lots of research data are not reusable, even when they are shared. In Chapter 3, we discussed Hardwicke et al. (2018)'s study of *analytic* reproducibility. But before we were even able to try and reproduce the analytic results, we had to look at the data. When we did that, we found that only 64% of shared datasets were both complete and understandable.

How can you make sure that your data are managed so as to enable effective sharing? We make four primary recommendations:

1. save your raw data
2. document your data collection process
3. organize your raw data for later analysis
4. document your data using a codebook or other metadata

Let's look at each in turn.

13.2.1 Save your raw data

Raw data take many forms. For many of us, the raw data are those returned by the experimental software; for others, the raw data are videos of the experiment being carried out. Regardless of the form of these data, save them! They are often the only way to check issues in whatever processing pipeline brings these data from their initial state to the form you analyze. They also can be invaluable for addressing critiques or questions about your methods or results later in the process. If you need to correct something about your raw data, *do not alter the original files*. Make a copy, and make a note about how the copy differs from the original.[14]

Raw data are often not anonymized – or even anonymizable. Anonymizing them sometimes means altering them (e.g., in the case of downloaded logs from a service that might include IDs or IP addresses). Or in some cases, anonymization is difficult or impossible without significant effort and loss of some value from the data, e.g. for video data or MRI data (Bischoff-Grethe et al. 2007). Unless you have specific permission for broad distribution of these identifiable data, the raw data may then need to be stored in a different way. In these cases, we recommend saving your raw data in a separate repository

[14] Future you will thank present you for explaining why there are two copies of subject 19's data after you went back and corrected a typo.

with the appropriate permissions. For example, in the ManyBabies 1 study we described above, the public repository does not contain the raw data contributed by participating labs, which the team could not guarantee was anonymized; these data are instead stored in a private repository.[15]

You can use your repository's README to describe what is and is not shared. For example, a README might state that "We provide anonymized versions of the files originally downloaded from Qualtrics" or "Participants did not provide permission for public distribution of raw video recordings, which are retained on a secure university server." Critically, if you share the derived tabular data, it should still be possible to reproduce the analytic results in your paper, even if checking the provenance of those numbers from the raw data is not possible for every reader.[16]

One common practice is the use of participant identifiers to link specific experimental data – which, if they are responses on standardized measures, rarely pose a significant identifiability risk – to demographic data sheets that might include more sensitive and potentially identifiable data.[17] Depending on the nature of the analyses being reported, the experimental data can then be shared with limited risk. Then a selected set of demographic variables – for example, those that do not increase privacy risks but are necessary for particular analyses – can be distributed as a separate file and joined back into the data later.

13.2.2 Document your data collection process

In order to understand the meaning of the raw data, it's helpful to share as much as possible about the context in which they were collected. This practice also helps communicate the experience that participants had in your experiment. Documentation of this experience can take many forms.

If the experimental experience was a web-based questionnaire, archiving this experience can be as simple as downloading the questionnaire source.[18] For more involved studies, it can be more difficult to reconstruct what participants went through. This kind of situation is where video data can shine (Gilmore and Adolph 2017). A video recording of a typical experimental session can provide a valuable tutorial for other experimenters – as well as good context for readers of your paper. This is doubly true if there is a substantial interactive element to your experimental experience, as is often the case for experiments with children. For example, in our ManyBabies case study, the project shared "walk through" videos of experimental sessions[19] for many of the participat-

[15] The precise repository you use for this task is likely to vary by the kind of data that you're trying to store and the local regulatory environment. For example, in the United States, to store deanonymized data with certain fields requires a server that is certified for HIPAA (the relevant privacy law). Many – but by no means all – universities provide HIPAA-compliant cloud storage.

[16] One way we organize the raw data in some of our paper is to have three different subfolders in the `data/` directory: `raw/`, for the original data; `processed/`, for the anonymized or otherwise preprocessed data; and `/scripts`, for the code that does the preprocessing. Since these folders are in a git repository, we can then add `raw/*` to the `.gitignore` file, ensuring that they are never added to the public version of the repository even though they sit within our local file hierarchy in the appropriate place.

[17] A word about subject identifiers. These should be anonymous identifiers, like randomly generated numbers, that cannot be linked to participant identities (like data of birth) and are unique. You laugh, but one of us was in a lab where all the subject IDs were the date of test and the initials of the participant. These were neither unique nor anonymous. One common convention is to give your study a code-name and to number participants sequentially, so your first participant in a sequence of experiments on information processing might be `INFO-1-01`.

[18] If it's in a proprietary format like a Qualtrics `.QSF` file, a good practice is to convert it to a simple plain text format as well so it can be opened and re-used by folks who do not have access to Qualtrics (which may include future you!).

[19] https://nyu.databrary.org/volume/896

ing labs, creating a repository of standard experiences for infant development studies. If nothing else, a video of an experimental session can sometimes be a very nice archive of a particular context.[20]

[20] Videos of experimental sessions also are great demos to show in a presentation about your experiment, provided you have permission from the participant.

Regardless of what specific documentation you keep, it's critical to create some record linking your data to the documentation. For a questionnaire study, for example, this documentation might be as simple as a README that says that the data in the `data/raw/` directory were collected on a particular date using the file named `experiment1.qsf`. This kind of "connective tissue" linking data to materials can be very important when you return to a project with questions. If you spot a potential error in your data, you will want to be able to examine the precise version of the materials that you used to gather those data in order to identify the source of the problem.

13.2.3 Organize your data for later analysis: Spreadsheets

Data come in many forms, but chances are that at some point during your project you will end up with a spreadsheet full of information. Well-organized spreadsheets cam mean the difference between project success and failure! A wonderful article by Broman and Woo (2018) lays out principles of good spreadsheet design. We highlight some of their principles here (with our own, opinionated ordering):

1. *Make it a rectangle.*[21] Nearly all data analysis software, like SPSS, Stata, Jamovi and JASP (and many R packages), require data to be in a tabular format.[22] If you are used to analyzing data exclusively in a spreadsheet, this kind of tabular data isn't quite as readable, but readable formatting gets in the way of almost any analysis you want to do. Figure 13.5 gives some examples of non-rectangular spreadsheets. All of these will cause any analytic package to choke because of inconsistencies in how rows and columns are used!

[21] Think of your data like a well-ordered plate of sushi, neatly packed together without any gaps.

[22] Tabular data is a precursor to "tidy" data, which we describe in more detail in Appendix Appendix D.

	A	B	C	D	E	F
1						
2		101	102	103	104	105
3	sex	Male	Female	Male	Male	Male
4						
5		101	102	103	104	105
6	glucose	134.1	120.0	124.8	83.1	105.2
7						
8		101	102	103	104	105
9	insulin	0.60	1.18	1.23	1.16	0.73

	A	B	C	D	E	F	G
1	1MIN						
2			Normal			Mutant	
3	B6	146.6	138.6	155.6	166	179.3	186.9
4	BTBR	245.7	240				
5							
6	5MIN						
7			Normal			Mutant	
8	B6	333.6	353.6	408.8	450.6	474.4	423.8
9	BTBR	514.4	610.6	597.9	412.1	447.4	446.5

	A	B	C	D	E	F	G
1							
2	Date	11/3/14					
3	Days on Diet	126					
4	Mouse #	43					
5	sex	f					
6	experiment		values			mean	SD
7	control		0.186	0.191	1.081	0.49	0.52
8	treatment A		7.414	1.468	2.254	3.71	3.23
9	treatment B		9.811	9.259	11.296	10.12	1.05
10							
11	fold change		values			mean	SD
12	treatment A		15.26	3.02	4.64	7.64	6.65
13	treatment B		20.19	19.05	23.24	20.83	2.17

	A	B	C	D	E	F
1		GTT date	GTT weight	time	glucose mg/dl	insulin ng/ml
2	321	2/9/15	24.5	0	99.2	lo off curve
3				5	349.3	0.205
4				15	286.1	0.129
5				30	312	0.175
6				60	99.9	0.122
7				120	217.9	lo off curve
8	322	2/9/15	18.9	0	185.8	0.251
9				5	297.4	2.228
10				15	439	2.078
11				30	362.3	0.775
12				60	232.7	0.5
13				120	260.7	0.523
14	323	2/9/15	24.7	0	198.5	0.151
15				5	530.6	off curve lo

Figure 13.5: Examples of non-rectangular spreadsheet formats that are likely to cause problems in analysis. Adapted from Broman and Woo (2018).

2. *Choose good names for your variables*. No one convention for name formatting is best, but it's important to be consistent. We tend to follow the tidyverse style guide[23] and use lowercase words separated by underscores (_). It's also helpful to give units where these are available, e.g., are reaction times in seconds or milliseconds. Table 13.1 gives some examples of good and bad variable names.

[23] https://style.tidyverse.org

Table 13.1: Examples of good and bad variable names. Adapted from Broman and Woo (2018).

Good name	Good alternative	Avoid
subject_id	SubID	subject #
sex	female	M/F
rt_msec	reaction_time_ms	reaction time (millisec.)

3. *Be consistent with your cell formatting*. Each column should have one *kind* of thing in it. For example, if you have a column of numerical values, don't all of a sudden introduce text data like "missing" into one of the cells. This kind of mixing of data types can cause havoc down the road. Mixed or multiple entries also don't work, so don't write "0 (missing)" as the value of a cell. Leaving cells blank is also risky because it is ambiguous. Most software packages have a standard value for missing data (e.g. `NA` is what R uses). If you are writing dates, please be sure to use the "global standard" (ISO 8601), which is YYYY-MM-DD. Anything else can be misinterpreted easily.[24]

4. *Decoration isn't data*. Decorating your data with bold headings or highlighting may seem useful for humans, but it isn't uniformly interpreted or even recognized by analysis software (e.g., reading an Excel spreadsheet into R will scrub all your beautiful highlighting and artistic fonts) so do not rely on it.

5. *Save data in plain text files*. The CSV (comma-delimited) file format is a common standard for data that is uniformly understood by most analysis software (it is an "interoperable" file format).[25] The advantage of CSVs is that they are not proprietary to Microsoft or another tech company and can be inspected in a text editor, but be careful: they do not preserve Excel formulas or formatting!

[24] Dates in Excel deserve special mention as a source of terribleness. Excel has an unfortunate habit of interpreting information that has nothing to do with dates as dates, destroying the original content in the process. Excel's issue with dates has caused unending horror in the genetics literature, where gene names are automatically converted to dates, sometimes without the researchers noticing (Ziemann, Eren, and El-Osta 2016). In fact, some gene names have had to be changed in order to avoid this issue!

[25] Be aware of some interesting differences in how these files are output by European vs. American versions of Microsoft Excel! You might find semi-colons instead of commas in some datasets.

Given the points above, we recommend that you avoid analyzing your data in Excel. If it is necessary to analyze your data in a spreadsheet program, we urge you to save the raw data as a separate CSV and then create a distinct analysis spreadsheet so as to be sure to retain the raw data unaltered by your (or Excel's) manipulations.

13.2.4 Organize your data for later analysis: Software

Many researchers do not create data by manually entering information into a spreadsheet. Instead they receive data as the output from a web platform, software package, or device. These tools typically provide researchers limited control over the format of the resulting tabular data export. Case in point is the survey platform Qualtrics, which – at least at hte moment – provides data with not one but two header rows, complicating import into almost all analysis software![26]

That said, if your platform *does* allow you to control what comes out, you can try to use the principles of good tabular data design outlined above. For example, try to give your variables (e.g., questions in Qualtrics) sensible names!

[26] The R package `qualtRics` can help with this.

ACCIDENT REPORT

Bad variable naming can lead to analytic errors!

In our methods class, students often try to reproduce the original analyses from a published study before attempting to replicate the results in a new sample of participants. When Kengthsagn Louis looked at the code for the study she was interested in, she noticed that the variables in the analysis code were named horribly (presumably because they were output this way by the survey software). For example, one piece of Stata code looked like this:

```
gen recall1=.
replace recall1=0 if Q21==1
replace recall1=1 if Q21==3 | Q21==5 | Q21==6
replace recall1=2 if Q21==2 | Q21==4 | Q21==7 | Q21==8
replace recall1=0 if Q69==1
replace recall1=1 if Q69==3 | Q69==5 | Q69==6
replace recall1=2 if Q69==2 | Q69==4 | Q69==7 | Q69==8
ta recall1
```

In the process of translating this code into R in order to reproduce the analyses, Kengthsagn and a course teaching assistant, Andrew Lampinen, noticed that some participant responses had been assigned to the wrong variables. Because the variable names were not human-readable, this error was almost impossible to detect. Since the problem affected some of the inferential conclusions of the article, the article's author – to their credit – issued an immediate correction (Petersen 2019).

The moral of the story: Obscure variable names can hide existing errors and create opportunities for further error! Sometimes you can adjust these within your experimental software, avoiding the issue. If not, make sure to create a "key" and translate the names immediately, double checking after you are done.

13.2.1 Document the format of your data

Even the best-organized tabular data are not always easy to understand by other researchers, or even yourself, especially after some time has passed. For that reason, you should make a **codebook** (also known as a **data dictionary**) that explicitly documents what each variable is. Figure 13.7 shows an example codebook for the trial-level data in the bottom of Figure 13.6. Each row represents one variable in the associated dataset. Codebooks often describe what type of variable a column is (e.g., numeric, string), and what values can appear in that column. A human-readable explanation is often given as well, providing providing units (e.g., "seconds") and a translation of numeric codes (e.g., "test condition is coded as 1") where relevant.

	A	B	C	D	E	F	G	H	I
1	lab	subid	method	RA	age_days	trial_order	session_error	session_error_type	notes
2	babylab_nijmegen	ba01_6-9	HPP	KM	245	1	noerror	NA	teeth may be painful
3	babylab_nijmegen	ba02_6-9	HPP	KM	206	4	noerror	NA	NA
4	babylab_nijmegen	ba03_6-9	HPP	KM	257	3	noerror	NA	NA
5	babylab_nijmegen	ba04_6-9	HPP	KM	245	2	error	baby cried	teeth may be painful
6	babylab_nijmegen	ba05_6-9	HPP	KM	208	2	noerror	NA	baby was sick 2 months ago

	A	B	C	D	E	F
1	lab	subid	trial_type	stimulus	trial_num	looking_time
2	babylab_nijmegen	ba01_6-9	training	train1	-2	18.02
3	babylab_nijmegen	ba01_6-9	training	train2	-1	9.05
4	babylab_nijmegen	ba01_6-9	IDS	IDS1	1	17.48
5	babylab_nijmegen	ba01_6-9	ADS	ADS1	2	5.51
6	babylab_nijmegen	ba01_6-9	IDS	IDS2	3	16.34
7	babylab_nijmegen	ba01_6-9	ADS	ADS2	4	13.9

Figure 13.6: Example participant (top) and trial (bottom) level data from the ManyBabies (2020) case study.

	A	B	C	D
1	Variable Name	Type	Possible Values	Explanation
2	lab	string	<your lab ID>	your unique lab ID
3	subid	string	<participant ID codes>	unique (within lab) ID for the participant
4	trial_type	string	`IDS`, `ADS`, and `training`	stimulus type on this trial
5	stimulus	string	IDS-x', 'ADS-x', 'training'	the actual sound file that was playing
6	trial_num	integer	-2, -1, 1-8	trial number, from 1 — 8 (with -2 and -1 denoting training trials)
7	looking_time	double	range 0-20	looking time in seconds

Figure 13.7: Codebook for trial-level data (see above) from the ManyBabies (2020) case study.

Creating a codebook need not require a lot of work. Almost any documentation is better than nothing! There are also several R packages that can automatically generate a codebook for you, for example `codebook`, `dataspice`, and `dataMaid` (Arslan 2019). Adding a codebook can substantially increase the reuse value of the data and prevent hours of frustration as future you and others try to decode your variable names and assumptions.

13.3 Sharing Research Products

As we've been discussing throughout this chapter, if you've managed your research products effectively, sharing them with others is a far less daunting prospect, and usually just requires uploading them to an online

repository like the Open Science Framework. This section addresses some potential limitations on sharing that you should bear in mind and discusses where and how to share research products.

13.3.1 What you can and can't share

We've been advocating that you share all of your research products, especially your data. In practice, however, **participant privacy** (as well as a few other constraints) limits what you can share. Luckily, there are some concrete steps you can take to make sure that you protect participants and comply with your obligations while still realizing the benefits of data sharing.

Unless they explicitly waive their rights, participants in psychology experiments have the expectation of privacy – that is, no one should be able to identify them from the data they have provided. Protecting participant privacy is an important part of researchers' ethical responsibilities (Ross, Iguchi, and Panicker 2018), and needs to be balanced against the ethical imperatives to share (see Chapter 4).[27]

[27] Meyer (2018) gives an excellent overview of how to navigate various legal and ethical issues around data sharing in the US context.

Furthermore, there are legal regulations that protect participants' data, though these vary from country to country. In the US, the relevant regulation is **HIPAA**, the Health Insurance Portability and Accountability Act, which limits disclosures of private health information (**PHI**). In the European Union, the relevant regulation is the European **GDPR** (General Data Protection Regulation). It's beyond the scope of this book to give a full treatment of these regulatory frameworks; you should consult with your local ethics board regarding compliance, but here is the way we have navigated this situation while still sharing data.

Under both frameworks, **anonymization** (or equivalently **de-identification**) of data is a key concept, such that data sharing is generally just fine if the data meet the relevant standard. Under US guidelines, researchers can follow the "safe harbor" standard[28] under which data are considered to be anonymized if they do not contain identifiers like names, telephone numbers, email addresses, social security numbers, dates of birth, faces, etc. Thus, data that only contain participant IDs and nothing from this list can typically be shared without participant consent without a problem.[29]

[28] As described on the relevant DHHS page (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html).

[29] US IRBs are a very de-centralized bunch and their interpretations often vary considerably. For reasons of liability or ethics, they may not allow data sharing even though it is permitted by US law. If you feel like arguing with an IRB that takes this kind of stand, you could mention that the DHHS rule actually doesn't consider de-identified data to be "human subjects" data at all, and thus the IRB may not have regulatory authority over it. We're not lawyers, and we're not sure if you'll succeed but it could be worth a try.

The EU's GDPR also allows fully anonymized data sharing, with one big complication. Putting anonymous identifiers in a data file and removing identifiable fields does not itself suffice for GDPR anonymization if the data are still **in-principle re-identifiable** because you have maintained documentation linking IDs to identifiable data like names or email addresses. Only when the key linking identifiers to data has

been destroyed are the data truly de-identified according to this standard.

Accident report

Really anonymous?

When we first began teaching Psych 251, our experimental methods course at Stanford, one of the biggest contributions of the course was simply showing students how to do experiments online. Amazon's Mechanical Turk crowdsourcing service was relatively new, and our IRB did not have a good sense of what this service really was. We proposed that we would share data from the class and received approval for this practice. Our datasets were downloaded directly from Mechanical Turk and included participants' MTurk IDs (long alphanumeric strings that seemed completely anonymous). Several experiences caused us to reconsider this practice!

First, we discovered that MTurk IDs were in some cases linked to study participants' public Amazon "wish lists," which could both inadvertently provide information about the participant and also even potentially provide a basis for reidentification (in rare cases). This discovery led us to consult with our IRB and provide more explicit consent language in our class experiments, linking to instructions for making Amazon profiles private.

Then, a little later we received an irate email from an MTurk participant who had discovered their data on github via a search for their MTurk ID. Although they were not identified in this dataset, it convinced us that at least some participants would not like this ID shared. After another consultation with the IRB, we apologized to this individual and removed their and others' IDs from our github commit histories across that and other repositories. Prior to posting data, we now take care to anonymize IDs by creating a secret mapping between the IDs we post and the actual MTurk IDs.

De-identification is not always enough. As datasets get richer, **statistical reidentification risks** go up substantially such that, with a little bit of outside information, data can be matched with a unique individual. These risks are especially high with linguistic, physiological, and geospatial data, but they can be present even for simple behavioral experiments. In one influential demonstration, knowing a person's location on two occasions was often enough to identify their data uniquely in a huge database of credit card transactions (De Montjoye et al. 2015).[30] Thus, simply removing fields from the data is a good starting point – but if you are collecting richer data about participants' behavior you may need to consult an expert.

Privacy issues are ubiquitous in data sharing, and almost every experimental research project will need to solve them before sharing data. For simple projects, often these are the only issues that preclude data sharing. However, in more complex projects, other concerns can arise. Funders may have specific mandates regarding where your data should be shared. Data use agreements or collaborator preferences may restrict where and when you can share. And certain data types require much more sensitivity since they are more consequential than, say, the reaction times on a Stroop task. We include here a set of questions to walk through to plan

[30] For an example closer to home, many of the contributing labs in the ManyBabies project logged the date of test for each participant. This useful and seemingly innocuous piece of information is unlikely to identify any particular participant – but alongside a social media post about a lab visit or a dataset about travel records, it could easily reveal a particular participant's identity.

your sharing (Figure 13.8). When in doubt, it's often a good idea to consult with the relevant local authority, e.g. your ethics board for ethical issues or your research management office for regulatory issues.

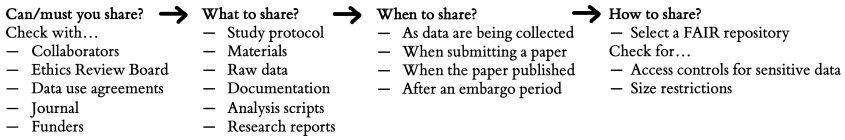


Figure 13.8: A decision chart for thinking about sharing research products. Adapted from Klein et al. (2018).

13.3.1 Where and how to share: the FAIR principles

For shared research products[31] to be usable by others, they should meet the FAIR standard by being Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016).

- **Findable** products are easily discoverable to both humans and machines. That means linking to them in research reports using unique persistent identifiers (e.g. a digital object identifier [DOI]).[32] and attaching them with metadata describing what they are so they can be indexed by search engines.
- **Accessibility** means that research products need to be preserved across the long-term and are retrievable via their standardized identifier.
- **Interoperability** means that the research products needs to be in a format that people and machines (e.g., search engines and analysis software) can understand.
- **Reusable** means that the research products need to be well organized, documented, and licensed so that others know how to use them.

If you've followed the guidance in the rest of this chapter, then you will already be well on your way to making your research products FAIR. There are a few final steps to consider. An important decision is where you are going to share the research products. We recommend uploading the files to a repository that's designed according to support FAIR principles. Personal websites don't cut it, since these sites tend to go out of date and disappear. There's also no easy way to find research products on personal sites unless you know who created them. Github, though it's a great platform for collaboration, isn't a FAIR repository – for one thing, products there don't have DOIs[33] – and there are no archival guarantees on files that are shared there. Perhaps surprisingly for some researchers, journal supplementary materials are also not a great place to put research products. Often supplementary materials are assigned no

[31] Most of this discussion is about data, because that's where the community has focused its efforts. That said, almost everything here applies to other research products as well!

[32] DOIs are those long URL-like things that are often used to link to papers. Turns out they can also be associated with datasets and other research products. Critically, they are guaranteed to work to find stuff, whereas standard web URLs often go stale after several years when people refactor their website. Most online repositories, like the Open Science Framework, will issue DOIs for the research products you store there.

[33] You can get a DOI for github software through a partnership with Zenodo (zenodo.org), a FAIR-compliant repository.

unique DOI or metadata, have limited supported formats, and have no persistence guarantees (Evangelou, Trikalinos, and Ioannidis 2005).

Fortunately, there are many repositories that help you conform to FAIR standards. Zenodo, Figshare, the Open Science Framework (OSF), and the various Dataverse sites are designed for this purpose, though there are many other domain-specific repositories that are particularly relevant for different research fields. We often use the OSF as it makes it easy to share all research products connected to a project in one place. OSF is FAIR compatible and allows users to assign DOIs to their data and provide appropriate metadata.

We recommend you attach a license to your research products. Academic culture is (usually) unburdened by discussion of intellectual property and legal rights and instead relies on scholarly norms about citation and attribution. The basic expectation is that if you rely on someone else's research, you explicitly acknowledge the relevant journal article through a citation. Although norms are still evolving, using research products created by others generally adheres to the same scholarly principle. Research products can also be useful in non-academic contexts, however. Perhaps you created software that a company would like to use. Maybe a pediatrician would like to use a research instrument you've been working on to assess their patients. These applications (and many other reuses of the data) require a legal license. In practice, there are a number of simple, open source licenses that permit reuse. We tend to favor Creative Commons licenses[34], which come in a variety of flavors such as CC0[35] (which allows all reuse), CC-BY[36] (which allows reuse as long as there is attribution), and CC-BY-NC[37] (which only allows attributed, non-commercial reuse).[38] Regardless of what license you choose, having a license means that your products won't be in a "not sure what I'm allowed to do with this" limbo for others who are interested in reusing them.

As we have discussed, you may want to consider storing your work in a public repository from the outset of the project. If you are using Github to manage your project, you can link the Git repository to the Open Science Framework so it automatically syncs. This provides a valuable incentive to organize your work properly throughout your project and makes sharing super easy, because you've already done it! On the other hand, this way of working can feel exposed for some researchers, and it does carry some risks, however small, of "scooping" or pre-emption by other groups working in the same space. Fortunately you can set up the same Git-OSF workflow and keep it private until your ready to make it public later on.

[34] https://creativecommons.org

[35] https://creativecommons.org/share-your-work/public-domain/cc0/

[36] https://creativecommons.org/licenses/by/4.0/

[37] https://creativecommons.org/licenses/by/4.0/

[38] Klein et al. (2018) recommend the CC0 license, which puts no limits on what can be done with your data. At first blush it may seem like a license that requires attribution is useful. But academic norms, rather than the threat of litigation, lead to good citation practices. In addition, more restrictive licenses can mean that some legitimate uses of your data or research can be blocked.

The next stage at which you should consider sharing your research products is when you submit your study to a journal. If you're still hesitant to make the project entirely public, many repositories (including OSF) will allow you to create special links that facilitate limited access to, for example, reviewers and editors. In general, the earlier you share your research products the better because there are more opportunities for others to learn from, build on, and verify your research.[39] But if neither of these options seem appealing, please do share your research products once your paper is accepted. Doing so will increase the value (and the impact) of your publication.

[39] If there are errors in our work, we'd certainly love to hear about it *before* the article is published in a journal rather than after!

13.4 Chapter summary

All of the hard work you put into your experiments – not to mention the contributions of your participants – can be undermined by bad data and project management. As our accident reports and case study show, bad organizational practices can at a minimum cause huge headaches. Sometimes the consequences can be even worse. On the flip side, starting with a firm organizational foundation sets your experiment up for success. These practices also make it easier to share all of the products of your research, not just your findings. Such sharing is both useful for individual researchers and for the field as a whole.

DISCUSSION QUESTIONS

1. Find an Open Science Framework repository that corresponds to a published paper. What is their strategy for documenting what is shared? How easy is it to figure out where everything is and if the data and materials sharing is complete?
2. Open up the US Department of Health and Human Services "safe harbor" standards (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html) and navigate to the section called "The De-identification Standard." Go through the list of identifiers that must be removed. Are there any on this list that you would need to include in your dataset in order to conduct your own research? Can you think of any others that do not fall on this list?

READINGS

– A more in-depth tutorial on various aspects of scientific openness: Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., Ijzerman, H., Nilsonne, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*, 20. https://doi.org/10.1525/collabra.158.

Part V

Reporting

References

Arslan, Ruben C. 2019. "How to Automatically Document Data with the Codebook Package to Facilitate Data Reuse." *Advances in Methods and Practices in Psychological Science* 2 (2): 169–87.

Bischoff-Grethe, Amanda, I Burak Ozyurt, Evelina Busa, Brian T Quinn, Christine Fennema-Notestine, Camellia P Clark, Shaunna Morris, et al. 2007. "A Technique for the Deidentification of Structural Brain MR Images." *Human Brain Mapping* 28 (9): 892–903.

Blischak, John D, Emily R Davenport, and Greg Wilson. 2016. "A Quick Introduction to Version Control with Git and GitHub." *PLoS Computational Biology* 12 (1): e1004668.

Broman, Karl W, and Kara H Woo. 2018. "Data Organization in Spreadsheets." *The American Statistician* 72 (1): 2–10.

Byers-Heinlein, Krista, Christina Bergmann, Catherine Davies, Michael C Frank, J Kiley Hamlin, Melissa Kline, Jonathan F Kominsky, et al. 2020. "Building a Collaborative Psychological Science: Lessons Learned from ManyBabies 1." *Canadian Psychology/Psychologie Canadienne* 61 (4): 349.

De Montjoye, Yves-Alexandre, Laura Radaelli, Vivek Kumar Singh, et al. 2015. "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata." *Science* 347 (6221): 536–39.

Evangelou, Evangelos, Thomas A Trikalinos, and John PA Ioannidis. 2005. "Unavailability of Online Supplementary Scientific Information from Articles Published in Major Journals." *The FASEB Journal* 19 (14): 1943–44.

Gilmore, Rick O, and Karen E Adolph. 2017. "Video Can Make Behavioural Science More Reproducible." *Nature Human Behaviour*.

Hardwicke, Tom E, Manuel Bohn, Kyle MacDonald, Emily Hembacher, Michèle B. Nuijten, Benjamin N. Peloquin, Benjamin E. deMayo, Bria Long, Erica J. Yoon, and Michael C. Frank. 2021. "Analytic Reproducibility in Articles Receiving Open Data Badges at the Journal Psychological Science: An Observational Study." *Royal Society Open Science* 8 (1): 201494. https://doi.org/10.1098/rsos.201494.

Hardwicke, Tom E, and John P. A. Ioannidis. 2018. "Populating the Data Ark: An Attempt to Retrieve, Preserve, and Liberate Data from the Most Highly-Cited Psychology and Psychiatry Articles." *PLOS ONE* 13 (8): e0201856. https://doi.org/10.1371/journal.pone.0201856.

Hardwicke, Tom E, Maya B Mathur, Kyle Earl MacDonald, Gustav Nilsonne, George Christopher Banks, Mallory Kidwell, Alicia Hofelich Mohr, et al. 2018. "Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the Journal Cognition."

Houtkoop, Bobby Lee, Chris Chambers, Malcolm Macleod, Dorothy V. M. Bishop, Thomas E. Nichols, and Eric-Jan Wagenmakers. 2018. "Data Sharing in Psychology: A Survey on Barriers and Preconditions." *Advances in Methods and Practices in Psychological Science* 1 (1): 70–85. https://doi.org/10.1177/2515245917751886.

King, Gary, and Stuart Shieber. 2013. "Office Hours: Open Access." 2013. https://www.youtube.com/watch?v=jD6CcFxRelY/.

Klein, Olivier, Tom E Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsonne, Wolf Vanpaemel, and Michael C Frank. 2018. "A Practical Guide for Transparency in Psychological Science."

Meyer, Michelle N. 2018. "Practical Tips for Ethical Data Sharing." *Advances in Methods and Practices in Psychological Science* 1 (1): 131–44.

Munafò, Marcus R., Brian A Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1 (1): 1–9. https://doi.org/10.1038/s41562-016-0021.

Nosek, Brian A, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. "Promoting an Open Research Culture." *Science* 348 (6242): 1422–25. https://doi.org/10.1126/science.aab2374.

Petersen, Michael Bang. 2019. "Corrigendum: Healthy Out-Group Members Are Represented Psychologically as Infected in-Group Members." *Psychological Science* 30 (12): 1792–94. https://doi.org/https://doi.org/10.1177/0956797619887750.

Piwowar, Heather A, and Todd J Vision. 2013. "Data Reuse and the Open Data Citation Advantage." *PeerJ* 1: e175.

Ross, Michael W, Martin Y Iguchi, and Sangeeta Panicker. 2018. "Ethical Aspects of Data Sharing and Research Participant Protections." *American Psychologist* 73 (2): 138.

Rouder, Jeffrey N. 2015. "The What, Why, and How of Born-Open Data." *Behavior Research Methods* 48 (3): 1062–69. https://doi.org/10.3758/s13428-015-0630-z.

Simonsohn, Uri. 2013. "Just Post It: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone." *Psychological Science* 24 (10): 1875–88. https://doi.org/10.1177/0956797613480366.

Tenopir, Carol, Natalie M. Rice, Suzie Allard, Lynn Baird, Josh Borycz, Lisa Christian, Bruce Grant, Robert Olendorf, and Robert J. Sandusky. 2020. "Data Sharing, Management, Use, and Reuse: Practices and Perceptions of Scientists Worldwide." Edited by Sergi Lozano. *PLOS ONE* 15 (3): e0229003. https://doi.org/10.1371/journal.pone.0229003.

The ManyBabies Consortium, Michael C Frank, Katherine Jane Alcock, Natalia Arias-Trejo, Gisa Aschersleben, Dare Baldwin, Stéphanie Barbu, et al. 2020. "Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference." *Advances in Methods and Practices in Psychological Science*.

Voytek, Bradley. 2016. "The Virtuous Cycle of a Data Ecosystem." *PLOS Computational Biology* 12 (8): e1005037. https://doi.org/10.1371/journal.pcbi.1005037.

Wilkinson, Mark D, Michel Dumontier, I Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Sci Data* 3 (March): 160018.

Ziemann, Mark, Yotam Eren, and Assam El-Osta. 2016. "Gene Name Errors Are Widespread in the Scientific Literature." *Genome Biology* 17 (1): 1–3.

14 Writing

Learning goals

- Write clearly by being concise, using structure, and adjusting to your audience
- Write reproducibly by interleaving writing and analysis code
- Write responsibly by acknowledging limitations, correcting errors, and calibrating your conclusions

You've designed and run your experiment, and you have even analyzed your data. This final section of Experimentology discusses reporting your results. We begin by thinking through how to write clearly, reproducibility, and responsibly (this chapter); then we turn to the question of designing informative and pretty data visualizations (Chapter 15). Our final chapter in the section introduces meta-analysis as a tool for research synthesis, allowing us to contextualize research results. These chapters focus on themes of TRANSPARENCY as well as (especially for meta-analysis) BIAS REDUCTION and MEASUREMENT PRECISION.

All of the effort you put into designing and running an effective experiment may be wasted if you cannot clearly communicate what you did. Writing is a powerful tool – though you contribute to the conversation only once, it enables you to speak to a potentially infinite number of readers. So it's important to get it right! In this chapter, we'll provide some guidance on how to write scientific papers – the primary method for reporting on experiments – clearly, reproducibly, and responsibly.[1]

14.1 Writing clearly

What is the purpose of writing? "Telepathy, of course" says Stephen King (King 2000). The goal of writing is to transfer information from your mind to the reader's as effectively as possible. Unfortunately, for most of us, writing clearly does not come naturally; it is a craft we need to work at.

One of the most effective ways to learn to write clearly is to read and to imitate the writing you admire. Many scientific articles are not clearly written, so you will need to be selective in which models you imitate.

[1] Clarity of communication was a founding principle of modern science. Early proto-scientists conducting alchemical experiments often made their work deliberately obscure - even writing in cryptic codes - so that others could not discover the "powerful secrets of nature." Pioneers of scientific methodology, like Francis Bacon and Robert Boyle, pushed instead for transparency and clarity. Notoriously, Issac Newton (originally an alchemist and later a scientist), continued to write in a deliberately obscure fashion in order to "protect" his work (Heard 2016).

Fortunately, as a reader, you will know good writing when you see it – you will feel like the writer is sending ideas directly from their mind to yours. When you come across writing like that, try to find more work by the same author. The more good scientific writing you are exposed to, the more you will develop a sense of what works and what does not. You may pick up bad habits as well as good ones (we sure have!), but over time, your writing will improve if you make a conscious effort to weed out the bad, and keep the good.

There are no strict rules of clear writing, but there are some generally accepted conventions that we will share with you here, drawing from both general style guides and those specific to scientific writing (Zinsser 2006; Heard 2016; Gernsbacher 2018; Savage and Yeh 2019).

14.1.1 The structure of a scientific paper

A scientific paper is not a novel. Rather than reading from beginning to end, readers typically jump between sections to extract information efficiently (Doumont 2009). This "random access" is possible because research articles typically follow the same conventional structure (see Figure 14.1). The main body of the article includes four main sections: Introduction, Methods, Results, and Discussion (IMRaD).[2] This structure has a narrative logic: what's the knowledge gap? (introduction); how did you address it? (methods); what did you find? (results); what do the results mean? (discussion).

Structure helps writers as well as readers. Try starting the writing process with section headings as a structure, then flesh it out, layer by layer. In each section, start by making a list of the key points you want to convey, each representing the first sentence of a new paragraph. Then add the content of each paragraph and you'll be well on your way to having a full first draft of your article.

Imagine that the breadth of focus in the body of your article has an "hourglass" structure (Figure 14.1). The start of the introduction should have a broad focus, providing the reader with the general context of your study. From there, the focus of the introduction should get increasingly narrow until you are describing the specific knowledge gap or problem you will address and (briefly how you are going to address it. The methods and results sections are at the center of the hourglass because they are tightly focused on your study alone. In the discussion section, the focus shifts in the opposite direction, from narrow to broad. Begin by summarizing the results of your study, discuss limitations, then integrate the findings with existing literature and describe practical and theoretical implications.

[2] In the old old days, there were few conventions – scientists would share their latest findings by writing letters to each other. But as the number of scientists and studies increased, this approach became unsustainable. The IMRaD structure gained traction in the 1800s and became dominant in the mid-1900s as scientific productivity rapidly expanded in the post-war era. We think IMRaD style articles are a big improvement, even if it is nice to receive a letter every now and again.

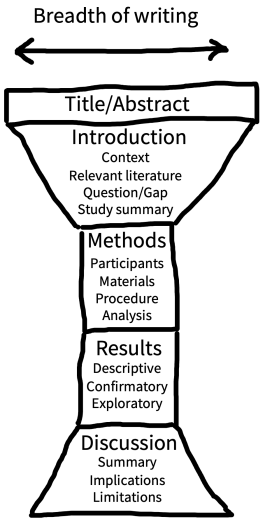


Figure 14.1: Conventional structure of a research article. The main body of the article consists of Introduction, Methods, Results, and Discussion (IMRaD) sections.

Research articles are often packed with complex information; it is easy for readers to get lost. A "cross reference" is a helpful signpost that tells readers where they can find relevant additional information without disrupting the flow of your writing. For example, you can refer the reader to data visualizations by cross referencing to figures or tables (e.g., "see Figure 1"), or additional methodological information in the supplementary information (e.g., "see Supplementary Information A").

One useful trick for structuring complex arguments is to cross reference your research aims/hypotheses with your results. For example, you could introduce numbered hypotheses in the introduction of an article and then refer to them directly when reporting the relevant analyses and results. These cross references can serve to remind readers how different results or analyses relate back to your research goals.

14.1.2 Paragraphs, sentences, and words

Writing an article is like drawing a human form. If you begin by sketching the clothes, you risk adding beautiful textures onto an impossible shape. Instead, you have to start by understanding the underlying skeleton and then gradually adding layers until you can visualize how cloth hangs on the body. The structure of an article is the "skeleton" and the paragraphs and sentences are the "flesh". Only start thinking about paragraphs and sentences once you have a solid outline in place.

Ideally, each paragraph should correspond to a single point in the article's outline, with the specifics necessary to convince the reader embedded within. "P-E-E-L" (Point - Explain - Evidence - Link) is a useful paragraph structure, particularly in the introduction and discussion sections. First, state the paragraph's message succinctly in the first sentence (P). The core of the paragraph is dedicated to further explaining the point and providing evidence (E-E; you can also include a third "E" – an example). At the end of the paragraph, take a couple of sentences to remind the reader of your point and set up a link to the next paragraph.

Since each sentence in a paragraph has a purpose, you can compose and edit the sentence by asking how its form serves that purpose. For example, short sentences are great for making strong initial points. On the other hand, if you only use short sentences your writing may come across as monotonous and robotic. Try varying sentence lengths to give your writing a more natural rhythm. Just avoid cramming too much information into the same sentence; very long sentences can be confusing and difficult to process.

You can also use sentence structure as a scaffold to support the reader's thinking. Start sentences with something the reader already knows. For example, rather than writing "We performed a between-subjects t-test comparing performance in the experimental and control groups to address the cognitive dissonance hypothesis", write "To address the cognitive dissonance hypothesis, we compared performance in the experimental group and control group using a between-subjects t-test."

Human readers are good at processing narratives about people. Yet often scientists compromise the research narrative by removing themselves from the process, sometimes even using awkward grammatical constructions to do so. For example, scientists sometimes write "the data were analysed" or, worse, "an analysis of the data was carried out." Many of us were taught to write sentences like these, but it's much clearer to say "we analyzed the data."

Similarly, many of us tend to hide our views with frames and caveats: "[It is believed that/Research indicates that/Studies show that] money leads to increased happiness (Frog & Toad, 1963)." If you truly do believe that money causes happiness, simply assert it – with a citation if necessary. Save caveats for cases where *someone* believes that money causes happiness, but it's *not* you. Emphasize uncertainty where you in fact feel that uncertainty is warranted and readers will take your doubts more seriously.

14.2 Advice

Scientific writing has a reputation for being dry, dull, and soulless. While it's true that writing research articles is more constrained than writing fiction, there are still ways to surprise and entertain your reader with metaphor, alliteration, and even humor. As long as your writing is clear and accurate, we see no reason why you cannot also make it enjoyable. Enjoyable articles are easier to read and more fun to write.[3]

Here are a few more pieces of advice about expressing yourself clearly.

Be explicit. Avoid vagueness and ambiguity. The more you leave the meaning of your writing to your reader's imagination the greater the danger that different readers will imagine different things! So be direct and specific.

Be concise. Maximize the signal to noise ratio in your writing by omitting needless words and removing clutter (Zinsser 2006). For example, say *we investigated* rather than *we performed an investigation of* and say *if*

[3] One of our favorite examples of an enjoyable article is Cutler (1994), a delightful piece that uses the form of the article to make a point about human language processing. Read it: you'll see!

rather than *in the event that*. Don't try to convey everything you know about a topic – a research report is not an essay. Include only what you need to achieve the purpose of the article and exclude everything else.

Be concrete. Concrete examples make abstract ideas easier to grasp. But some ideas are just hard to express in prose, and diagrams can be very helpful in these cases. For example, it may be clearer to illustrate a complex series of exclusion criteria using a flow chart rather than text. You can even use photos, videos, and screenshots to illustrate experimental tasks (Heycke and Spitzer 2019).

Be consistent. Referring to the same concept using different words can be confusing because it may not be clear if you are referring to a different concept or just using a synonym. For example, in everyday conversation, "replication" and "reproducibility" may sound like two different ways to refer to the same thing, but in scientific writing, these two concepts have different technical definitions, so we should not use them interchangeably. Define each technical term once and then use the same term throughout the manuscript.

Adjust to your audience. Most of us adjust our conversation style depending on who we're talking to; the same principle applies to good writing. Knowing your audience is more difficult with writing, because we cannot see the reader's reactions and adjust accordingly. Nevertheless, we can make some educated guesses about who our readers might be. For example, if you are writing an introductory review article, you may need to pay more attention to explaining technical termsn than if you are writing a research article for a specialty journal.

Check your understanding. Unclear writing can be a symptom of unclear thinking. If an idea doesn't make sense in your head, how will it ever make sense on the page? In fact, trying to communicate something in writing is an excellent way to probe your understanding and expose logical gaps in your arguments. So if you are finding it difficult to write clearly, stop and ask yourself *do I know what I want to say*? If the problem is unclear thinking, then it might be worth talking out the ideas with a colleague or advisor before you try to write them down.

Use acronyms sparingly. It's tempting to replace lengthy terminology with short acronyms — why say "cognitive dissonance theory" when you can say "CDT"? Unfortunately, acronyms can increase the reader's cognitive burden and cause misunderstandings.[4] For example, if you shorten "odds ratio" to "OR", the reader has to take the extra step of translating "OR" back to "odds ratio" every time they encounter it. The problem multiplies as you introduce more acronyms into your article.

[4] Barnett and Doubleday (2020) found that acronyms are widely used in research articles and argued that they undermine clear communication. Here is one example of text Barnett and Doubleday extracted from a 2019 publication to illustrate the point: "Applying PROBAST showed that ADO, B-AE-D, B-AE-D-C, extended ADO, updated ADO, updated BODE, and a model developed by Bertens et al. were derived in studies assessed as being at low risk of bias."

Worse, for some readers, "OR" tends to mean "operating room", not "odds ratio." Acronyms can be useful, but usually only when they are widely used and understood.

14.2.1 Drafting and revision

The clearest and most effortless-seeming scientific writing has probably gone through extensive revision to appear that way. It can surprise many students to know the amount of revision that has gone into many "breezy" articles. For example, Tversky and Kahneman repeatedly drafted and re-drafted each word of their famous (and highly readable) articles on judgment and decision-making, hunched over the typewriter together (Lewis 2016).

Think of the article you are writing as a garden. Your first draft may be an unruly mess of intertwined fronds and branches. Several rounds of pruning and sculpting will be needed before your writing reaches its most effective form. You'll be amazed how often you find words you can omit or elaborate sentences you can simplify.

It can be difficult to judge if your own writing has achieved its telepathic goal, especially after several rounds of revision. Try to get feedback from somebody in your target audience. Their comments – even if not wholly positive – will give you a good sense of how much of your argument they understood (and agreed with).[5]

[5] Seek out people who are willing to tell you that your writing is not good! They may not make you feel good, but they will help you improve.

14.3 Writing reproducibly

Many research results are not reproducible -— that is, the numbers and graphs that they report can't be recreated by repeating the original analyses – even on the original data. As we discussed in Chapter 3, a lack of reproducibility is a big problem for the scientific literature; if you can't trust the numbers in the articles you read, it's much harder to build on the literature.

Fortunately, there are number of tools and techniques available that you can use to write fully reproducible research reports. The basic idea is to create an unbroken chain that links every single part of the data analysis pipeline, from the raw data through to the final numbers reported in your research article. This linkage enables you – and hopefully others as well – to trace the provenance of every number and recreate (reproduce) it from scratch.

14.3.1 Why write reproducible reports?

There are (at least) three reasons to write reproducible reports. First, data analysis is an error-prone activity. Without safeguards in place, it can be easy to accidentally overwrite data, mislabel experimental conditions, or copy and paste the wrong statistics. As we discussed in Chapter 3, one study found that nearly half of a sample of psychology papers contained obvious statistical reporting errors (Nuijten et al. 2016). You can reduce opportunities for error by adopting a reproducible analysis workflow that avoids error-prone manual actions, like copying and pasting.

Second, technical information about data analysis can be difficult to communicate in writing. Prose is often ambiguous and authors can inadvertently leave out important details (Hardwicke et al. 2018). By contrast, a reproducible workflow documents the entire analysis pipeline from raw data to research report exactly as it was implemented, describing the origin of any reported values and allowing readers to assess, verify, and repeat the analysis process.

Finally, reproducible workflows are typically more efficient workflows. For example, you may realize you forgot to perform data exclusions and need to rerun the analysis. You may produce a graph and then decide you'd prefer a different color scheme. Or perhaps you want to output the same results table in a PDF document and in a PowerPoint slide. In a reproducible workflow, all of the analysis steps are scripted, and can be easily re-run at the click of a button. You (and others) can also reuse parts of your code in other projects, rather than having to write from scratch.

14.3.2 Principles of reproducible writing

Below we outline some general principles of reproducible writing. These can be put in practice in a number of different software ecosystems. We recommend RMarkdown and its successor, Quarto, which are ways of writing data analysis code in R so that it compiles into spiffy documents or even websites. (This book was written in Quarto). Appendix C gives an introduction to the nuts and bolts of using these tools to create scientific papers.

- **Never break the chain**. Every part of the analysis pipeline – from raw data[6] to final product – should be present in the project repository. By consulting the repository documentation, a reader should be able to follow the steps to go from the raw data to the final manuscript, including tables and figures.

[6] Modulo the privacy concerns discussed in Chapter 13, of course.

- **Script everything**. Try to ensure that each step of the analysis pipeline is executed by computer code rather than manual actions such as copying and pasting or directly editing spreadsheets. This practice ensures that every step is documented via executable code rather than ambiguous description, ensuring it can be reproduced. Imagine, for example, that you decided to re-code a variable in your dataset. You could use the "find and replace" function in Excel, but this action would not be documented – you might even forget that you did it! A better option would be to write an R script. While a scripted pipeline can be a pain to set up the first time, by the third time you rerun it, it will save you time.
- **Use literate programming**. The meaning of a chunk of computer code is not always obvious to another user, especially if they're not an expert. Indeed, we frequently look at our own code and scratch our heads, wondering what on earth it's doing. To avoid this problem, try to structure your code around plain language comments that explain what it should be doing, a technique known as "literate programming" (Knuth 1992).
- **Use defensive programming**. Errors can still occur in scripted analyses. Defensive programming is a series of strategies to help anticipate, detect, and avoid errors in advance. A typical defensive programming tool is the inclusion of **tests** in your code, snippets that check if the code or data meet some assumptions. For example, you might test if a variable storing reaction times has taken on values below zero (which should be impossible). If the test passes, the analysis pipeline continues; if the test fails, the pipeline halts and an error message appears to alert you to the problem.
- **Use free/open-source software and programming languages**. If possible, avoid using commercial software, like SPSS or Matlab, and instead use free, open-source software and programming languages, like JASP, Jamovi, R, or Python. This practice will make it easier for others to access, reuse, and verify your work – including yourself![7]
- **Use version control**. In Chapter 13, we introduced the benefits of version control – a great way to save your analysis pipeline incrementally as you build it, allowing you to roll back to a previous version if you accidentally introduce errors.
- **Preserve the computational environment**. Even if your analysis pipeline is entirely reproducible on your own computer, you still need to consider whether it will run on somebody else's computer,

[7] Several of us have libraries of old Matlab code. While discounted licenses are available to students, a full-price software license can be a major barrier to researchers with limited resources. if you move away from Matlab, it's also terrible to have to ask yourself whether it's worth the price of another year's license just to rerun one old analysis.

or even your own computer after software updates. You can address this issue by documenting and preserving the computational environment in which the analysis pipeline runs successfully. Various tools are available to help with this, including Docker, Code Ocean, renv (for R), and pip (for Python).[8]

[8] If you are interested in going in this direction, we recommend Peikert and Brandmaier (2021), which gives an advanced tutorial for complete computational reproducibility using Docker and `make` as tools to supplement git and R Markdown.

14.3.3 The reproducibility-collaboration trade-off

We would love to leave it there and watch you walk off into the sunset with a spring in your step and a reproducible report under your arm. Unfortunately, we have to admit that writing reproducibly can create a few practical difficulties when it comes to collaboration.

A major aspect of collaboration is exchanging comments and inline text edits with co-authors. You can do this exchange with R Markdown files and Git, but these tools are not as user-friendly as, say, Word or Google Docs, and some collaborators will be completely unfamiliar with them. Most journals also expect articles to be submitted as Word documents. Outputting R Markdown files to Word can often introduce formatting issues, especially for moderately complex tables. So until more user-friendly tools are introduced, some compromise between reproducibility and collaboration may be necessary. Here are two workflow styles for you to consider.

First, the **maximal reproducibility** approach. If your collaborators are familiar with R Markdown and you don't mind exchanging comments and edits via Git – or if they don't mind giving you lists of comments and changes that you implement in the R Markdown document – then you can maintain a fully reproducible workflow for your project. The journal submission and publication process may still introduce some issues such as incorporating changes made by the copy editor, but at least your submitted manuscript (and the preprint you have hopefully posted) will be fully reproducible.

Second, the **two worlds** approach. This workflow is a bit clunky, but it facilitates collaboration and maintains reproducibility. First, write your results section in R Markdown and generate a Word document (see Appendix C). Then, write the remainder of the manuscript in Word, including incorporating comments and changes from collaborators. When you have a final version, copy and paste the abstract, introduction, methods, and discussion into the R Markdown document.[9] Integrating any changes made to the results section back into the R Markdown requires a bit more effort, either using manual checking or Word's "compare documents" feature.[10] The advantage of this approach is that you have a reproducible document and your collaborators have

[9] You can also incorporate Google Docs into this workflow – we find that cloud platforms like Docs are especially useful when gathering comments from multiple collaborators on the same document. Unfortunately, you cannot generate a Google Doc from R Markdown, so you will need to import and convert or else copy and paste.

[10] New packages such as "trackdown" could help as well: https://claudiozandonella.github.io/trackdown/.

not had to deviate from their preferred workflow. Unfortunately, it requires more effort from you and is slightly more error-prone than the maximal reproducibility approach.

14.4 Writing responsibly

As a scientific writer, you have both professional and ethical responsibilities. You must communicate all relevant information about your research so as to enable proper evaluation and verification by other scientists. It is also important not to overstate your findings and calibrate your conclusions to the available evidence (Hoekstra and Vazire 2021). If errors are found in your work, you must respond and correct them when possible (Bishop 2018). Finally, you must meet scholarly obligations with regards to authorship and citation practices.

14.4.1 Responsible disclosure and interpretation

Back in school, we all learned that getting the right answer is not enough – you need to demonstrate how you arrived at that answer in order to get full marks. The same expectation applies to research reports. Don't just tell the reader what you found, tell them how you found it.[11] That means describing the methods in full detail, as well as sharing data, materials, and analysis scripts.

[11] It can be easy to overlook important details, especially when you reach the end of a project. Looking back at your study preregistration can be a helpful reminder. Reporting guidelines for different research designs can also provide useful checklists (Appelbaum et al. 2018).

In a journal article, you typically have some flexibility in terms of how much detail you provide in the main body of the article and how much you relegate to the supplementary information. Readers have different needs; some may just want to know the highlights, and some will need detailed methodological information in order to replicate your study. As a rule of thumb, try to make sure there is nothing relegated to the supplementary information that might surprise the reader. You certainty should not use the supplementary information to hide important details deliberately or use it as a disorganized dumping ground – the principles of clear writing still apply!

Here are a few more guidelines for responsible writing:

- **Don't overclaim**. Scientists often feel they are (and unfortunately, often are) evaluated based on the *results* of their research, rather than the *quality* of their research. Consequently, it can be tempting to make bigger and bolder claims than are really justified by the evidence. Think carefully about the limitations of your research and calibrate your conclusions to the evidence, rather than

what you wish you were able to claim. Ensure that your conclusions are appropriately stated throughout the manuscript, especially in the title and abstract.

- **Acknowledge limitations on interpretation and generalizability.** Even if you calibrate your claims appropriately throughout, there are likely specific limitations that are worth discussing, either as you introduce the design of the study in the introduction or as you interpret it in the discussion section. For example, if your experiment used one particular manipulation to instantiate a construct of interest, you might discuss this limitation and how it might be addressed by future work. Think carefully about the limitations of your study, state them clearly, and consider how they impact your conclusions (Clarke et al. 2023).[12] Discussions of limitations are a great point to make an explicit statement about the *generalizability* of your findings (see Simons, Shoda, and Lindsay 2017 for guidance about these kinds of "Constraints on Generality" statements).

- **Discuss, don't debate.** The purpose of the discussion section is to help the reader interpret your research. Importantly, a journal article is not a debate – don't feel the need to argue dogmatically for a particular position or interpretation. You should discuss the strengths and weaknesses of the evidence, and the relative merits of different interpretations. For example, perhaps there is a potential confounding variable that you were unable to eliminate with your research design. The reader might be able to spot this themselves, but regardless, its your responsibility to highlight it. Perhaps on balance you think the confound is unlikely to explain the results – that's fine, but you need to explain your reasoning to the reader.

- **Disclose conflicts of interest and funding.** Researchers are usually personally invested in the outcomes of their research and this investment can lead to bias (for example, overclaiming or selective reporting). But sometimes your potential personal gains from a piece of research rise above a threshold and are considered **conflicts of interest**. Where this threshold lies is not always completely clear. The most obvious conflicts of interest occur when you stand to benefit financially from the outcomes of your research (for example a drug developer evaluating their own drug). If you are in doubt about whether you have a potential conflict of interest, then you should disclose it. You should also disclose any funding you received for the research, partly because this is often a requirement of the funder, and partly because it may represent

[12] Should you just make your claims more modest, and avoid writing about your study's limitations? The balance between claims and limitations is tricky. One way to navigate this issue is to ask yourself, "is it OK to say X in the abstract of my article, if I later go on to say state a limitation relevant to that claim, or will the reader feel tricked?"

a conflict of interest, for example if the funder has a particular stake in the outcome of the research. To avoid ambiguity, you should also disclose when you do *not* have a conflict of interest or funding to declare.

- **Report transparently**. In Chapter 11, you learned about the problem of selective reporting and how this practice can bias the research literature. There are several ways to avoid this issue in your own work. First, assuming you *have* reported everything, include a statement in the methods section that explicitly says so. A statement suggested by Simmons, Nelson, and Simonsohn (2012) is "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study." If you have preregistered your study, clearly link to the preregistration and state whether you deviated from your original plan. You can include a detailed preregistration disclosure table in the supplementary information and highlight any major deviations in the methods section. In the results section, clearly identify (e.g., with sub-headings) which analyses were pre-planned and included in the preregistration (confirmatory) and which were not planned (exploratory).

14.4.2 Responsible handling of errors

It is not your responsibility to never make mistakes. But it *is* your responsibility to respond to errors in a timely, transparent, and professional manner (Bishop 2018).[13] Regardless of how the error was identified (e.g., by yourself or by a reader), we recommend contacting the journal and requesting that they publish a correction statement (sometimes called an **erratum**). Several of us have corrected papers in the past. If the error is serious and cannot be fixed, you should consider retracting the article.

[13] As jazz musician Miles Davis once said, "If you hit a wrong note, it's the next note that you play that determines if it's good or bad."

A correction/retraction statement should include the following information:

1. **Acknowledge the error**. Be clear that an error has occurred.
2. **Describe the error**. Readers need to know the exact nature of the error.
3. **Describe the implications of the error**. Readers need to know how the error might affect their interpretation of the results.
4. **Describe how the error occurred**. Knowing how the error happened may help others avoid the same error.
5. **Describe what you have done to address the error**. Others may learn from solutions you've implemented.

6. **Acknowledge the person who identified the error.** Identifying errors can take a lot of work; if the person is willing to be identified, give credit where credit is due.

Accident report

In 2018, at a crucial stage of her career, Julia Strand published an important study in the prestigious journal *Psychonomic Bulletin & Review*. She presented the work at conferences and received additional funding to do follow-up studies. But several months later, her team found that they could not replicate the result.

Puzzled, she began searching for the cause of the discrepant results. Eventually, she found the culprit – a programming error. As she sat staring at her computer in horror, she realized that it was unlikely anyone else would ever find the bug. Hiding the error must have seemed like the easiest thing to do.

But she did the right thing. She spent the next day informing her students, her co-authors, the funding officer, the department chair overseeing her tenure review, and the journal – to initiate a retraction of the article. And… it didn't ruin her career. Everybody was understanding and appreciated that she was doing the right thing. The journal corrected the article. She didn't lose her grant. She got tenure. And a lot of scientists, including us, admire her for what she did.

Honest mistakes happen – it's how you respond to them that matters (Strand 2021). In fact, survey research with both scientists and the general public suggests that scientists' reputations are built on the perception that they try to "get it right," not just to "be right" (Ebersole, Axt, and Nosek 2016).

14.4.3 Responsible citation

Citing prior work that your study builds upon ensures that researchers receive credit for their contributions and helps readers to verify the basis of your claims. You should certainly avoid copying the work of others and presenting it as your own (see Chapter 4 for more on plagiarism). Try to be explicit about why you are citing a source. For example, does it provide evidence to support your point? Is it a review paper that gives the reader useful background? Or is it a description of a theory you are testing?

Make sure you read articles before you cite them. Stang, Jonas, and Poole (2018) reports a cautionary tale in which a commentary criticizing a methodological tool was frequently cited as *supporting* the use of that tool! It seems that many authors had not read the paper they were citing, which is both misleading and embarrassing.

Try to avoid selective or uncritical citation. It is misleading to cite only research that supports your argument and ignoring research that doesn't. You should provide a balanced account of prior work, including contradictory evidence. Make sure to evaluate and integrate evidence from prior studies, rather than simply describing them. Remember – every study has limitations.

14.4.4 Responsible authorship practices

It is an ethical responsibility to credit the individuals who worked on a research project – both so that they can reap the benefits if the work is influential, but also so that they can take responsibility for errors.[14]

Currently in academia, the *authorship model* is dominant. Under this model, authorship and authorship order are important signals about researchers contributions to a project. It is generally expected that to qualify for authorship, an individual should have made a substantial contribution to the research (e.g., design, data collection, analysis), assisted with writing the research report, and takes joint responsibility for the research along with the other co-authors. Individuals who worked on the project who do not reach this threshold are instead mentioned in a separate acknowledgements section and not considered authors.

Authorship order is often understood to signal the nature and extent of an author's contribution. In psychology (and neighboring disciplines), the first author and last author are typically the project leaders. Typically – though certainly not always! – the first author is a junior colleague who implements the project and the last author is a senior colleague who supervises the project.

It has been argued that the authorship model should be replaced with a more inclusive *contributorship* model in which all individuals who worked on the project are acknowledged as 'contributors'. Unlike the authorship model, there is no arbitrary threshold for contributorship. The actual contributions of each individual are explicitly described, rather than relying on the implicit conventions of authorship order. The contributorship model may facilitate collaboration and ensure student assistants are properly credited.

You will probably find that most journals still expect you to use the authorship model. Nevertheless, it is usually possible – and sometimes required – to include a contributorship statement in your article that describes what everybody did. For example, the CREDIT taxonomy provides a structured taxonomy of research tasks, making for uniform contributorship reporting.[15]

Because authorship is such an important signal in academia, it's important to agree on an authorship plan with your collaborators (particularly who will be the first and last authors) as early as possible.[16]

[14] In 1975, physicist and mathematician Jack H. Hetherington wrote a paper he intended to submit to the journal *Physical Review Letters*. We're not sure why, but Hetherington wrote the paper in first person plural (i.e., referring to himself as "we" rather than "I"). He subsequently discovered that the journal would not accept the use of "we" for single-authored articles. Hetherington had painstakingly tapped out the article on his typewriter, an exercise he was not keen to repeat. Instead, he opted for a less taxing solution and named his cat – a feline by the name of F. D. C. Willard – as a coauthor. The paper was accepted and published (Hetherington and Willard 1975).

[15] For larger projects, the tool Tenzing allows for CREDIT statements to be generated automatically from standardized forms (Holcombe et al. 2020).

[16] If you have find yourself in a situation where all authors have contributed equally, you may have to draw inspiration from historical examples and determine authorship order based on a 25 game croquet series (Hassell and May 1974), rock, paper, scissors (Kupfer, Webbeking, and Franklin 2004), or a brownie bake-off (Young and Young 1992). Alternatively, you can adopt the method of Lakens, Scheel, and Isager (2018) and randomize the authorship order in R!

14.5 Chapter summary: Writing

Writing a scientific article can be a rewarding endpoint for the process of doing experimental research. But writing is a craft, and writing clearly – especially about complex and technical topics – can require substantial practice and many drafts. Further, writing about research comes with ethical and professional responsibilities that are different than the burdens of other kinds of writing. A scientific author must work to ensure the reproducibility of their findings and report on those findings responsibly, noting limitations and weaknesses as well as strengths.

Discussion questions

1. Find a writing buddy and exchange feedback on a short piece of writing (the abstract of a paper in progress, a conference abstract, or even a class project proposal would be good examples). Think about how to improve each other's writing using the advice offered in this chapter.
2. Identify a published research article with openly available data and see if you can reproduce an analysis in their paper by recovering the exact numerical values they report. You can find support for this exercise at the Social Science Reproduction Platform (https://www.socialsciencereproduction.org) or ReproHack (https://www.reprohack.org). Discuss with a friend what challenges you faced in this exercise and how they might be avoided in your own work.

Readings

- Zinsser, W. (2006). *On writing well: The classic guide to writing nonfiction [7th ed]*. Harper Collins.
- Gernsbacher, M. A. (2018). Writing empirical articles: Transparency, reproducibility, clarity, and memorability. *Advances in Methods and Practices in Psychological Science*, *1*, 403–14. https://doi.org/10.1177/2515245918754485.

References

Appelbaum, Mark, Harris Cooper, Rex B. Kline, Evan Mayo-Wilson, Arthur M. Nezu, and Stephen M. Rao. 2018. "Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report." *American Psychologist* 73 (1): 3. https://doi.org/10.1037/amp0000191.

Barnett, Adrian, and Zoe Doubleday. 2020. "The Growth of Acronyms in the Scientific Literature." *eLife* 9 (July): e60080. https://doi.org/10.7554/eLife.60080.

Bishop, D. V. M. 2018. "Fallibility in Science: Responding to Errors in the Work of Oneself and Others." *Advances in Methods and Practices in Psychological Science* 1 (3): 432–38. https://doi.org/10.1177/2515245918776632.

Clarke, Beth, Lindsay Alley, Sakshi Ghai, Jessica Kay Flake, Julia M. Rohrer, Joseph P. Simmons, Sarah R. Schiavone, and Simine Vazire. 2023. "Looking Our Limitations in the Eye: A Tutorial for Writing about Research Limitations in Psychology." PsyArXiv. https://doi.org/10.31234/osf.io/386bh.

Cutler, Anne. 1994. "The Perception of Rhythm in Language." *Cognition* 50: 79–81.

Doumont, Jean-Luc. 2009. *Trees, Maps, and Theorems*. Principiae.

Ebersole, Charles R, Jordan R Axt, and Brian A Nosek. 2016. "Scientists' Reputations Are Based on Getting It Right, Not Being Right." *PLoS Biology* 14 (5): e1002460.

Gernsbacher, Morton Ann. 2018. "Writing Empirical Articles: Transparency, Reproducibility, Clarity, and Memorability." *Advances in Methods and Practices in Psychological Science* 1 (3): 403–14. https://doi.org/10.1177/2515245918754485.

Hardwicke, Tom E, Maya B Mathur, Kyle Earl MacDonald, Gustav Nilsonne, George Christopher Banks, Mallory Kidwell, Alicia Hofelich Mohr, et al. 2018. "Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the Journal Cognition."

Hassell, M. P., and R. M. May. 1974. "Aggregation of Predators and Insect Parasites and Its Effect on Stability." *Journal of Animal Ecology* 43 (2): 567–94. https://doi.org/10.2307/3384.

Heard, Stephen B. 2016. *The Scientist's Guide to Writing: How to Write More Easily and Effectively Throughout Your Scientific Career*. Princeton, New Jersey: Princeton University Press.

Hetherington, J. H., and F. D. C. Willard. 1975. "Two-, Three-, and Four-Atom Exchange Effects in Bcc $^{3}\mathrm{he}$." *Physical Review Letters* 35 (21): 1442–44. https://doi.org/10.1103/PhysRevLett.35.1442.

Heycke, Tobias, and Lisa Spitzer. 2019. "Screen Recordings as a Tool to Document Computer Assisted Data Collection Procedures." *Psychologica Belgica* 59 (1): 269–80. https://doi.org/10.5334/pb.490.

Hoekstra, Rink, and Simine Vazire. 2021. "Aspiring to Greater Intellectual Humility in Science." *Nature Human Behaviour* 5 (12): 1602–7.

Holcombe, Alex O., Marton Kovacs, Frederik Aust, and Balazs Aczel. 2020. "Documenting Contributions to Scholarly Articles Using CRediT and Tenzing." Edited by Cassidy R. Sugimoto. *PLOS ONE* 15 (12): e0244611. https://doi.org/10.1371/journal.pone.0244611.

King, Stephen. 2000. *On Writing: A Memoir of the Craft*. Scribner.

Knuth, Donald Ervin. 1992. *Literate Programming*. no. 27. Center for the Study of Language; Information.

Kupfer, John A, Amy L Webbeking, and Scott B Franklin. 2004. "Forest Fragmentation Affects Early Successional Patterns on Shifting Cultivation Fields Near Indian Church, Belize." *Agriculture, Ecosystems & Environment* 103 (3): 509–18. https://doi.org/10.1016/j.agee.2003.11.011.

Lakens, Daniël, Anne M. Scheel, and Peder M. Isager. 2018. "Equivalence Testing for Psychological Research: A Tutorial." *Advances in Methods and Practices in Psychological Science* 1 (2): 259–69. https://doi.org/10.1177/2515245918770963.

Lewis, Michael. 2016. *The Undoing Project: A Friendship That Changed the World*. Penguin UK.

Nuijten, Michèle B, Chris H J Hartgerink, Marcel A L M van Assen, Sacha Epskamp, and Jelte M Wicherts. 2016. "The Prevalence of Statistical Reporting Errors in Psychology (1985–2013)." *Behavior Research Methods* 48 (4): 1205–26.

Peikert, Aaron, and Andreas M Brandmaier. 2021. "A Reproducible Data Analysis Workflow with r Markdown, Git, Make, and Docker." *Quantitative and Computational Methods in Behavioral Sciences*, 1–27.

Savage, Van, and Pamela Yeh. 2019. "Novelist Cormac McCarthy's Tips on How to Write a Great Science Paper." *Nature* 574 (7777): 441–43.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2012. "A 21 Word Solution." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2160588.

Simons, Daniel J, Yuichi Shoda, and D Stephen Lindsay. 2017. "Constraints on Generality (COG): A Proposed Addition to All Empirical Papers." *Perspectives on Psychological Science* 12 (6): 1123–28.

Stang, Andreas, Stephan Jonas, and Charles Poole. 2018. "Case Study in Major Quotation Errors: A Critical Commentary on the Newcastle–Ottawa Scale." *European Journal of Epidemiology* 33 (11): 1025–31. https://doi.org/10.1007/s10654-018-0443-3.

Strand, Julia. 2021. "Error Tight: Exercises for Lab Groups to Prevent Research Mistakes." PsyArXiv. https://doi.org/10.31234/osf.io/rsn5y.

Young, Helen J., and Truman P. Young. 1992. "Alternative Outcomes of Natural and Experimental High Pollen Loads." *Ecology* 73 (2): 639–47. https://doi.org/10.2307/1940770.

Zinsser, William. 2006. *On Writing Well: The Classic Guide to Writing Nonfiction*. 30th anniversary ed., 7th ed., rev. and updated. New York: HarperCollins.

15 Visualization

Learning goals

- Analyze the principles behind informative visualizations
- Incorporate visualization into an analysis workflow
- Learn to make "the design plot": a standard visualization of experimental data
- Select different visualizations of variability and distribution
- Connect visualization concepts to measurement principles

What makes visualizations so useful, and what role do they play in the experimenter's toolkit? Simply put, data visualization is the act of "making the invisible visible." Our visual systems are remarkably powerful pattern detectors, and relationships that aren't at all clear when scanning through rows of raw data can immediately jump out at us when presented in an appropriate graphical form (Zacks and Franconeri 2020). Good visualizations aim to deliberately harness this power and put it to work at every stage of the research process, from the quick sanity checks we run when first reading in our data to the publication-quality figures we design when we are ready to communicate our findings.

Yet our powerful pattern detectors can also be a liability; if we're not careful, we can easily be fooled into seeing patterns that are unreliable or even misleading. As psychology moves into an era of bigger data and more complex behaviors, we become increasingly reliant on **data visualization literacy** (Börner, Bueckle, and Ginda 2019) to make sense of what is going on. Further, as a researcher reporting about your data, creating appropriate visualizations that are aligned with your analyses (as well as your design and preregistration) is an important part of TRANSPARENCY and BIAS REDUCTION in your reporting.

Case study

Mapping a pandemic

In 1854, a deadly outbreak of cholera was sweeping through London. The scientific consensus at the time was that diseases like cholera spread through breathing poisonous and foul-smelling vapors, an idea known as the "miasma theory" (Halliday 2001). An obstetrician and anesthesiologist named John Snow, however, had proposed an alter-

native theory: rather than spreading through foul air, he thought that cholera was spreading through a polluted water supply (Snow 1855). To make a public case for this idea, he started counting cholera deaths. He marked each case on a map of the area, and indicated the locations of the water pumps for reference. Furthermore, a line could be drawn representing the region that was closest to each water pump, a technique which is now known as a Voronoi diagram (https://en.wikipedia.org/wiki/Voronoi_diagram). The resulting illustration clearly reveals that cases clustered around an area called Golden Square, which received water from a pump on Broad Street (Figure 15.1). Although the precise causal role of these maps in Snow's own thinking is disputed, and it is likely that he produced them well after the incident (Brody et al. 2000), they nonetheless played a significant role in the history of data visualization (Friendly and Wainer 2021).

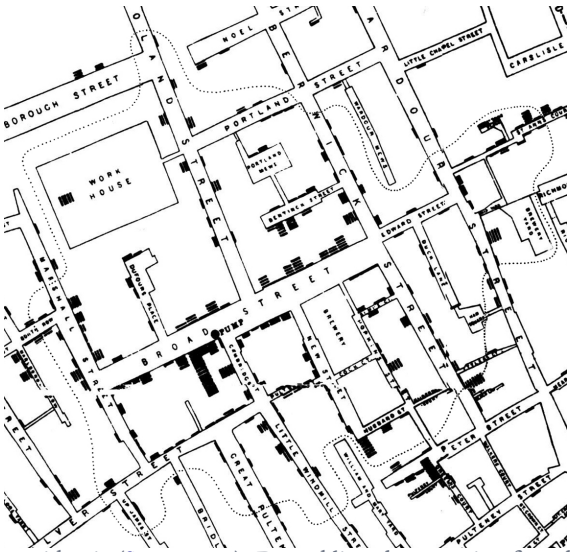


Figure 15.1: Mapping out a cholera epidemic (Snow 1854). Dotted line shows region for which Broad Street pump is nearest.

Nearly two centuries later, as the COVID-19 pandemic swept through the world, governmental agencies like the CDC produced maps of the outbreak that became much more familiar (Figure 15.2).

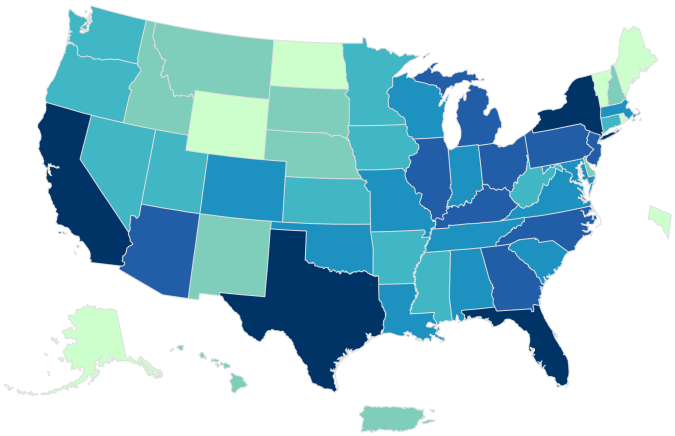


Figure 15.2: Map showing the counts of COVID hospitalizations by state since August 2020 as of January 2024 (from the CDC COVID Data Tracker, https://covid.cdc.gov/covid-data-tracker). *Usage does not constitute endorsement or recommendation by the U.S. Government, Department of Health and Human Services, or Centers for Disease Control and Prevention.*

These maps make abstract statistics visible: By assigning higher cumulative case rates to darker colors, we can see at a glance which areas have been most affected. And we're not limited by the spatial layout of a map. We're now also used to seeing the horizontal axis correspond to *time* and the vertical axis correspond to some value at that time. Curves like the following, showing the weekly counts of new cases, allow us to see other patterns, like the *rate of change*. Even though more and more cases accumulate every day, we can see at a glance the different "waves" of cases, and when they peaked (Figure 15.3).

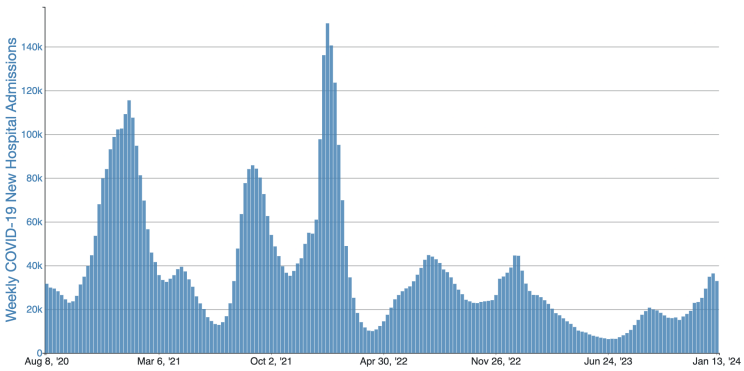


Figure 15.3: Weekly counts of new reported COVID hospital admissions in the US between August 2020 and January 2024 (from the CDC COVID Data Tracker, https://covid.cdc.gov/covid-data-tracker). *Usage does not constitute endorsement or recommendation by the U.S. Government, Department of Health and Human Services, or Centers for Disease Control and Prevention.*

While these visualizations capture purely descriptive statistics, we often want our visualizations to answer more specific questions. For example, we may ask about the effectiveness of vaccinations: how do case rates differ across vaccinated and unvaccinated populations? In this case, we may talk about "breaking out" a curve by some other variable, like vaccination status (Figure 15.4).

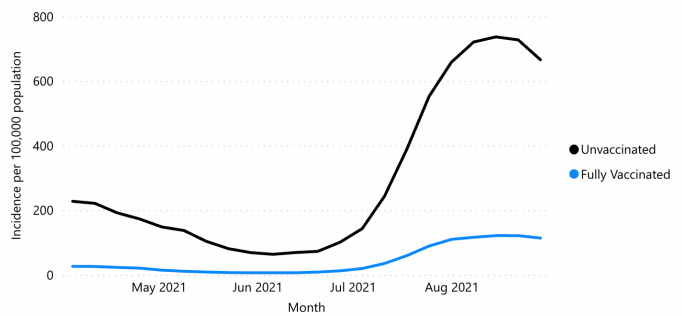


Figure 15.4: Rates of COVID cases by vaccination status (from the CDC COVID Data Tracker, https://covid.cdc.gov/covid-data-tracker). *Usage does not constitute endorsement or recommendation by the U.S. Government, Department of Health and Human Services, or Centers for Disease Control and Prevention.*

From this visualization, we can see that unvaccinated individuals are about 6x more likely to test positive. At the same time, these visualizations were produced using *observational* data, which makes it challenging to draw causal inferences. For example, people were not randomly assigned to vaccination conditions, and those who have avoided vaccinations may differ in other ways than those who sought out vaccinations. Additionally, you may have noticed that these visualizations typically do not give a sense of the raw data, the sample sizes of each group, or uncertainty about the estimates. In this chapter, we will explore how to use visualizations to communicate the results of carefully controlled psychology experiments, which license stronger causal inferences.

15.1 Basic principles of (confirmatory) visualization

In this section, we begin by introducing a few simple guidelines to keep in mind when making informative visualizations in the context of experimental psychology.[1] Remember that our needs may be distinct from other fields, such as journalism or public policy. You may have seen beautiful and engaging full-page graphics with small print and a wealth of information. The art of designing and producing these graphics is typically known as **infoviz** and should be distinguished from what we call **statistical visualization** (Gelman and Unwin 2013).

[1] For the purposes of understanding the examples in this chapter, it should be sufficient to work through the tutorials on data manipulation and visualization in Appendix D and Appendix E.

Roughly, infoviz aims to construct rich and immersive worlds to visually explore: a reader can spend hours pouring over the most intricate graphics and continue to find new and intriguing patterns. **Statistical visualization**, on the other hand, aims to crisply convey the logic of a specific inference at a glance. These visualizations are the production-ready figures that anchor the results section of a paper and accompany the key, pre-registered analyses of interest. In this section, we review several basic principles of making statistical visualizations. We then return below to the role of visualization in more exploratory analyses.

15.1.1 Principle 1: Show the design

There are so many different kinds of graphs (bar graphs, line graphs, scatter plots, and pie charts) and so many different possible attributes of those graphs (colors, sizes, line types). How do we begin to decide how to navigate these decisions? The first principle guiding good statistical visualizations is to *show the design* of your experiment.

The first confirmatory plot you should have in mind for your experiment is the **design plot**. Analogous to the "default" or "saturated" model in Chapter 7, the design plot should show the key dependent variable of the experiment, broken down by all of the key manipulations. Critically, design plots should neither omit particular manipulations because they didn't yield an effect or include extra covariates because they seemed interesting after looking at the data. Both of these steps are the visual analogue of p-hacking! Instead, the design plot is the "preregistered analysis" of your visualization: it illustrates a first look at the estimated causal effects from your experimental manipulations. In the words of Coppock (2019), "visualize as you randomize"!

It can sometimes be a challenge to represent the full pattern of manipulations from an experiment in a single plot. Below we give some tricks for maximizing the legible information in your plot. But if you have tried these and your design plot still looks crowded and messy, that could be

an indication that your experiment is manipulating too many things at once!

There are strong (unwritten) conventions about how your confirmatory analysis is expected to map onto graphical elements, and following these conventions can minimize confusion. Start with the variables you manipulate, and make sure they are clearly visible. Conventionally, the primary manipulation of interest (e.g. condition) goes on the x-axis, and the primary measurement of interest (e.g. responses) goes on the y-axis. Other critical variables of interest (e.g. secondary manipulations, demographics) are then assigned to "visual variables" (e.g. color, shape, or size).

CODE

The visualization library `ggplot` (see Appendix E) makes the mapping of variables in the data to visual data. Part of a ggplot call is an `aes` (short for aesthetics) mapping:

```
aes(
  x = ...,
  y = ...,
  color = ...,
  linetype = ...,
)
```

The aesthetics argument serves as a statement of how data are mapped to "marks" on the plot. This transparent mapping makes it very easy to explore different plot types by changing that `aes()` statement, as we'll see below.

As an example, we will consider the data from Stiller, Goodman, and Frank (2015) that we explored back in Chapter 7. Because this experiment was a developmental study, the primary independent variable of interest was the age group of participants (ages 2, 3, or 4). So age gets assigned to the horizontal (x) axis. The dependent variable is accuracy: the proportion of trials that a participant made the correct response (out of 4 trials). So accuracy goes on the vertical (y) axis. Now, we have two other variables that we might want to show: the condition (experimental vs. control) and the type of stimuli (houses, beds, and plates of pasta). When we think about it, though, only condition is central to exposing the design. While we might be interested in whether some types of stimuli are systematically easier or harder than others, condition is more central for understanding the *logic* of the study.

CODE

As a reminder, here's our code for loading the Stiller, Goodman, and Frank (2015) data:

```
repo <- "https://raw.githubusercontent.com/langcog/experimentology/main/"
sgf <- read_csv(file.path(repo, "data/tidyverse/stiller_scales_data.csv")) |>
  mutate(age_group = cut(age, 2:5, include.lowest = TRUE),
         condition = condition |>
           fct_recode("Experimental" = "Label", "Control" = "No Label"))

sgf_cond_means <- sgf |>
  group_by(condition, age_group) |>
  summarise(rating = mean(correct))
```

15.1.2 Principle 2: Facilitate comparison

Now that you've mapped elements of your design to the figure's axes, how do you decide which graphical elements to display? You might think: well, in principle, these assignments are all arbitrary anyway. As long as we clearly label our choices, it shouldn't matter whether we use lines, points, bars, colors, textures, or shapes. It's true that there are many ways to show the same data. But being thoughtful about our choices can make it much easier for readers to interpret our findings. The second principle of statistical visualizations is to *facilitate comparison* along the dimensions relevant to our scientific questions. It is easier for our visual system to accurately compare the location of elements (e.g. noticing that one point is a certain distance away from another) than to compare their areas or colors (e.g. noticing that one point is bigger or brighter than another). Figure 15.5 shows an ordering of visual variables based on how accurate our visual system is in making comparisons.

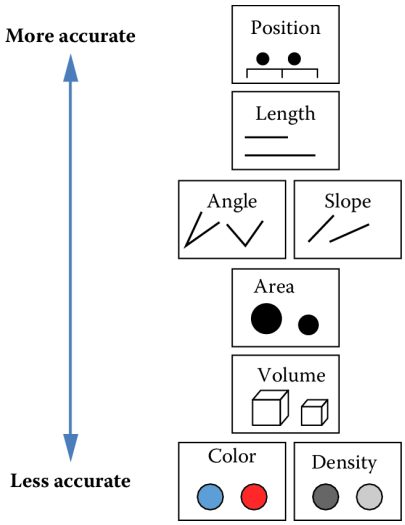


Figure 15.5: Principles of visual perception can help guide visualization choices. Based on Mackinlay (1986; see also Cleveland and McGill 1984).

For example, we *could* start by plotting the accuracy of each age group as colors (Figure 15.6).

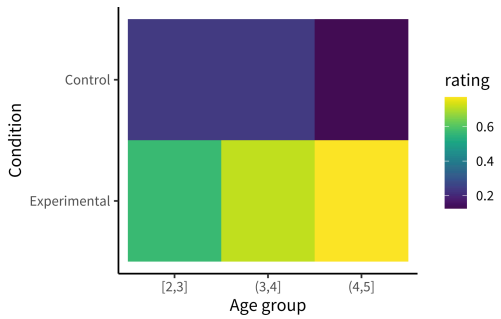


Figure 15.6: A first visualization of the Stiller, Goodman, and Frank (2015) data.

CODE

To make this (bad) visualization, we used a ggplot function called `geom_tile()`.

```
ggplot(sgf_cond_means, aes(x = age_group, y = condition, fill = rating)) +
  geom_tile() +
  labs(x = "Age group", y = "Condition")
```

`geom_tile()` is commonly used to make heat maps (https://en.wikipedia.org/wiki/Heat_map): for each value of some pair of variables (x, y), it shows a color representing the magnitude of a third variable (z). While a heat map is a silly way to visualize the Stiller, Goodman, and Frank (2015) data, consider using `geom_tile()` when you have a pair of continuous variables, each taking a large range of values. In these cases, bar plots and line plots tend to get extremely cluttered, making it hard to see the relationship between the variables. Heat maps help these relationships to pop out as clear "hot" and "cold" regions. For example, in Barnett, Griffiths, and Hawkins (2022), a heatmap was used to show a specific range of parameters where an effect of interest emerged (see Figure 15.7).

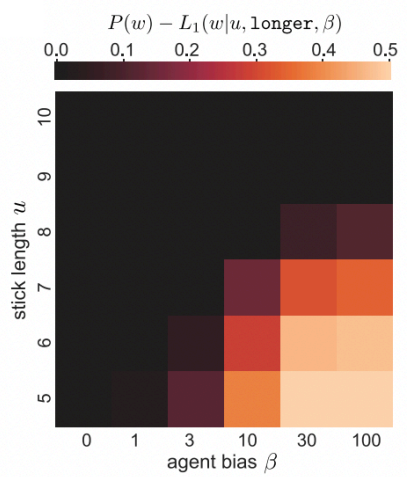


Figure 15.7: Heatmap showing a specific range of continuous parameters where an effect emerged. Barnett, Griffiths, and Hawkins (2022), Figure 3 (licensed under CC BY 4.0).

Or we could plot the accuracy of each age group as sizes/areas (Figure 15.8).

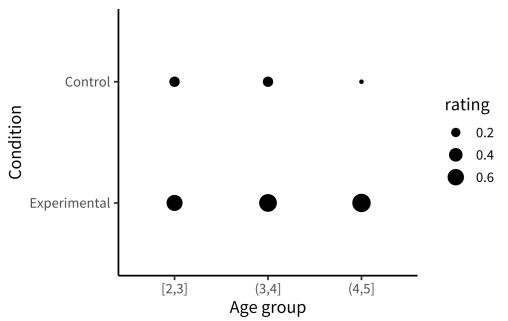


Figure 15.8: Iterating on the Stiller data using size.

CODE

To make this (bad) visualization, we mapped the rating DV to the `size` element in our `aes()` call.

```
ggplot(sgf_cond_means, aes(x = age_group, y = condition, size = rating)) +
  geom_point() +
  labs(x = "Age group", y = "Condition")
```

These plots allow us to see that one condition is (qualitatively) bigger than others, but it's hard to tell how much bigger. Additionally, this way of plotting the data places equal emphasis on age and condition, but we may instead have in mind particular contrasts, like the *change* across ages and how that change differs across conditions. An alternative is to show six bars: three on the left showing the 'experimental' phase and three on the right showing the 'control' phase. Maybe the age groups then are represented as different colors, as in Figure 15.9.

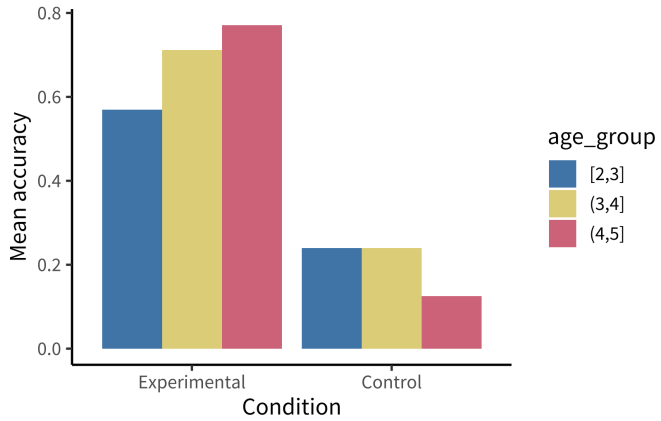


Figure 15.9: A bar graph of the Stiller data.

CODE

We make bar plots using the `ggplot` function `geom_col()`. By default, it creates "stacked" bar plots, where all values associated with the same x value (here, `condition`) get stacked up on top of one another. Stacked bar plots can be useful if, for example, you're plotting proportions that sum up to 1, or want to show how some big count is broken down into subcategories. It's also common to use `geom_area()` for this purpose, which connects adjacent regions. But the more common bar plot used in psychology puts the bars next to one another, or "dodges" them. To accomplish this, we use the `position = "dodge"` argument:

```
ggplot(sgf_cond_means, aes(x = condition, y = rating, fill = age_group)) +
  geom_col(position = "dodge") +
  labs(x = "Condition", y = "Mean accuracy")
```

This plot is slightly better: it's easier to compare the heights of bars than the 'blueness' of squares, and mapping age to color draws our eye to those contrasts. However, we can do even better by noticing that our experiment was designed to test an *interaction*. That statistic of interest is a difference of differences. To what extent does the developmental change in performance on the experimental condition different from developmental change in performance on the control condition? Some researchers have gotten proficient at reading off interactions from bar plots, but they also require a complex set of eye movements. We have to look at the pattern across the bars on the left, and then jump over to the bars on the right, and implicitly judge one difference against the other: the actual statistic isn't explicitly shown anywhere! What could help facilitate this comparison? Consider the line plot in Figure 15.10.

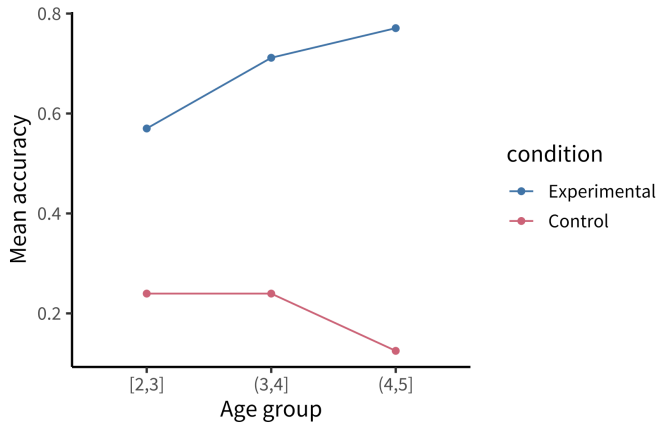


Figure 15.10: A line graph of the Stiller data promotes comparison.

Code

Using a combination of `geom_point()` and `geom_line()`:

```
ggplot(sgf_cond_means, aes(x = age_group, y = rating, color = condition)) +
  geom_point() +
  geom_line(aes(group = condition)) +
  labs(x = "Age group", y = "Mean accuracy")
```

The interaction contrast we want to interpret is highlighted visually in this plot. It is much easier to compare the slopes of two lines than mentally compute a difference of differences between bars. Here are a few corollaries of this principle (adapted from a presentation by Karl Broman[2]).

- It is easier to compare values that are *adjacent* to one another. This is especially important when there are many different conditions

[2] https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf

included on the same plot. If particular sets of conditions are of theoretical interest, place them close to one another. Otherwise, sort conditions by a meaningful value (rather than alphabetically, which is usually the default for plotting software).

- When possible, color-code labels and place them directly next to data rather than in a separate legend. Legends force readers to glance back and forth to remember what different colors or lines mean.
- When making histograms or density plots, it is challenging to compare distributions when they are placed side-by-side. Instead, facilitate comparison of distributions by vertically aligning them, or making them transparent and placed on the same axes.
- If the scale makes it hard to see important differences, consider transforming the data (e.g. taking the logarithm).
- When making bar plots, be very careful about the vertical y-axis. A classic "misleading visualization" mistake is to cut off the bottom of the bars by placing the endpoint of the y-axis at some arbitrary value near the smallest data point. This is misleading because people interpret bar plots in terms of the relative *area* of the bars (i.e. the amount of ink taken up by the bar), not just their absolute y-values.
- If a key variable from your design is mapped to color, choose the color scale carefully. For example, if the variable is binary or categorical, choose visually distinct colors to maximize contrast (e.g. black, blue, and orange). If the variable is ordinal or continuous, use a color gradient. If there is a natural midpoint (e.g. if some values are negative and some are positive), consider using a diverging scale (e.g. different colors at each extreme). Remember also that a portion of your audience may be color-blind.[3]

[3] Palettes like `viridis` have been designed to be colorblind-friendly and also perceptually uniform (i.e. the perceived difference between 0.1 and 0.2 is approximately the same as the difference between 0.8 and 0.9).

15.1.3 Principle 3: Show the data

Looking at older papers, you may be alarmed to notice how little information is contained in the graphs. The worst offenders might show just two bars, representing average values for two conditions. This kind of plot adds very little beyond a sentence in the text reporting the means, but it can also be seriously misleading. It hides real variation in the data, making a noisy effect based on a few data points look the same as a more systematic one based on a larger sample. Additionally, it collapses the *distribution* of the data, making a multi-modal distribution look the same as

a unimodal one. The third principle of modern statistical visualization is to *show the data* and visualize variability in some form.

The most minimal form of this principle is to *always include error bars*.[4] Error bars turn a purely descriptive visualization into an inferential one. They represent a minimal form of uncertainty about the possible statistics that might have been observed, not just the one that was actually observed. Figure 15.11 shows the Stiller data with (bootstrapped) error bars.

[4] And be sure to tell the reader what the error bars represent – a 95% confidence interval? A standard error of the mean? – without this information, error bars are hard to interpret (see Depth box below).

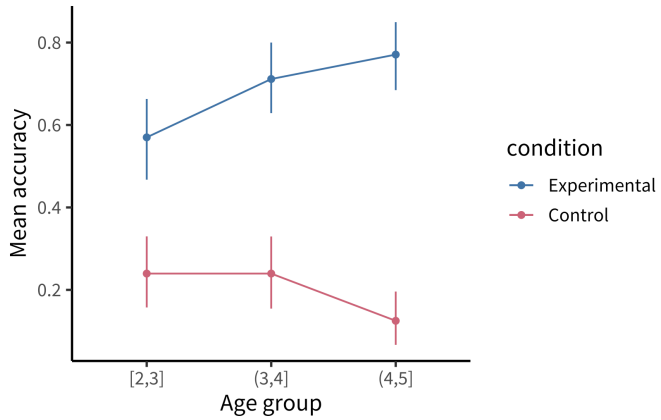


Figure 15.11: Error bars (95% CIs) added to the Stiller data line graph.

CODE

A common problem arises when we want to add error bars to a dodged bar plot. Naively, we'd expect we could just dodge the error bars in the same way we dodged the bars themselves:

```
geom_col(position = "dodge") +
geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper), position = "dodge")
```

But this doesn't work! The rationale is kind of technical, but the width of the error bars is much narrower than the width of the bars, so we need to manually specify how much to dodge the error bars with the `position_dodge()` function:

```
geom_col(position = position_dodge()) +
geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper),
              position = position_dodge(width = 0.9))
```

This does the trick!

But we can do even better. By overlaying the distribution of the actual data points on the same plot, we can give the reader information not just about the statistical inferences but the underlying data supporting those inferences. In the case of the Stiller, Goodman, and Frank (2015) study,

data points for individual trials are binary (correct or incorrect). It's technically possible to show individual responses as dots at 0s and 1s, but this doesn't tell us much (we'll just get a big clump of 0s and a big clump of 1s). The question to ask yourself when 'showing the data' is: what are the theoretically meaningful *units* of variation in the data? This question is closely related to our discussion of mixed-effects models in Chapter 7, when we considered which random effects we should include. Here, a reader is likely to wonder how much variance was found across *different children* in a given age group. To show such variation, we aggregate to calculate an accuracy score for each participant.[5]

There are many ways of showing the resulting distribution of participant-level data. For example, a **boxplot** shows the median (a horizontal line) in the center of a box extending from the lower quartile (25%) to the upper quartile (75%). Lines then extend out to the biggest and smallest values (excluding outliers, which are shown as dots). Figure 15.12 gives the boxplots for the Stiller data, which don't look that informative – perhaps because of the coarseness of individual participant averages due to the small number of trials.

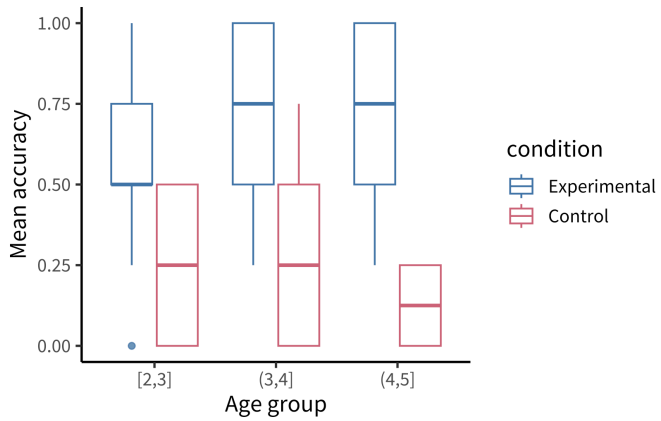


Figure 15.12: Boxplot of the Stiller data.

[5] While participant-level variation is a good default, the relevant level of aggregation may differ across designs. For example, collective behavior studies may choose to show the data point for each *group*. This choice of unit is also important when generating error bars: if you have a small number of participants but many observations per participant, you are faced with a choice. You may either bootstrap over the flat list of all individual observations (yielding very small error bars), or you may first aggregate within participants (yielding larger error bars that account for the fact that repeated observations from the same participant are not independent).

CODE

In `ggplot`, we can make box plots using `geom_boxplot()`:

```
geom_boxplot(alpha = 0.8)
```

A common problem to run into is that `geom_boxplot()` requires the variable assigned to `x` to discrete. If you have discrete levels of a numeric variable (e.g. age groups), make sure you've actually converted that variable to a `factor`. Otherwise, if it's still coded as `numeric`, `ggplot` will collapse all of the levels together!

It is also common to show the raw data as jittered values with low trans-

parency. In Figure 15.13, we jitter the points because many participants have the same numbers (e.g. 50% and if they overlap it is hard to see how many points there are.

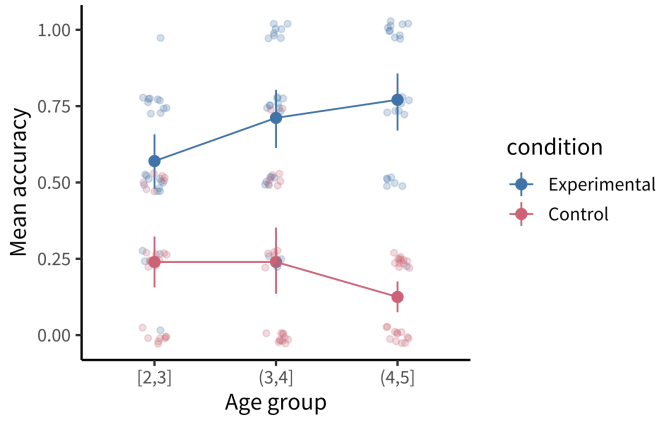


Figure 15.13: Jittered points representing the data distribution of the Stiller data.

Code

Adding the jittered points is simple using `geom_jitter()`, but we are starting to have a fairly complex plot so maybe it's worth taking stock of how we get there.

To plot both *condition* means and *participant* means, we need to create two different data frames. Here `sgf_subj_means` is a data frame of means for each participant; `sgf_subj_ci` is a data frame with means and confidence intervals *across* participants. For this purpose, we use the `tidyboot` package and the `tidyboot_mean()` function, which gives us bootstrapped 95% confidence intervals for the means.

```
sgf_subj_means <- sgf |>
  group_by(condition, age_group, subid) |>
  summarize(rating = mean(correct))

sgf_subj_ci <- sgf_subj_means |>
  group_by(condition, age_group) |>
  tidyboot::tidyboot_mean(rating) |>
  rename(rating = empirical_stat)

ggplot(sgf_subj_ci, aes(x = age_group, y = rating, color = condition)) +
  geom_pointrange(aes(ymin = ci_lower, ymax = ci_upper)) +
  geom_line(aes(group = condition)) +
  geom_jitter(data = sgf_subj_means, alpha = 0.25, width = 0.1, height = .03) +
  labs(x = "Age group", y = "Mean accuracy")
```

The most noteworthy aspect of this code is that the `geom_jitter()` function doesn't just take a different aesthetic, it also takes a different dataframe altogether! Mixing dataframes can be an important tool for creating complex plots.

Perhaps the format that takes this principle the furthest is the so-called

"raincloud plot" (Allen et al. 2019) shown in Figure 15.14. A raincloud plot combines the raw data ("rain") with a smoothed density ("cloud") and a boxplot giving the median and quartiles of the distribution.

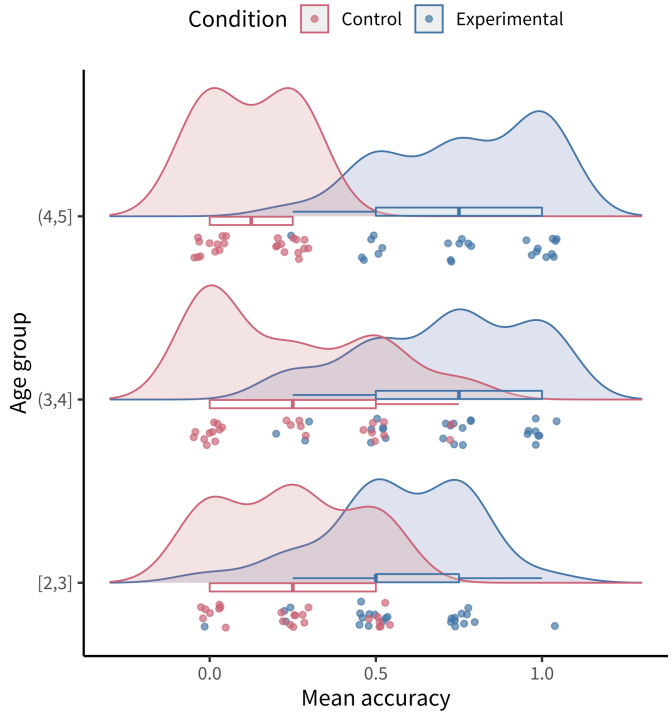


Figure 15.14: Raincloud plot of the Stiller data.

CODE

This raincloud plot requires two additional plotting packages: `ggridges` for the densities and `ggstance` for the horizontal boxplots.

```
library(ggridges)
library(ggstance)
ggplot(sgf_subj_means, aes(y = age_group, x = rating, color = condition)) +
  geom_density_ridges(aes(fill = condition), alpha = 0.2, scale = 0.7,
                      jittered_points = TRUE, point_alpha = 0.7,
                      position = position_raincloud(width = 0.05, height = 0.15,
                                                    ygap = 0.1)) +
  geom_boxploth(width = 0.1, alpha = 0.2, outlier.shape = NA, show.legend = FALSE) +
  scale_y_discrete(expand = expansion(mult = c(0.2, 0.4))) +
  guides(fill = "none", color = guide_legend(reverse = TRUE)) +
  labs(x = "Mean accuracy", y = "Age group", color = "Condition") +
  theme(legend.position = "top")
```

DEPTH

Visualizing uncertainty with error bars

One common misconception is that error bars are a measure of variance *in the data*, like the standard deviation of the response variable. Instead, they typically represent a measure of precision extracted from the statistical model. In older papers, for example, it was common to use the standard error of the mean (SEM; see Chapter 6). Remember that this is not the standard deviation of the data distribution but of the *sampling distribution* of the mean that is being estimated. Given the central limit theorem, which tells us that this sampling distribution is asymptotically normal, it was straightforward to estimate the standard error analytically using the empirical standard deviation of the data divided by the square root of the sample size: `sd(x) / sqrt(length(x))`. Error bars based on the SEM often looked misleadingly small, as they only represent a 68% interval of the sampling distribution and go to zero quickly as a function of sample size. As a result, it became more common to show the 95% confidence interval instead: [-1.96 × SEM, 1.96 × SEM].

While these analytic equations remain common, an increasingly popular alternative is to *bootstrap* confidence intervals (see Depth box in Chapter 6 for more on bootstrapping). The bootstrap is a powerfully generic technique, especially when you want to show error bars for summary statistics that are more complex than means, where we do not have such convenient asymptotic guarantees and "closed-form" equations. For example, if you're working with a skewed response variable or a dataset with clear outliers, and you want to estimate medians and quartiles.

Or suppose you want to estimate proportions from categorical data, or a more *ad hoc* statistic like the AUC (area underneath the curve) in a hierarchical design where it is not clear how to aggregate across items or participants in a mixed-effects model. Analytic estimators of confidence intervals can in principle be derived for these statistics, subject to different assumptions, but it is often more transparent and reliable in practice to use the bootstrap. As long as you can write a code snippet to compute a value from a dataset, you can use the bootstrap.

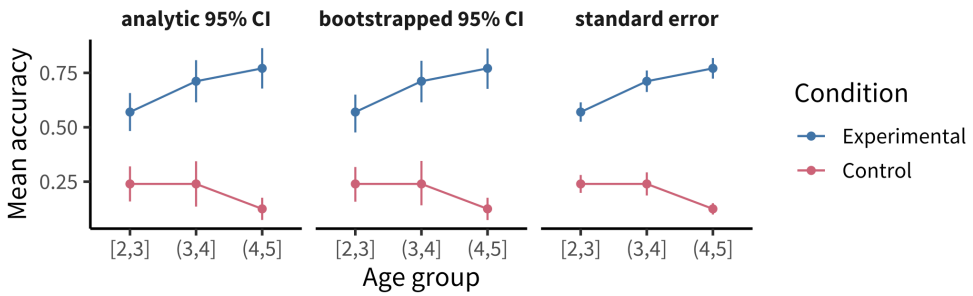


Figure 15.15: Three different error bars for the Stiller data: bootstrapped 95% confidence intervals (left), standard error of the mean (middle), and analytically computed confidence intervals (right).

As we can see, the bootstrapped 95% CI looks similar to the analytic 95% CI derived from the standard error, except the upper and lower limits are slightly asymmetric (reflecting outliers in one direction or another). Of course, the bootstrap is not a silver bullet and can be abused in particularly small samples. This is because the bootstrap is fundamentally limited to the sample we run it on. It can be expected to be reasonably accurate if the sample is reasonably representative of the population. But at the end of the day, as they say, "there's no such thing as a free lunch." In other words, we cannot magically pull more information out of a small sample without making additional assumptions about the data generating process.

15.1.1 Principle 4: Maximize information, minimize ink

Now that we have the basic graphical elements in place to show our design and data, it might seem like the rest is purely a matter of aesthetic preference, like choosing a pretty color scheme or font. Not so.

There are well-founded principles to make the difference between an effective visualization and a confusing or obfuscating one. Simply put, we should try to use the simplest possible presentation of the maximal amount of information: we should maximize the "data-ink ratio". To calculate the amount of information shown, Tufte (1983) suggested a measure called the "data density index," the "numbers plotted per square inch". The worst offenders have a very low density while also using a lot of excess ink (e.g., Figure 15.16 and Figure 15.17)

The defaults in modern visualization libraries like `ggplot` prevent some of the worst offenses, but are still often suboptimal. For example: consider whether the visual complexity introduced by the default grey background and grid lines in Figure 15.18) is justified, or whether a more minimal theme would be sufficient (see the `ggthemes`[6] package for a good collection of themes).

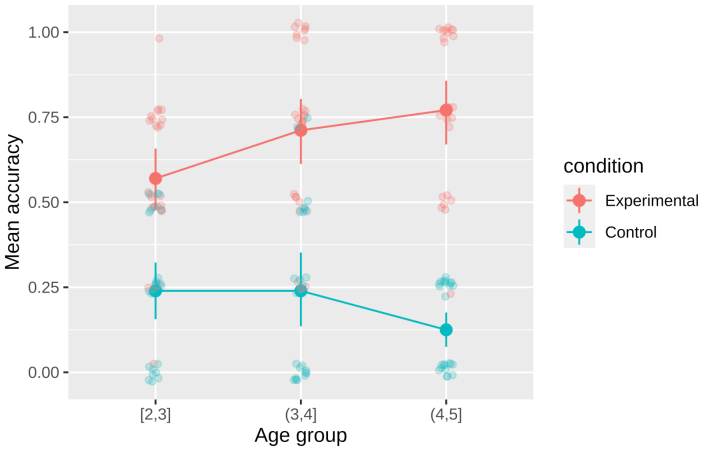


Figure 15.18: Standard "gray" themed Stiller figure.

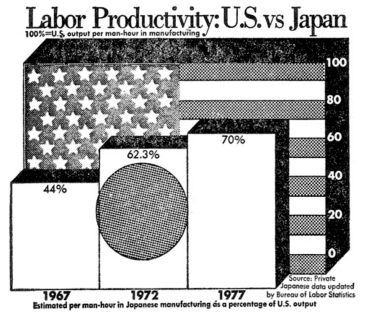


Figure 15.16: This figure uses a lot to ink to show three numbers, for a "ddi" of 0.2 (from the Washington Post, 1978; see Wainer (1984) for other examples).

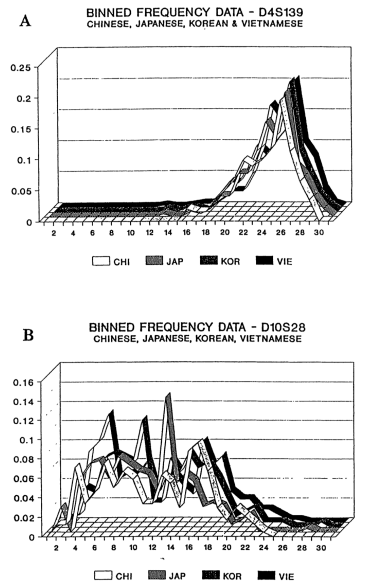


Figure 15.17: This figure uses complicated 3D ribbons to compare distributions across four countries (from Roeder 1994). How could the same data have been presented more legibly?

6 https://yutannihilation.github.io/allYourFigureAreBelongToUs/ggthemes/

Figure 15.19 shows a slightly more "styled" version of the same plot with labels directly on the plot and a lighter-weight theme.

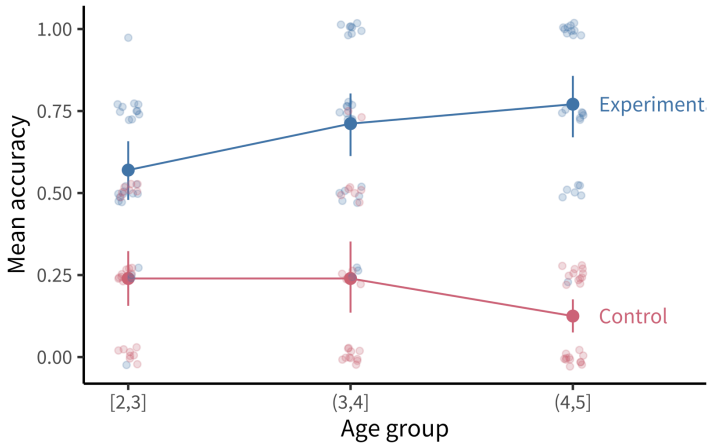


Figure 15.19: Custom themed Stiller figure with direct labels.

CODE

To produce the plot above, we've added a few styling elements including:

- The nice and minimal custom theme, with a larger font size.
- A more accessible color palette (`scale_colour_ptol()`) from the `ggthemes` package.
- Direct labels using `geom_dl()` from the `directlabels` package.

```
geom_dl(aes(label = condition), method = list("last.points", dl.trans(x = x + 0.5)))
```

Here are a few final tips for making good confirmatory visualizations:

- Make sure the font size of all text in your figures is legible and no smaller than other text in your paper (e.g. 10pt). This change may require, for example, making the axis breaks sparser, rotating text, or changing the aspect ratio of the figure.
- Another important tool to keep in your visualization arsenal is the **facet plot**. When your experimental design becomes more complex, consider breaking variables out into a *grid* of facets instead of packing more and more colors and line-styles onto the same axis. In other words, while higher information density is typically a good thing, you want to aim for the sweet spot before it becomes too dense and confusing. Remember Principle 2. When there is too much going on in every square inch, it is difficult to guide your reader's eye to the comparisons that actually matter, and spreading it out across facets gives you additional control over the salient patterns.

- Sometimes these principles come into conflict, and you may need to prioritize legibility over, for example, showing all of the data. For example, suppose there is an outlier orders of magnitude away from the summary statistics. If the axis limits are zoomed out to show that point, then most of the plot will be blank space! It is reasonable to decide that it is not worth compressing the key statistical question of your visualization into the bottom centimeter just to show one point. It may suffice to truncate the axes and note in the caption that a single point was excluded.
- Fix the axis labels! A common mistake is to keep the default shorthand you used to name variables in your plotting software instead of more descriptive labels (e.g., "RT" instead of "Reaction Time"). Use consistent terminology for different manipulations and measures in the main text and figures. If anything might be unclear in the figure, explain it in the caption.
- Different audiences may require different levels of detail. Sometimes it is better to collapse over secondary variables (even if they are included in your statistical models) in order to control the density of the figure and draw attention to the key question of interest.

15.2 Exploratory visualization

So far in this chapter we have focused on principles of *confirmatory* data visualization: how to make production-quality figures that convey the key pre-registered analyses without hiding sources of variability or misleading readers about the reliability of the results. Yet this is only one role that data visualization plays when doing science. An equally important role is called *exploratory visualization*: the more routine practice of understanding one's own data by visualizing it. This role is analogous to the sense of exploratory data analyses discussed in Chapter 11. We typically do not pre-register exploratory visualizations, and when we decide to include them in a paper they are typically in the service of a secondary argument (e.g., checking the robustness of an effect or validating that some assumption is satisfied).

This kind of visualization plays a ubiquitous role in a researcher's day-to-day activities. While confirmatory visualization is primarily audience-driven and concerned with visual communication, exploratory visualization is first and foremost a "cognitive tool" for the researcher. The first time we load in a new dataset, we start up a new feedback loop — we ask ourselves questions and answer them by making visualizations. These visualizations then raise further questions and are often our best

tool for debugging our code. In this section, we consider some best practices for exploratory visualization.

15.2.1 Examining distributional information

The primary advantage of exploratory visualization – the reason it is uniquely important for data science – is that it gives us access to holistic information about the distribution of the data, that cannot be captured in any single summary statistic. The most famous example is known as "Anscombe's quartet," a set of four datasets with identical statistics (Figure 15.20). They have the same means, the same variances, the same correlation, the same regression line, and the same R^2 value. Yet when they are plotted, they reveal striking structural differences. The first looks like a noisy linear relationship – the kind of idealized relationship we imagine when we imagine a regression line. But the second is a perfect quadratic arc, the third is a perfectly noiseless line with a single outlier, and the fourth is nearly categorical: every observation except one shares exactly the same x-value.

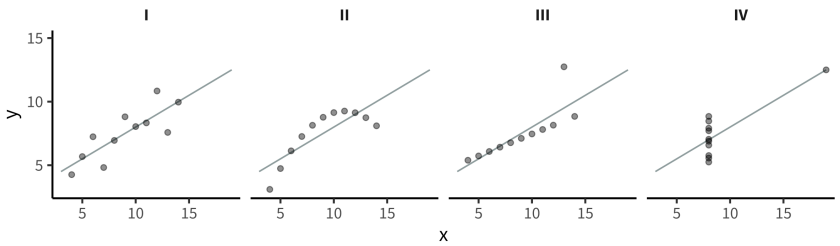


Figure 15.20: Anscombe's quartet (Anscombe 1973).

If our analyses are supposed to help us distinguish between different data-generating processes, corresponding to different psychological theories, it is clear that these four datasets would correspond to dramatically different theories even though they share the same statistics. Of course, there are arbitrarily many datasets with the same statistics, and most of these differences don't matter (this is why they are called "summary" statistics, after all!). Figure 15.21 and Table 15.1 show just how bad things can get when we rely on summary statistics. When we operationalize a theory's predictions in terms of a single statistic (e.g., a difference between groups or a regression coefficient) we can lose track of everything else that may be going on. Good visualizations force us to zoom out and take in the bigger picture.

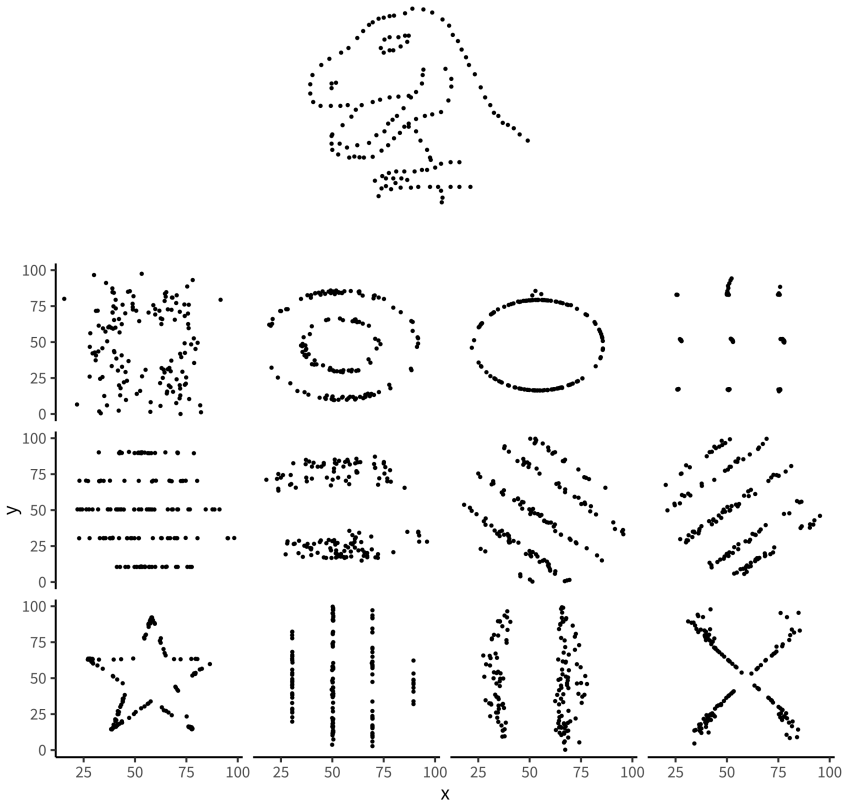


Figure 15.21: Originally inspired by the Datasaurus figure constructed by @albertocairo on Twitter (Cairo 2016) using the DrawMyData tool (http://robertgrantstats.co.uk/drawmydata.html), we can construct an arbitrary number of different graphs with exactly the same statistics (Matejka and Fitzmaurice 2017; Murray and Wilson 2021), such as the Datasaurus Dozen (Matejka and Fitzmaurice 2017).

Table 15.1: Summary statistics for each dataset in the Datasaurus Dozen (Matejka 2017).

dataset	mean_x	mean_y	sd_x	sd_y	cor_xy
away	54.3	47.8	16.8	26.9	-0.064
bullseye	54.3	47.8	16.8	26.9	-0.069
circle	54.3	47.8	16.8	26.9	-0.068
dino	54.3	47.8	16.8	26.9	-0.064
dots	54.3	47.8	16.8	26.9	-0.060
h_lines	54.3	47.8	16.8	26.9	-0.062
high_lines	54.3	47.8	16.8	26.9	-0.069
slant_down	54.3	47.8	16.8	26.9	-0.069
slant_up	54.3	47.8	16.8	26.9	-0.069
star	54.3	47.8	16.8	26.9	-0.063
v_lines	54.3	47.8	16.8	26.9	-0.069
wide_lines	54.3	47.8	16.8	26.9	-0.067
x_shape	54.3	47.8	16.8	26.9	-0.066

ACCIDENT REPORT

[Distributional] gorillas in our midst.

Many data scientists don't bother checking what their data looks like before proceeding to test specific hypotheses. Yanai and Lercher (2020) cleverly designed an artificial dataset for their students to test for such oversight. Each row of the dataset contained an individual's body mass index (BMI) and the number of steps they walked on a given day. While the spreadsheet looked innocuous, the data was constructed such that simply plotting the raw data revealed a picture of a gorilla. One group of 19 students was given an explicit set of hypotheses to test (e.g. about the relationship between BMI and steps). Fourteen of these students failed to notice a gorilla, suggesting that they evaluated these hypotheses without ever visualizing their data. Another group of 14 students were simply asked what, if anything, they could conclude (without being given explicit hypotheses). More of these students apparently made the visualization, but five of them still failed to notice the gorilla (Figure 15.22)!

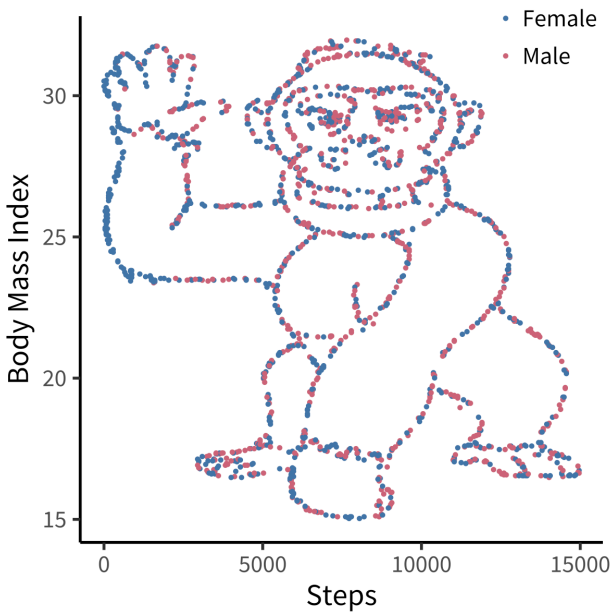


Figure 15.22: A dataset constructed by Yanai and Lercher (2020) which revealed a picture of a gorilla when the raw data were plotted.

While it may not be surprising that a group of students would take the shortest path to completing their assignment, similar concerns have been raised in much more serious cases concerning how experienced researchers could fail to notice obviously fraudulent data. For example, when the Datacolada bloggers -Datacolada (2021) made a simple histogram of the car mileage data reported in Shu et al. (2012; released publicly by Kristal et al. 2020), they were immediately able to observe that it followed a perfectly uniform distribution, truncated at exactly 50,000 miles (Figure 15.23). Given a little thought, this pattern should be extremely puzzling. Over a given period of time, we would typically expect something more bell-shaped: a small number of people will drive very little (e.g., 1000 miles), a small number of people will drive a lot (e.g., 50,000 miles), and most people will fall between these tails. So it is highly surprising to find exactly the same number of drivers in every mileage bin. While further specialized analyses revealed additional evidence of fraud (e.g. based on patterns of rounding and pairs of duplicated data points), this humble histogram was already enough to set off alarm bells. A recurring regret raised by the co-authors of this

paper is that they never thought to make this visualization before reporting their statistical tests.

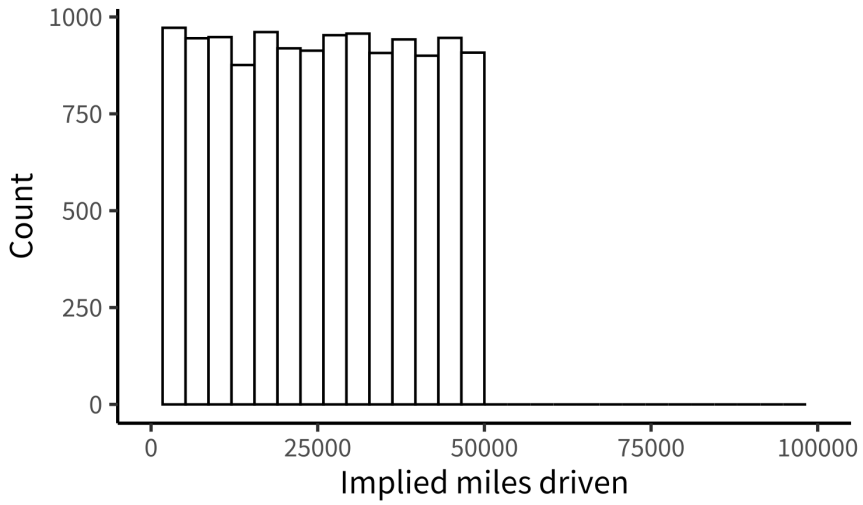


Figure 15.23: A suspiciously uniform distribution abruptly cutting off at 50k miles. Ring the alarm!

Our data are always messier than we expect. There might be a bug in our coding scheme, a column might be mislabeled, or might contain a range of values that we didn't expect. Maybe our design wasn't perfectly balanced, or something went wrong with a particular participant's keyboard presses. Most of the time, it's not tractable to manually scroll through our raw data looking for such problems. Visualization is our first line of defense for the all-important process of running "data diagnostics." If there is a weird artifact in our data, it will pop out if we just make the right visualizations.

15.2.1 Data diagnostics

So which visualizations should we start with? The best practice is to always start by making histograms of the raw data. As an example, let's consider the rich and interesting dataset shared by Blake, McAuliffe, and colleagues (2015) in their article "Ontogeny of fairness in seven societies." This article studies the emergence of children's reasoning about fairness – both when it benefits them and when it harms them – across cultures.

CODE

If you want to follow along with this example at home, you can load the data from our repository!

```
repo <- "https://raw.githubusercontent.com/langcog/experimentology/main/"
fairness_raw <- read_csv(file.path(repo, "data/viz/ontogeny_of_fairness.csv"))

fairness <- fairness_raw |>
  mutate(trial_num = trial |> str_remove("t") |> as.numeric(),
         trial_type = eq.uneq |> fct_recode("Equal" = "E", "Unequal" = "U"),
         condition = condition |> fct_recode("Advantageous" = "AI",
                                             "Disadvantageous" = "DI"),
         age = floor(actor.age.years),
         reject = decision == "reject") |>
  select(subj_id = actor.id, age, country, condition, trial_num, trial_type, reject) |>
  arrange(country, condition, subj_id, trial_num)
```

In this study, pairs of children played the "inequity game": they sat across from one another and were given a particular allocation of snacks. On some trials, each participant was allocated the same amount (Equal trials) and on some trials they were allocated different amounts (Unequal trials). One participant was chosen to be the "actor" and got to choose whether to accept or reject the allocation: in the case of rejection, neither participant got anything. The critical manipulation was between two forms of inequity. Some pairs were assigned to the Disadvantageous condition, where the actor was allocated less than their partner on Unequal trials (e.g. 1 vs. 4). Others were assigned to the Advantageous condition, where they were allocated more (e.g. 4 vs. 1).

The confirmatory **design plot** for this study would focus on contrasting developmental trajectories for Advantageous vs. Disadvantageous inequality. However, this is a complex, multivariate dataset, including 866 pairs from different age groups and different testing sites across the world which used subtly different protocols. How might we go about the process of exploratory visualization for this dataset?

15.2.2 Plot data collection details

Let's start by getting a handle on some of the basic sample characteristics. For example, how many participants were in each age bin (Figure 15.24)?

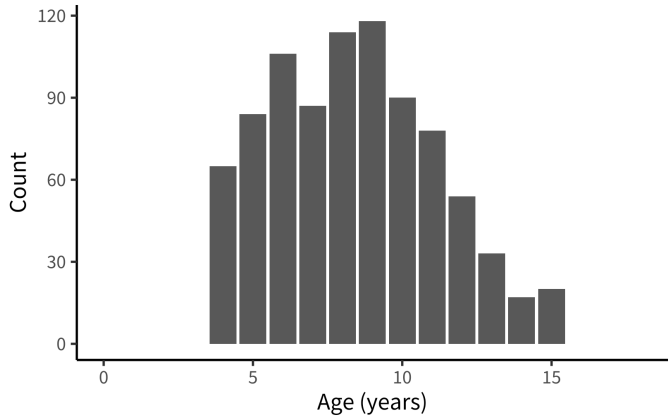


Figure 15.24: Participants by age in the Blake data.

CODE

Exploratory histograms are often a combination of an aggregation step and a plotting step. In the aggregation step, we make use of the convenience `count()` function, which gives the number (`n`) of rows in a particular grouping. Here we `count()` twice in order to get first one row per participant and then count the number of participants within each age group.

```
fairness_by_age <- fairness |>
  count(age, subj_id) |>
  count(age)
```

And then we plot using `ggplot()`:

```
ggplot(fairness_by_age, aes(x = age, y = n)) +
  geom_col() +
  xlim(0, 18) +
  labs(x = "Age (years)", y = "Count")
```

An alternative (perhaps more elegant) workflow here would be to use a histogram:

```
fairness_by_age <- fairness |>
  count(age, subj_id)

ggplot(fairness_by_age, aes(x = age)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Age (years)", y = "Count")
```

Histograms are intended by `ggplot` to be for continuous data, however, and so they don't give the discrete bars that our earlier workflow did.

How many participants were included from each country (Figure 15.25)?

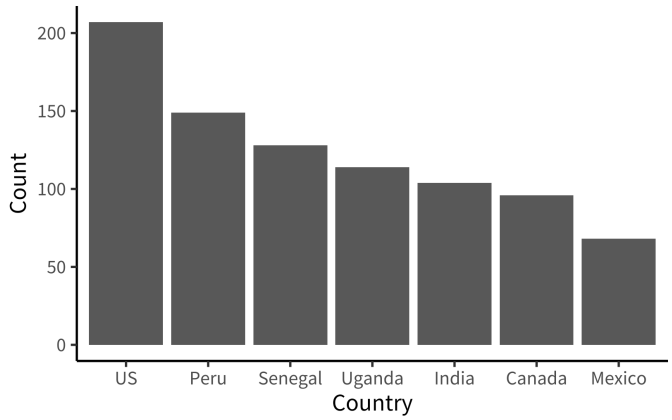


Figure 15.25: Participants by country in the Blake data.

CODE

Here we are going to make things even terser and use a pipe chain that *includes* the `ggplot()` call, just so we are writing only a single call to produce our plot. It's up to you whether you think this enhances the readability of your code or decreases it. We find that it's sometimes useful when you don't plan on keeping the intermediate data frame for any other use than plotting.

```
fairness |>
  count(country, subj_id) |>
  count(country) |>
  mutate(country = fct_reorder(country, -n)) |>
  ggplot(aes(x = country, y = n)) +
    geom_col() +
    labs(x = "Country", y = "Count")
```

If you use this technique, be careful to use pipe (`|>` or `%>%`) between function calls but use (`+`) between `ggplot` layers!

The only other trick to point out here is that we use the `fct_reorder()` call to order the levels of the `country` factor in descending order. This function is found in the very useful `forcats` package of the `tidyverse`, which contains all sorts of functions for working with factors.

Are ages roughly similar across each country (Figure 15.26)?

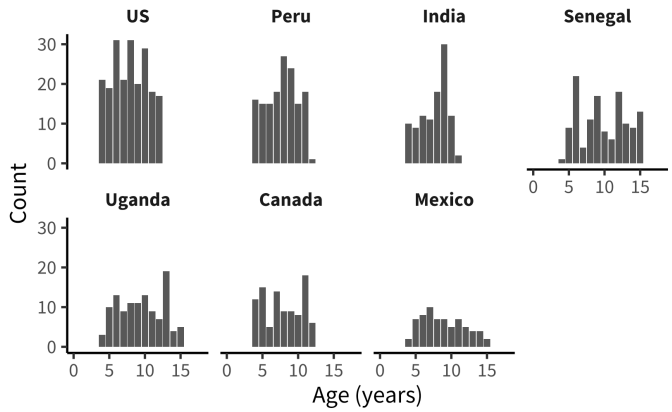


Figure 15.26: Age distribution across countries in the Blake data.

CODE

This next plot simply combines the grouping factors of each of the last two plots, and uses `facet_wrap()` to show a separate histogram by country:

```
fairness |>
  count(country, age, subj_id) |>
  count(country, age) |>
  mutate(country = fct_reorder(country, -n)) |>
  ggplot(aes(x = age, y = n)) +
    facet_wrap(vars(country), ncol = 4) +
    geom_col() +
    xlim(0, 18) +
    labs(x = "Age (years)", y = "Count")
```

These exploratory visualizations help us read off some descriptive properties of the sample. For example, we can see that age ranges differ somewhat across sites: the maximum age is 11 in India but 15 in Mexico. We can also see that age groups are fairly imbalanced: in Canada, there are 18 11-year-olds but only 5 6-year-olds.

None of these properties are problematic, but seeing them gives us a degree of awareness that could shape our downstream analytic decisions. For example, if we did not appropriately model random effects, our estimates would be dominated by the countries with larger sample sizes. And if we were planning to compare specific groups of 6-year-olds (for some reason), this analysis would be underpowered.

15.2.3 Explorating distributions

Now that we have a handle on the sample, let's get a sense of the dependent variable: the participant's decision to accept or reject the allocation. Before we start taking means, let's look at how the "rejection rate" variable is distributed. We'll aggregate at the participant level, and check the frequency of different rejection rates, overall (Figure 15.27).

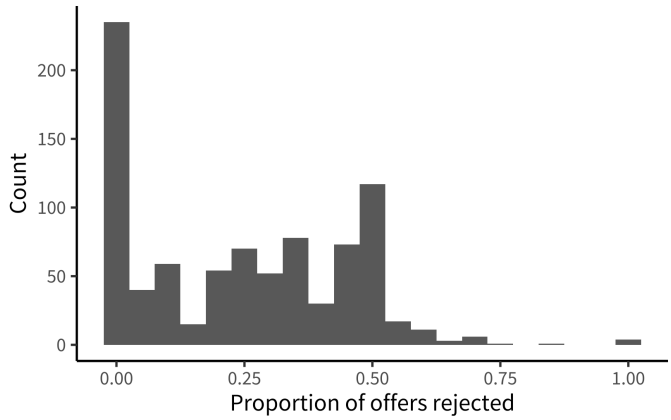


Figure 15.27: Rejection rates in the Blake data.

CODE

Rejection rate is a continuous variable, so we switch to using a histogram in this case, choosing .05 as a reasonable bin width to see the distribution.

```
fairness_by_subj <- fairness |>
  filter(!is.na(trial_type)) |>
  group_by(subj_id) |>
  summarise(mean_reject = mean(reject, na.rm = TRUE))

ggplot(fairness_by_subj, aes(x = mean_reject)) +
  geom_histogram(binwidth = .05) +
  labs(x = "Proportion of offers rejected", y = "Count")
```

We notice that many participants (27%) never reject in the entire experiment. This kind of "zero-inflated" distribution is not uncommon in psychology, and may warrant special consideration when designing the statistical model. We also notice that there is clumping around certain values. This clumping leads us to check how many trials each participant is completing (Figure 15.28).

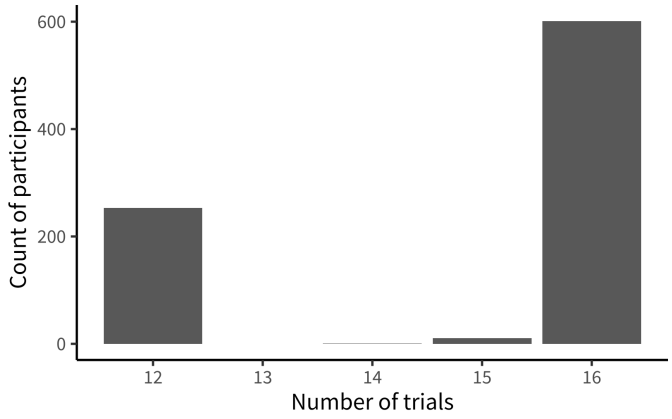


Figure 15.28: Trials per participant in the Blake data.

CODE

This histogram is very similar to the ones above; however, we now use `count()` twice, first getting the trial counts for each participant and then counting how many times each count occurs overall!

```
fairness |>
  filter(!is.na(trial_type)) |>
  count(subj_id) |>
  count(n) |>
  ggplot(aes(x = n, y = nn)) +
    geom_col() +
    labs(x = "Number of trials", y = "Count of participants")
```

There's some variation here: most participants completed 17 trials, but some participants completed 8 trials, and a small number of participants have 14 or 15. Given the logistical complexity of large multi-site studies, it is common to have some changes in experimental protocol across data collection. Indeed, looking at the supplement for the study, we see that while India and Peru had 12 trials, additional trials were added at the other sites. In a design where the number of trials was carefully controlled, seeing unexpected numbers here (like the 14 or 15 trial bins) are clues that something else may be going on in the data. In this case, it was a small number of trials with missing data. More generally, seeing this kind of signal in a visualization of our own data typically leads us to look up the participant IDs in these bins and manually inspect their data to see what might be going on.

15.2.4 Hypothesis-driven exploration

Finally, we can make a few versions of the design plot that are broken out by different variables. Let's start by just looking at the data from the largest site (Figure 15.29).

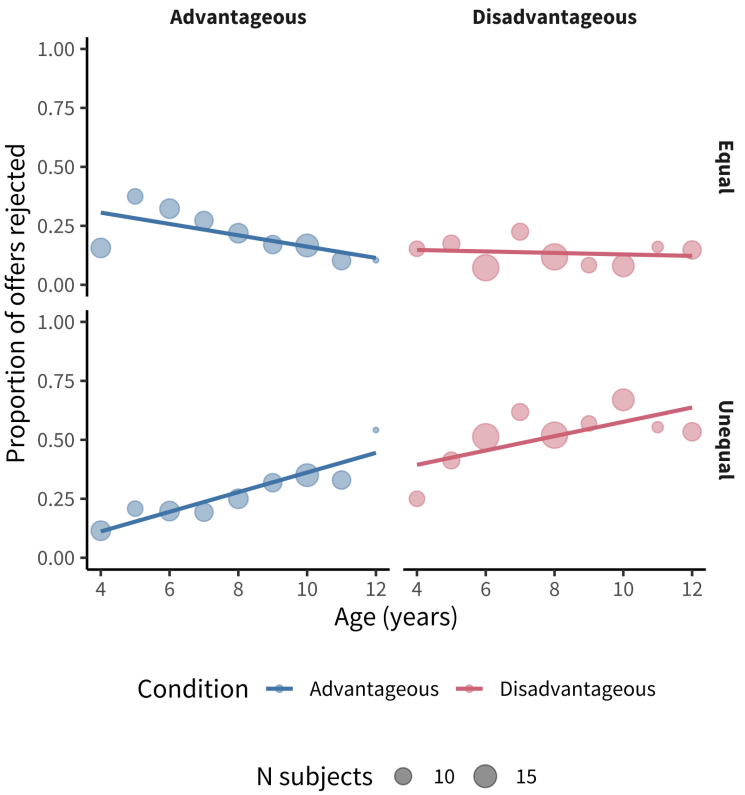


Figure 15.29: Rejection rates in the US data from Blake, plotted by age.

Figure 15.29 is not a figure we'd put in a paper, but it helps us get a sense of the pattern in the data. There appears to be an age trend that's specific to the Unequal trials, with rejection rates rising over time (compared to roughly even or decreasing rates in the Equal trials). Meanwhile, rejection rates for the Disadvantageous group also seem slightly higher than those in the Advantageous group.

CODE

Here, we are using `geom_smooth()` to overlay regression trends over the raw data. `geom_smooth()` takes a number of different options corresponding to different smoothing techniques. Non-parametric smoothing can be a good choice for exploratory visualizations if you have a lot of data and want to make minimal assumptions about the form of the trend.

Here, however, we show the linear regression trend, `geom_smooth(method = "lm")`, which better corresponds to the predictions of the study and the statistical model being used (see Chapter 7). Other regression forms can be specified with the `formula` argument. For example, we could show quadratic smoothing with `geom_smooth(method = "lm", formula = y ~ poly(x, 2))`. The form of smoothing you use may differ across exploratory and confirmatory visualizations. In a confirmatory visualization — if you are going to include a smoothing curve — it is typically best to use the one specified by your statistical model, as the slopes will correspond to the inferences being testing.

We begin by making a summary dataset:

```
fairness_by_age <- fairness |>
  filter(!is.na(reject)) |>
  group_by(country, trial_type, condition, age, subj_id) |>
  summarise(mean_reject_subj = mean(reject, na.rm = TRUE)) |>
  group_by(country, trial_type, condition, age) |>
  summarise(mean_reject_age = mean(mean_reject_subj, na.rm = TRUE),
            n_subj = n()) |>
  ungroup()
```

Then we can create the visualization:

```
fairness_by_age |> filter(country == "US") |>
  ggplot(aes(x = age, y = mean_reject_age, color = condition)) +
    facet_grid(vars(trial_type), vars(condition)) +
    geom_smooth(method = "lm", se = FALSE) +
    geom_point(aes(size = n_subj), alpha = .5) +
    ylim(c(0, 1)) +
    labs(x = "Age (years)", y = "Proportion of offers rejected",
         color = "Condition", size = "N subjects") +
    theme(legend.position = "bottom", legend.box = "vertical")
```

We often find it convenient to filter the summary dataset in the plotting call, so that we can reuse it again.

Now let's re-bin the data into two-year age groups so that individual point estimates are a bit more reliable, and add the other countries back in.[7]

[7] Binning data is a trick that we often use for reducing complexity in a plot when data are noisy. It should be used with care, however, since different binning decisions can sometimes lead to different conclusions. Here we tried several binning intervals and decided that two-year age bins showed the underlying trends pretty well.

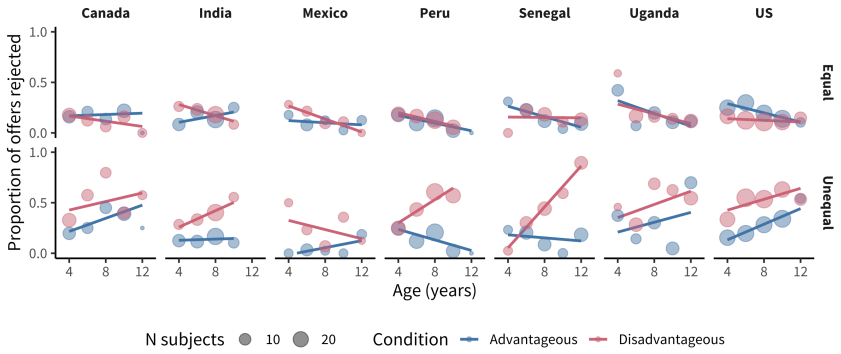


Figure 15.30: Rejection rates by age for all data in the Blake dataset.

Figure 15.30 is now looking much closer to a quick-and-dirty version of a "design plot" we might include in a paper. The DV (rejection rate) is on the y-axis, and the primary variable of interest (age) is on the x-axis. Other elements of the design (country and trial type) are mapped to color and facets, respectively.

CODE

Despite the difference between the plot above and this one, the code to produce them is actually very similar. The only difference is the creation of the binned variable and a slight shift of aesthetic and faceting variables.

```
fairness_by_age_binned <- fairness |>
  filter(!is.na(reject)) |>
  mutate(age_binned = floor(age / 2) * 2) |>
  group_by(country, trial_type, condition, age_binned, subj_id) |>
  summarise(mean_reject_subj = mean(reject, na.rm = TRUE)) |>
  group_by(country, trial_type, condition, age_binned) |>
  summarise(mean_reject_age = mean(mean_reject_subj, na.rm = TRUE),
            n = n())  |>
  ungroup()

ggplot(fairness_by_age_binned,
       aes(x = age_binned, y = mean_reject_age, color = condition)) +
  facet_grid(vars(trial_type), vars(country)) +
  geom_smooth(method = "lm", se = FALSE, aes(weight=n)) +
  geom_point(alpha = .5, aes(size = n)) +
  scale_x_continuous(breaks = seq(4, 12, 4), limits = c(3,13)) +
  scale_y_continuous(limits = c(0, 1), breaks = c(0, .5, 1)) +
  labs(x = "Age (years)", y = "Proportion of offers rejected",
       color = "Condition", size = "N subjects") +
  theme(legend.position = "bottom")
```

15.2.5 *Visualization as debugging*

The point of exploratory visualization is to converge toward a better understanding of what's going on in your data. As you iterate through different exploratory visualizations, *stay vigilant*! Think about what you expect to see before making the plot, then ask yourself whether you got what you expected. You can think of this workflow as a form of "visual debugging". You might notice a data point with an impossible value, such as a proportion greater than 1 or a reaction time less than 0. Or you might notice weird clusters or striations, which might indicate heterogeneity in data entry (perhaps different coders used slightly different rubrics or rounded in different ways). You might notice that an attribute is missing for some values, and trace it back to a bug reading in the data or merging data frames (maybe there was a missing comma in our `csv` file). If you see anything that looks weird, track it down until you understand why it's happening. Bugs that are subtle and invisible in other parts of the analysis pipeline will often pop out as red flags in visualizations.

15.3 *Chapter summary: Visualization*

This chapter has given a short review of the principles of data visualization, especially focusing on the needs of experimental psychology, which are often quite different than those of other fields. We particularly focused on the need to make visualization part of the experimenter's analytic workflow. Picking up the idea of a "default model" from Chapter 7, we discussed a default "design plot" that reflects the key choices made in the experimental design. Within this framework, we then discussed different visualizations of distribution and variability that better align our graphics with the principles of measurement and attention to raw data that we have been advocating throughout.

DISCUSSION QUESTIONS

1. Choose a recent piece of research that you've heard about and try to sketch the "design plot" with pencil and paper. What does and doesn't work? How does your sketch differ from the visualizations in the paper?
2. The "design plot" idea that we've discussed here can run into problems when an experimental design is too complex to show on a single plot. Imagine you had data from a trial of attention deficit hyperactivity disorder (ADHD) treatment that manipulated both whether a medication was given and whether patients received therapy in a crossed design. The researchers measured two different outcomes: parent report symptom severity and teacher report symptom severity in four different time-points (baseline, 3 months, 6 months, and 9 months). How could you show the data from such an experiment in a transparent way?

READINGS

There are many good introductions to data visualization. Here are two social-science focused books whose advice we agree with and that also contain a lot of practical information and helpful R code for the same packages we use here.

- Healy, K. (2019). *Data Visualization: A Practical Introduction*. Princeton University Press. Princeton University Press. Available free online at https://socviz.co.
- Wilke, C. O. (2019). *Fundamentals of Data Visualization*. O'Reilly Media. Available free online at https://clauswilke.com/dataviz/.

For a more classical treatment, see:

- Tukey, J. W. (1977). *Exploratory data analysis*. Pearson.
- Tufte, E. R. (1997). *The Visual Display of Quantitative Information*. Graphics Press.

References

Allen, Micah, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, and Rogier A Kievit. 2019. "Raincloud Plots: A Multi-Platform Tool for Robust Data Visualization." *Wellcome Open Research* 4.

Anscombe, Francis J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21.

Barnett, Samuel A, Thomas L Griffiths, and Robert D Hawkins. 2022. "A Pragmatic Account of the Weak Evidence Effect." *Open Mind*, 1–14.

Blake, PR, K McAuliffe, J Corbit, TC Callaghan, O Barry, A Bowie, L Kleutsch, et al. 2015. "The Ontogeny of Fairness in Seven Societies." *Nature* 528 (7581): 258–61.

Börner, Katy, Andreas Bueckle, and Michael Ginda. 2019. "Data Visualization Literacy: Definitions, Conceptual Frameworks, Exercises, and Assessments." *Proceedings of the National Academy of Sciences* 116 (6): 1857–64.

Brody, Howard, Michael Russell Rip, Peter Vinten-Johansen, Nigel Paneth, and Stephen Rachman. 2000. "Map-Making and Myth-Making in Broad Street: The London Cholera Epidemic, 1854." *The Lancet* 356 (9223): 64–68.

Cairo, Alberto. 2016. "Download the Datasaurus: Never Trust Summary Statistics Alone; Always Visualize Your Data." 2016. http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html.

Cleveland, William S, and Robert McGill. 1984. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." *Journal of the American Statistical Association* 79 (387): 531–54.

Coppock, Alexander. 2019. "Visualize as You Randomize: Design-Based Statistical Graphs for Randomized Experiments." In *Advances in Experimental Political Science*, edited by James N. Druckman and Donald P. Green. Cambridge University Press.

Datacolada. 2021. "Evidence of Fraud in an Influential Field Experiment about Dishonesty." https://datacolada.org/98.

Friendly, Michael, and Howard Wainer. 2021. *A History of Data Visualization and Graphic Communication*. Harvard University Press.

Gelman, Andrew, and Antony Unwin. 2013. "Infovis and Statistical Graphics: Different Goals, Different Looks." *J. Comput. Graph. Stat.* 22 (1): 2–28.

Halliday, Stephen. 2001. "Death and Miasma in Victorian London: An Obstinate Belief." *British Medical Journal* 323 (7327): 1469–71.

Kristal, Ariella S, Ashley V Whillans, Max H Bazerman, Francesca Gino, Lisa L Shu, Nina Mazar, and Dan Ariely. 2020. "Signing at the Beginning Versus at the End Does Not Decrease Dishonesty." *Proceedings of the National Academy of Sciences* 117 (13): 7103–7.

Mackinlay, Jock. 1986. "Automating the Design of Graphical Presentations of Relational Information." *ACM Transactions On Graphics* 5 (2): 110–41.

Matejka, Justin, and George Fitzmaurice. 2017. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1290–94.

Murray, Lori L, and John G Wilson. 2021. "Generating Data Sets for Teaching the Importance of Regression Analysis." *Decision Sciences Journal of Innovative Education* 19 (2): 157–66.

Roeder, Kathryn. 1994. "DNA Fingerprinting: A Review of the Controversy." *Statistical Science*, 222–47.

Shu, Lisa L, Nina Mazar, Francesca Gino, Dan Ariely, and Max H Bazerman. 2012. "Signing at the Beginning Makes Ethics Salient and Decreases Dishonest Self-Reports in Comparison to Signing at the End." *Proceedings of the National Academy of Sciences* 109 (38): 15197–200.

Snow, John. 1854. "Dr. Snow's Report." In *Report on the Cholera Outbreak in the Parish of St. James, Westminster, During the Autumn of 1854*. John Churchill. https://johnsnow.matrix.msu.edu/work.php?id=15-78-55.

———. 1855. *On the Mode of Communication of Cholera*. John Churchill.

Stiller, Alex J, Noah D Goodman, and Michael C Frank. 2015. "Ad-Hoc Implicature in Preschool Children." *Language Learning and Development* 11 (2): 176–90.

Tufte, E. R. 1983. *The Visual Display of Quantitative Information*. Graphics Press.

Wainer, Howard. 1984. "How to Display Data Badly." *The American Statistician* 38 (2): 137–47.

Yanai, I, and M Lercher. 2020. "A Hypothesis Is a Liability." *Genome Biology* 21 (1): 231–31.

Zacks, Jeffrey M, and Steven L Franconeri. 2020. "Designing Graphs for Decision-Makers." *Policy Insights from the Behavioral and Brain Sciences* 7 (1): 52–63.

16 Meta-analysis

Learning goals

- Discuss the benefits of synthesizing evidence across studies
- Conduct a simple fixed- or random-effects meta analysis
- Reason about the role of within- and across-study biases in meta-analysis

Throughout this book, we have focused on how to design individual experiments that maximize measurement precision and minimize bias. But even when we do our best to get a precise, unbiased estimate in an individual experiment, one study can never be definitive. Variability in participant demographics, stimuli, and experimental methods may limit the generalizability of our findings. Additionally, even well-powered individual studies have some amount of statistical error, limiting their precision. Synthesizing evidence across studies is critical for developing a balanced and appropriately evolving view of the overall evidence on an effect of interest and for understanding sources of variation in the effect.

Synthesizing evidence rigorously takes more than putting a search term into Google Scholar, downloading articles that look topical or interesting, and qualitatively summarizing your impressions of those studies. While this ad-hoc method can be an essential first step in performing a literature review (Grant and Booth 2009), it is not systematic and doesn't provide a *quantitative* summary of a particular effect. Further, it doesn't tell you anything about potential biases in the literature – for example, a bias for the publication of positive effects.

To address these issues, a more systematic, quantitative review of the literature is often more informative. This chapter focuses on a specific type of quantitative review called **meta-analysis**: a method for combining effect sizes across different studies. (If you need a refresher on effect size, see Chapter 5, where we introduce the concept).[1] We incude a chapter on meta-analysis in Experimentology because we believe it's an important tool that can focus experimental researchers on issues of measurement precision and bias reduction, two of our key themes.

[1] We'll primarily be using Cohen's d, the standardized difference between means, which we introduced in Chapter 5. There are many more varieties of effect size available, but we focus here on d because it's common and easy to reason about in the context of the statistical tools we introduced in the earlier sections of the book.

By combining information from multiple studies, meta-analysis often provides more precise estimates of an effect size than any single study. In addition, meta-analysis also allows the researcher to look at the extent to which an effect varies across studies. If an effect does vary across studies, meta-analysis also can be used to test whether certain study characteristics systematically produce different results (e.g., whether an effect is larger in certain populations).

CASE STUDY

Towel reuse by hotel guests

Imagine you are staying in a hotel and you have just taken a shower. Do you throw the towels on the floor or hang them back up again? In a widely-cited study on the power of social norms, Goldstein, Cialdini, and Griskevicius (2008) manipulated whether a sign encouraging guests to reuse towels focused on environmental impacts (e.g., "help reduce water use") or social norms (e.g., "most guests re-use their towels"). Across two studies, they found that guests were significantly more likely to reuse their towels after receiving the social norm message (Study 1: odds ratio [OR] = 1.46, 95% CI [1.00, 2.16], $p = .05$; Study 2: OR = 1.35, 95% CI [1.04, 1.77], $p = .03$).

However, five subsequent studies by other researchers did not find significant evidence that social norm messaging increased towel reuse. (ORs ranged from 0.22 to 1.34, and no hypothesis-consistent p-value was less than .05). This caused many researchers to wonder if there is any effect at all. To examine this question, Scheibehenne, Jamil, and Wagenmakers (2016) statistically combined evidence across the studies via meta-analysis. This meta-analysis indicated that using social norm messages did significantly increase hotel towel reuse, on average (OR = 1.26, 95% CI [1.07, 1.46], $p < .005$). This case study demonstrates an important strength of meta-analysis: by pooling evidence from multiple studies, meta-analysis can generate more powerful insights than any one study alone. We will also see how meta-analysis can be used to assess variability in effects across studies.

Meta-analysis often teaches us something about a body of evidence that we do not intuitively grasp when we casually read through a bunch of articles. In the above case study, merely reading the individual studies might give the impression that social norm messages do not increase hotel towel re-use. But meta-analysis indicated that the average effect is beneficial, although there might be substantial variation in effect sizes across studies.[2]

[2] Given the billions of hotel bookings worldwide every year, even a small effect might have lead to a substantial environmental impact!

16.1 The basics of evidence synthesis

As we explore the details of conducting a meta-analysis, we'll turn to another running example: a meta-analysis of studies investigating the "contact hypothesis" on intergroup relations.

According to the contact hypothesis, prejudice towards members of minority groups can be reduced through intergroup contact interventions, in which members of majority and minority groups work together to

pursue a common goal (Allport, Clark, and Pettigrew 1954). To aggregate the evidence on the contact hypothesis, Paluck, Green, and Green (2019) meta-analyzed studies that tested the effects of randomized intergroup contact interventions on long-term prejudice-related outcomes.

Using a systematic literature search, Paluck, Green, and Green (2019) searched for all papers that tested these effects and then extracted effect size estimates from each paper.[3] Because not every paper reports standardized effect sizes – or even means and standard deviations for every group – this process can often involve scraping information from plots, tables, and statistical tests to try to reconstruct effect sizes.[4]

Following best practices for meta-analysis (where there are almost never privacy concerns to worry about), Paluck, Green, and Green (2019) shared their data openly. The first few lines are shown in Table 16.1. We'll use these data as our running example throughout.

Table 16.1: First few lines of extracted effect sizes (d) and their variances (var_d) in the Paluck, Green, and Green (2019) meta-analysis.

name	pub_date	target	n_total	d	var_d
Boisjoly 06 B	2006	race	1243	0.030	0.006
Sorensen 10	2010	race	597	0.302	0.007
Scacco 18	2018	religion	474	0.000	0.010
Finseraas 2017	2017	foreigners	577	0.000	0.011
Sheare 74	1974	disability	400	1.059	0.011
Barnhardt 09	2009	religion	312	0.395	0.015

As we've seen throughout this book, visualizing data before and after analysis helps benchmark and check our intuitions about the formal statistical results. In a meta-analysis, a common way to plot effect sizes is the **forest plot**, which depicts individual studies' estimates and confidence intervals.[5] In the forest plot in Figure 16.1, the larger squares correspond to more precise studies; notice how much narrower their confidence intervals are than the confidence intervals of less precise studies.

[3] This book will not cover the process of conducting a systematic literature search and extracting effect sizes, but these topics are critical to understand if you plan to conduct a meta-analysis or other evidence synthesis. Our experience is that extracting effect sizes from papers with inconsistent reporting standards can be especially tricky, so it can be helpful to talk to someone with experience in meta-analysis to get advice about this.

[4] For example, if the outcome variable is continuous, we could estimate Cohen's d from group means and standard deviations reported in the paper, even without having access to raw data.

[5] You can ignore for now the column of percentages and the final line, "RE Model"; we will return to these later.

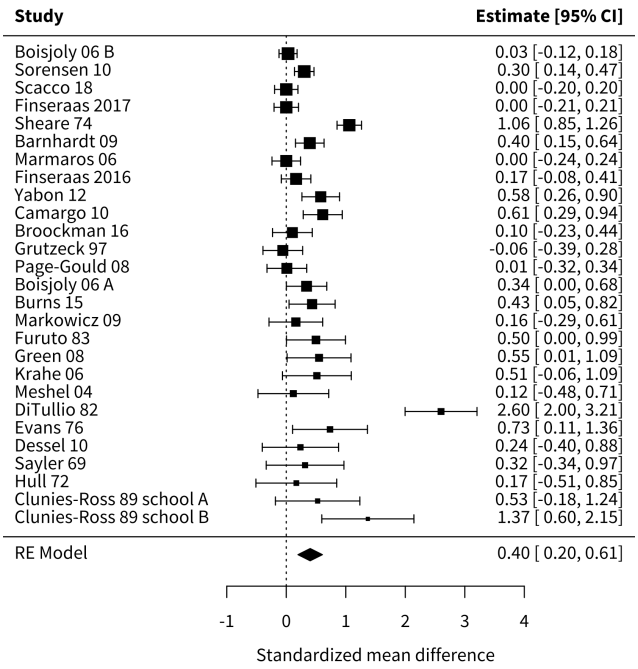


Figure 16.1: Forest plot for Paluck, Green, and Green (2019) meta-analysis. Studies are ordered from smallest to largest standard error.

CODE

In this chapter, we use the wonderful `metafor` package (Viechtbauer 2010). With this package, you must first fit your meta-analytic model. But once you've fit your model `mod`, you can simply call `forest(mod)` to create a plot like the one above.

16.1.1 How not to synthesize evidence

Many people's first instinct in evidence synthesis is to count how many studies supported versus did not support the hypothesis under investigation. This technique usually amounts to counting the number of studies with "significant" p-values, since – for better or for worse – "significance" is largely what drives the take-home conclusions researchers report (McShane and Gal 2017; Nelson, Rosenthal, and Rosnow 1986). In meta-analysis, we call this practice of counting the number of significant p-values **vote-counting** (Borenstein et al. 2021). For example, in the Paluck, Green, and Green (2019) meta-analysis, almost all studies had a positive effect size, but only 12 of 27 were significant. So, based on this vote-count, we would have the impression that most studies do not support the contact hypothesis.

Many qualitative literature reviews use this vote-counting approach, although often not explicitly. Despite its intuitive appeal, vote-counting can be very misleading because it characterizes evidence solely in terms of dichotomized p-values, while entirely ignoring effect sizes. In Chapter 3, we saw how fetishizing statistical significance can mislead us when we consider individual studies. These problems also apply when considering multiple studies.

For example, small studies may consistently produce non-significant effects due to their limited power. But when many such studies are combined in a meta-analysis, the meta-analysis may provide strong evidence of a positive average effect. Inversely, many studies might have statistically significant effects, but if their effect sizes are small, then a meta-analysis might indicate that the average effect size is too small to be practically meaningful. In these cases, vote-counting based on statistical significance can lead us badly astray (Borenstein et al. 2021). To avoid these pitfalls, meta-analysis combines the effect size estimates from each study (not just their p-values), weighting them in a principled way.

16.1.2 Fixed-effects meta-analysis

If vote-counting is a bad idea, how should we combine results across studies? Another intuitive approach might be to average effect sizes from each study. For example, in Paluck et al.'s meta-analysis, the mean of the studies' effect size estimates is 0.44. This averaging approach is a step in the right direction, but it has an important limitation: averaging effect size estimates gives equal weight to each study. A small study (e.g., Clunies-Ross and O'Meara 1989 with N=30) contributes as much to the mean effect size as a large study (e.g., Boisjoly et al. 2006 with N=1243). Larger studies provide more precise estimates of effect sizes than small studies, so weighting all studies equally is not ideal. Instead, larger studies should carry more weight in the analysis.

To address this issue, **fixed-effects meta-analysis** uses a **weighted average** approach. Larger, more precise studies are given more weight in the calculation of the overall effect size. Specifically, each study is weighted by the inverse of its variance (i.e., the inverse of its squared standard error). This makes sense because larger, more precise studies have smaller variances, and thus get more weight in the analysis.

In general terms, the fixed-effect pooled estimate is:

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

where k is the number of studies, $\hat{\theta}_i$ is the point estimate of the i^{th} study, and $w_i = 1/\hat{\sigma}_i^2$ is study i's weight in the analysis (i.e., the inverse of its variance).[6]

Using the fixed-effects formula, we can estimate that the overall effect size in Paluck et al.'s meta-analysis is a standardized mean difference of $\hat{\mu}$ = 0.28; 95% confidence interval [0.23, 0.34]; $p < .001$. Because Cohen's d is our effect size index, this estimate would suggest that intergroup contact decreased prejudice by 0.28 standard deviations.

[6] If you are curious, the standard error of the fixed-effect $\hat{\mu}$ is $\frac{1}{\sum_{i=1}^{k} w_i}$. This standard error can be used to construct a confidence interval or p-value, as described in Chapter 6.

CODE

Fitting meta-analytic models in `metafor` is quite simple. For example, for the fixed effects model above, we simply ran the `rma()` function and specified that we wanted a fixed effects analysis.

```
fe_model <- rma(yi = d, vi = var_d, data = paluck, method = "FE")
```

Then `summary(fe_model)` gives us the relevant information about the fitted model.

16.1.3 Limitations of fixed-effects meta-analysis

One of the limitations of fixed-effect meta-analysis is that it assumes that the true effect size is, well, *fixed*! In other words, fixed-effect meta-analysis assumes that there is a single effect size that all studies are estimating. This is a stringent assumption. It's easy to imagine that it could be violated. Imagine, for example, that intergroup contact decreased prejudice when the group succeeded at its joint goal, but *increased* prejudice when the group failed. If we meta-analyzed two studies under these conditions – one in which intergroup contact substantially increased prejudice, and one in which intergroup contact substantially decreased prejudice – it might appear that the true effect of intergroup contact was close to zero, when in fact both of the meta-analyzed studies had large effects.

In Paluck et al.'s meta-analysis, studies differed in several ways that could lead to different true effects. For example, some studies recruited adult participants while others recruited children. If intergroup contact is more or less effective for adults versus children, then it is misleading to talk about a single (i.e., "fixed") intergroup contact effect. Instead, we would say that the effects of intergroup contact vary across studies, an idea called **heterogeneity**.

Does the concept of heterogeneity remind you of anything from when we analyzed repeated-measures data in Chapter 7 on models? Recall

that, with repeated-measures data, we had to deal with the possibility of heterogeneity across participants – and of the ways we did so was by introducing participant-level random intercepts to our regression model. It turns out that we can do a similar thing in meta-analysis to deal with heterogeneity across studies.

16.1.4 Random-effects meta-analysis

While fixed-effect meta-analysis essentially assumes that all studies in the meta-analysis have the same population effect size, μ, random-effects meta-analysis instead assumes that study effects come from a normal distribution with mean μ and standard deviation τ.[7] The larger the standard deviation, τ, the more heterogeneous the effects are across studies. A random-effects model then estimates both μ and τ, for example by maximum likelihood (DerSimonian and Laird 1986; Brockwell and Gordon 2001).

Like fixed-effect meta-analysis, the random-effects estimate of $\hat{\mu}$ is still a weighted average of studies' effect size estimates:

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

However, in random-effects meta-analysis, the inverse-variance weights now incorporate heterogeneity: $w_i = 1/\left(\hat{\tau}^2 + \hat{\sigma}_i^2\right)$. Where before we had one term in our weights, now we have two. That is because these weights represent the inverse of studies' *marginal* variances, taking into account both statistical error due to their finite sample sizes ($\hat{\sigma}_i^2$ as before) and also genuine effect heterogeneity ($\hat{\tau}^2$).

Conducting a random-effects meta-analysis of Paluck et al.'s dataset yields $\hat{\mu} = 0.4$; 95% confidence interval [0.2, 0.61]; $p < .001$. That is, *on average across studies*, intergroup contact was associated with a decrease in prejudice of 0.4 standard deviations, substantially larger than the estimate from the fixed effects model. This meta-analytic estimate is shown as the bottom line of Figure 16.1.

[7] Technically, other specifications of random-effects meta-analysis are possible. For example, robust variance estimation does not require making assumptions about the distribution of effects across studies (Hedges, Tipton, and Johnson 2010). These approaches also have other substantial advantages, like their ability to handle effects that are clustered [e.g., because some papers contribute multiple estimates; Hedges, Tipton, and Johnson (2010); Pustejovsky and Tipton (2021)] and their ability to provide better inference in meta-analyses with relatively few studies (Tipton 2015). For these reasons, we often use these robust methods.

Code

Fitting a random effects model requires only a small change to the methods argument of `rma()`. (We also include the `knha` flag that adds a correction to the computation of standard errors and p-values).

```
re_model <- rma(yi = d, vi = var_d, data = paluck, method = "REML", knha = TRUE)
```

Based on the random effects model, intergroup contact effects appear to differ across studies. Paluck et al. estimated that the standard deviation of effects across studies was $\hat{\tau} = 0.44$; 95% confidence interval [0.25, 0.57]. This estimate indicates a substantial amount of heterogeneity! To visualize these results, we can plot the estimated density of the population effects, which is just a normal distribution with mean $\hat{\mu}$ and standard deviation $\hat{\tau}$ (Figure 16.2).

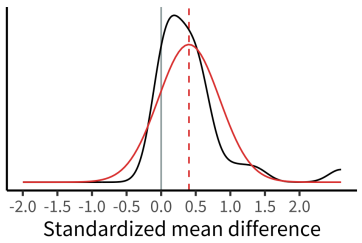


Figure 16.2: Estimated distribution of population effects from random-effects meta-analysis of Paluck et. al's dataset (heavy red curve) and estimated density of studies' point estimates (thin black curve).

This meta-analysis highlights an important point:that the overall effect size estimate $\hat{\mu}$ represents only the *mean* population effect across studies. It tells us nothing about how much the effects *vary* across studies. Thus, we recommend always reporting the heterogeneity estimate $\hat{\tau}$, preferably along with other related metrics that help summarize the distribution of effect sizes across studies (Riley, Higgins, and Deeks 2011; Wang and Lee 2019; Mathur and VanderWeele 2019, 2020a). Reporting the heterogeneity helps readers know how consistent or inconsistent the effects are across studies, which may point to the need to investigate *moderators* of the effect (i.e., factors that are associated with larger or smaller effects, such as whether participants were adults or children).[8]

[8] One common approach to investigating moderators in meta-analysis is meta-regression, in which moderators are included as covariates in a random-effects meta-analysis model (Thompson and Higgins 2002). As in standard regression, coefficients can then be estimated for each moderator, representing the mean difference in population effect between studies with versus without the moderator.

DEPTH

Single-paper meta-analysis and pooled analysis

Thus far, we have described meta-analysis as a tool for summarizing results reported across multiple papers. However, some people have argued that meta-analysis should also be used to summarize the results of multiple studies reported in a single paper (Goh, Hall, and Rosenthal 2016). For instance, in a paper where you describe 3 different experiments on a hypothesis, you could (1) extract summary information (e.g., *M*'s and *SD*'s) from each study, (2) compute the effect size, and then (3) combine the effect sizes in a meta-analysis.

Single-paper meta-analyses come with many of the same strengths and weaknesses we have discussed thus far. One unique weakness, though, is that having a small number of studies means that you typically have low power to detect heterogeneity and moderators. This lack of power sometimes leads researchers to claim that there are no significant differences between their studies. But an alternative explanation is that there simply wasn't enough power to detect those differences!

As an alternative, you can also pool the actual data from the three studies, as opposed to just pooling summary statistics. For example, if you have data from 10 participants in each of the 3 experiments, you could pool them into

a single dataset with 30 participants and include random effects of your condition manipulation across experiments (as described in Chapter 7). This strategy is often referred to as **pooled** or **integrative** data analysis (and occasionally as "mega-analysis", which sounds cool).

Study 1

Study	Participant	Group	Prejudice	Age
1	1	Treatment	2	18
1	2	Treatment	2	18
1	3	Treatment	2	18
1	4	Treatment	2	18
1	5	Treatment	2	18
1	6	Control	10	18
1	7	Control	10	18
1	8	Control	10	18
1	9	Control	10	18
1	10	Control	10	18

Study 2

Study	Participant	Group	Prejudice	Age
2	1	Treatment	5	24
2	2	Treatment	5	24
2	3	Treatment	5	24
2	4	Treatment	5	24
2	5	Treatment	5	24
2	6	Control	10	24
2	7	Control	10	24
2	8	Control	10	24
2	9	Control	10	24
2	10	Control	10	24

Study 3

Study	Participant	Group	Prejudice	Age
3	1	Treatment	9	45
3	2	Treatment	9	45
3	3	Treatment	9	45
3	4	Treatment	9	45
3	5	Treatment	9	45
3	6	Control	10	45
3	7	Control	10	45
3	8	Control	10	45
3	9	Control	10	45
3	10	Control	10	45

Pooled data analysis

Study	Participant	Group	Prejudice	Age
1	1	Treatment	2	18
1	2	Treatment	2	18
1	3	Treatment	2	18
1	4	Treatment	2	18
1	5	Treatment	2	18
1	6	Control	10	18
1	7	Control	10	18
1	8	Control	10	18
1	9	Control	10	18
1	10	Control	10	18
2	1	Treatment	5	24
2	2	Treatment	5	24
2	3	Treatment	5	24
2	4	Treatment	5	24
2	5	Treatment	5	24
2	6	Control	10	24
2	7	Control	10	24
2	8	Control	10	24
2	9	Control	10	24
2	10	Control	10	24
3	1	Treatment	9	45
3	2	Treatment	9	45
3	3	Treatment	9	45
3	4	Treatment	9	45
3	5	Treatment	9	45
3	6	Control	10	45
3	7	Control	10	45
3	8	Control	10	45
3	9	Control	10	45
3	10	Control	10	45

Meta-analysis

Study	Effect size (d)	Age
1	8	18
2	5	24
3	1	45

Figure 16.3: Meta-analysis vs. pooled data analysis.

One of the benefits of pooled data analysis is that it can give you more power to detect moderators. For instance, imagine that the effect of an intergroup contact treatment is moderated by age. If we performed a traditional meta-analysis, we would only have three observations in our data set, yielding very low power. However, we have many more observations (and much more variation in the moderator) in the pooled data analysis, which can lead to higher power (Figure 16.3).

Pooled data analysis is not without its own limitations (Cooper and Patall 2009). And, of course, sometimes it doesn't make as much sense to pool datasets (e.g., when measures are different from one another). Nonetheless, we believe that pooled data analysis and meta-analysis are both useful tools to keep in mind in a paper reporting multiple experiments!

16.2 Bias in meta-analysis

Meta-analysis is a great tool for synthesizing evidence across studies, but the accuracy of a meta-analysis can be compromised by bias. We'll talk

about two categories of bias here: **within-study** and **across-study** biases. Either type can lead to meta-analytic estimates that are too large, too small, or even in the wrong direction altogether.

16.2.1 Within-study biases

Within-study biases – such as demand characteristics, confounds, and order effects, all discussed in Chapter 9 – not only impact the validity of individual studies, but also any attempt to synthesize those studies. And of course, if individual study results are affected by analytic flexibility (p-hacking), meta-analyzing these will result in inflated effect size estimates. In other words: garbage in, garbage out.

For example, Paluck, Green, and Green (2019) noted that early studies on intergroup contact almost exclusively used non-randomized designs. Imagine a hypothetical study where researchers studied a completely ineffective intergroup contact intervention, and non-randomly assigned low-prejudice people to the intergroup contact condition and high-prejudice people to the control condition. In a scenario like this, the researcher would of course find that the prejudice was lower in the intergroup contact condition. But the effect would not be a true contact intervention effect, but rather a spurious effect of non-random assignment (i.e., confounding). Now imagine meta-analyzing many studies with similarly poor designs. The meta-analyst might find impressive evidence of an intergroup contact effect, even if none existed.

To mitigate this problem, meta-analysts often exclude studies that may be especially affected by within-study bias. (For example, Paluck, Green, and Green 2019 excluded non-randomized studies). Of course, these decisions can't be made on the basis of their effects on the meta-analytic estimate or else this post-hoc exclusion itself will lead to bias! For this reason, inclusion and exclusion criteria for meta-analyses should be preregistered whenever possible.

Sometimes certain sources of bias cannot be eliminated by excluding studies – often because studies in a particular domain share certain fundamental limitations (for example, attrition in drug trials). After data have been collected, meta-analysts should also assess studies' risks of bias qualitatively using established rating tools (Sterne et al. 2016). Doing so allows the meta-analyst to communicate how much within-study bias there may be.[9]

Meta-analysts can also conduct sensitivity analyses to assess how much results might be affected by different within-study biases or by excluding certain types of studies (Mathur and VanderWeele 2022). For example, if nonrandom assignment is a concern, a meta-analyst may run the

[9] If you're interested in assessing within-study bias, you can take a look at the Risk of Bias tool (https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool) developed by Cochrane, an organization devoted to evidence synthesis.

analyses including only randomized studies, versus including all studies, in order to determine how much including nonrandomized studies changes the meta-analytic estimate. These two options parallel our discussion of experimental preregistration in Chapter 11: To allay concerns about results-dependent meta-analysis, researchers can either preregister their analyses ahead of time or else be transparent about their choices after the fact. Sensitivity analyses can allay concerns that a specific choice of exclusion criteria is critically related to the reported results.

16.2.2 Across-study biases

Across-study biases occur if, for example, researchers **selectively report** certain types of findings or selectively publish certain types of findings (publication bias, as discussed in Chapter 3 and Chapter 11). Often, these across-study biases favor statistically-significant positive results, which means the meta-analytic estimate based on those studies will be inflated relative to the true effect.

ACCIDENT REPORT

Quantifying publication bias in the social sciences

It's typically very hard to quantify publication bias because you don't know how many studies are out there in researchers' "file drawers" – unpublished studies are by definition not available. But a recent study took advantage of a unique opportunity to try and quantify publication bias within a known pool of studies.

Time-sharing Experiments in the Social Sciences (TESS) is an innovative project that lets researchers apply to run experiments on nationally-representative samples in the U.S. In 2014, Franco, Malhotra, and Simonovits (2014) and colleagues took advantage of this application process by examining the entire population of 221 studies conducted through TESS.

Using this information, Franco and colleagues examined the records of these studies to determine whether the researchers found statistically significant results, a mixture of statistically significant and non-significant results, or only non-significant results. Then, they examined the likelihood that these results were published in the scientific literature.

Over 60% of studies with statistically significant results were published, compared to a mere 25% of studies that produced only statistically non-significant results. This finding was important because it quantified how strong publication bias actually was, at least in one particular population of studies. This estimate may not be general: for example, perhaps TESS studies were easier to put in the file drawer because they cost nothing for the researchers to run. But even a lower level of publication bias can have a substantial effect on a meta-analysis, meaning that it is crucial to check for – and potentially, correct for – publication bias.

Like within-study biases, meta-analysts often try to mitigate across-study biases by being careful about what studies make it into the meta-analysis. Meta-analysts don't only want to capture high-profile,

published studies on their effect of interest, but also studies published in low-profile journals and the so-called "gray literature" [i.e., unpublished dissertations and theses; Lefebvre et al. (2019)].[10]

There are also statistical methods to help assess how robust the results may be to across-study biases. Among the most popular tools to assess and correct for publication bias is the **funnel plot** (Duval and Tweedie 2000; Egger et al. 1997). A funnel plot shows the relationship between studies' effect estimates and their precision (usually their standard error). These plots are called "funnel plots" because if there is no publication bias, then as precision increases, the effects "funnel" towards the meta-analytic estimate. As the precision is smaller, they spread out more because of greater measurement error. Figure 16.4 is an example of one type of funnel plot (Mathur and VanderWeele 2020b) for a simulated meta-analysis of 100 studies with no publication bias.

[10] Evidence is mixed regarding whether including gray literature actually reduces across-study biases in meta-analysis (Tsuji et al. 2020; Mathur and VanderWeele 2021), but it is still common practice to try to include this literature.

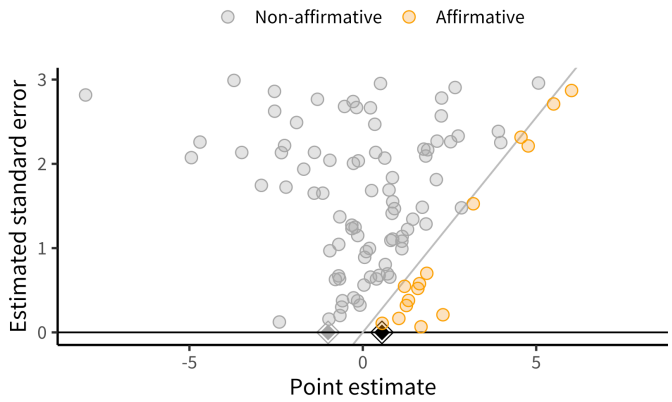


Figure 16.4: Significance funnel plot for a meta-analysis simulated to have no publication bias. Orange points: studies with $p < 0.05$ and positive estimates. Grey points: studies with $p \geq 0.05$ or negative estimates. Black diamond: random-effects estimate of $\widehat{\mu}$.

CODE

For this plot, we use the `PublicationBias` package and the `significance_funnel()` function. (An alternative function is the `metafor` function `funnel()`, which results in a more "classic" funnel plot.) We use our fitted model `re_model`:

```
significance_funnel(yi = re_model$yi, vi = re_model$vi)
```

Because meta-analysis is such a well-established method, many of the relevant operations are "plug and play."

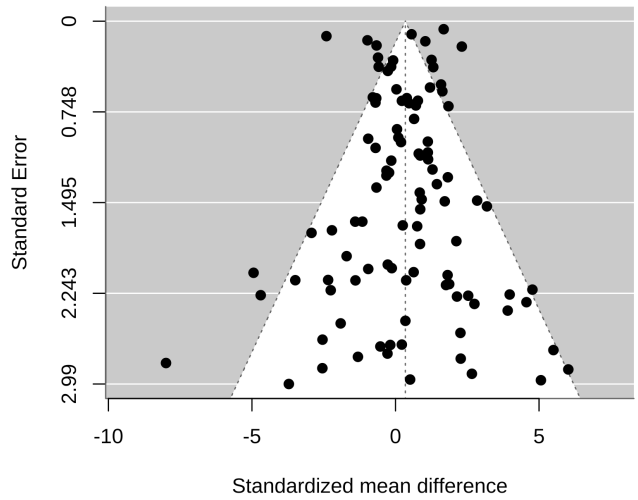


Figure 16.5: Classic funnel plot.

As implied by the "funnel" moniker, our plot looks a little like a funnel. Larger studies (those with smaller standard errors) cluster more closely around the mean of 0.34 than do smaller studies, but large and small studies alike have point estimates centered around the mean. That is, the funnel plot is symmetric.[11]

Not all funnel plots are symmetric! Figure 16.6 is what happens to our hypothetical meta-analysis if all studies with $p < 0.05$ and positive estimates are published, but only 10% of studies with $p \geq 0.05$ or with negative estimates are published. The introduction of publication bias dramatically inflates the pooled estimate from 0.34 to 1.15. Also, there appears to be a correlation between studies' estimates and their standard errors, such that smaller studies tend to have larger estimates than do larger studies. This correlation is often called **funnel plot asymmetry** because the funnel plot starts to look like a right triangle rather than a funnel. Funnel plot asymmetry *can* be a diagnostic for publication bias, though it isn't always a perfect indicator, as we'll see in the next subsection.

[11] Classic funnel plots look more like Figure 16.5). Our version is different in a couple of ways. Most prominently, we don't have the vertical axis reversed (which we think is confusing). We also don't have the left boundary highlighted, because we think folks don't typically select for negative studies.

16.2.1 Across-study bias correction

How do we identify and correct bias across studies? Given that some forms of publication bias induce a correlation between studies' point estimates and their standard errors, several popular statistical methods, such as Trim-and-Fill (Duval and Tweedie 2000) and Egger's regression (Egger et al. 1997) are designed to quantify funnel plot asymmetry.

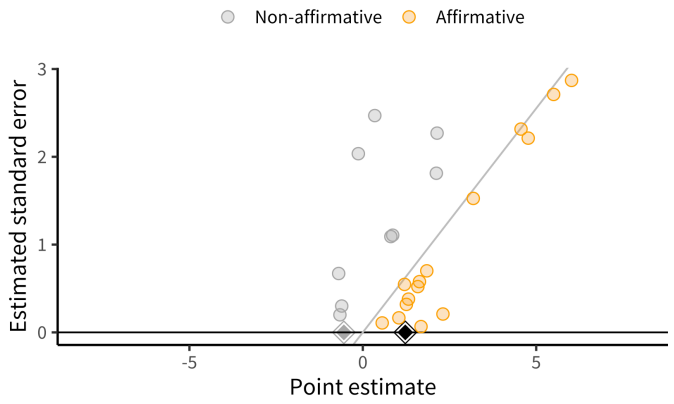


Figure 16.6: Significance funnel plot for the same simulated meta-analysis after publication bias has occurred. Orange points: studies with $p < 0.05$ and positive estimates. Grey points: studies with $p \geq 0.05$ or negative estimates. Black diamond: random-effects estimate of $\widehat{\mu}$.

Funnel plot asymmetry does not always imply that there is publication bias, though. Nor does publication bias always cause funnel plot asymmetry. Sometimes funnel plot asymmetry is driven by genuine differences in the effects being studied in small and large studies (Egger et al. 1997; Lau et al. 2006). For example, in a meta-analysis of intervention studies, if the most effective interventions are also the most expensive or difficult to implement, these highly effective interventions might be used primarily in the smallest studies ("small study effects").

Funnel plots and related methods are best suited to detecting publication bias in which (1) small studies with large positive point estimates are more likely to be published than small studies with small or negative point estimates; and (2) the largest studies are published regardless of the magnitude of their point estimates. That model of publication bias is sometimes what is happening, but not always!

A more flexible approach for detecting publication bias uses **selection models**. These models can detect other forms of publication bias that funnel plots may not detect, such as publication bias that favors *significant* results. We won't cover these methods in detail here, but we think they are a better approach to the question, along with related sensitivity analyses.[12]

You may also have heard of "p-methods" to detect across-study biases such as p-curve and p-uniform (Simonsohn, Nelson, and Simmons 2014; Van Assen, Aert, and Wicherts 2015). These methods essentially assess whether the significant p-values "bunch up" just under 0.05, which is taken to indicate publication bias. These methods are increasingly popular in psychology and have their merits. However, they are actually simplified versions of selection models (e.g., Hedges 1984) that work only under considerably more restrictive settings than the original selection models [for example, when there is not heterogeneity across

[12] High-level overviews of selection models are given in McShane, Böckenholt, and Hansen (2016) and Maier, VanderWeele, and Mathur (2022). For more methodological detail, see Hedges (1984), Iyengar and Greenhouse (1988), and Vevea and Hedges (1995). For a tutorial on fitting and interpreting selection models, see Maier, VanderWeele, and Mathur (2022). For sensitivity analyses, see Mathur and VanderWeele (2020b).

studies; McShane, Böckenholt, and Hansen (2016)]. For this reason, it is usually (although not always) better to use selection models in place of the more restrictive p-methods.

Going back to our running example, Paluck et al. used a regression-based approach to assess and correct for publication bias. This approach provided significant evidence of a relationship between the standard error and effect size (i.e., an asymmetric funnel plot). Again, this asymmetry could reflect publication bias or other sources of correlation between studies' estimates and their standard errors. Paluck et al. also used this same regression-based approach to try to correct for potential publication bias. Results from this model indicated that the bias-corrected effect size estimate was close to zero. In other words, even though all studies estimated that intergroup contact decreased prejudice, it is still possible that there are unpublished studies that did not find this (or found that intergroup contact increased prejudice).

ACCIDENT REPORT

Garbage in, garbage out? Meta-analyzing potentially problematic research

Botox can help eliminate wrinkles. But some researchers have suggested that, when used to paralyze the muscles associated with frowning, Botox may also help treat clinical depression. As surprising as this claim may sound, a quick examination of the literature would lead many to conclude that this treatment works. Studies that randomly assign depressed patients to receive either Botox or saline injections do indeed find that Botox recipients show decreased depression. And when you combine all available evidence in a meta-analysis, you find that this effect is quite large: $d = 0.83$, 95% CI [0.52, 1.14].

As Coles et al. (2019) argued though, this estimated effect may be impacted by both within- and between-study bias. First, participants are not supposed to know whether they have been randomly assigned to receive Botox or a control saline injections. But only one of these treatments leads the upper half of your face to be paralyzed! After a couple weeks, you're pretty likely to know whether you received the Botox treatment or control saline injection. Thus, the apparent effect of Botox on depression could instead be a placebo effect.

Second, only 50% of the outcomes that researchers measured were reported in the final publications, raising concerns about selective reporting. Perhaps researchers examining the effects of Botox on depression only reported the measures that showed a positive effect, not the ones that didn't.

Taken together, these two criticisms suggest that, despite the impressive meta-analytic estimate, the effect of Botox on depression is far from certain.

16.3 Chapter summary: Meta-analysis

Taken together, Paluck and colleagues' use of meta-analysis provided several important insights that would have been easy to miss in a non-quantitative review. First, despite a preponderance of non-significant findings, intergroup contact interventions were estimated to decrease

prejudice by on average 0.4 standard deviations. On the other hand, there was considerable heterogeneity in intergroup contact effects, suggesting important moderators of the effectiveness of these interventions. And finally, publication bias was a substantial concern, indicating a need for follow-up research using a registered report format that will be published regardless of whether the outcome is positive (Chapter 11).

Overall, meta-analysis is a key technique for aggregating evidence across studies. Meta-analysis allows researchers to move beyond the bias of naive techniques like vote counting and towards a more quantitative summary of an experimental effect. Unfortunately, a meta-analysis is only as good as the literature it's based on, so the aspiring meta-analyst must be aware of both within- and between-study biases!

DISCUSSION QUESTIONS

1. Imagine that you read the following result in the abstract of a meta-analysis: "In 83 randomized studies of middle school children, replacing one hour of class time with mindfulness meditation significantly improved standardized test scores (standardized mean difference $\hat{\mu} = 0.05$; 95% confidence interval: $[0.01, 0.09]$; $p < 0.05$)." Why is this a problematic way to report on meta-analysis results? Suggest a better sentence to replace this one.
2. As you read the rest of the meta-analysis, you find that the authors conclude that "These findings demonstrate robust benefits of meditation for children, suggesting that test scores improve even when the meditation is introduced as a replacement for normal class time." You recall that the heterogeneity estimate was $\hat{\tau} = 0.90$. Do you think that this result regarding the heterogeneity tends to support, or rather tends to undermine, the concluding sentence of the meta-analysis? Why?
3. What kinds of within-study biases would concern you in the meta-analysis described in the prior two questions? How might you assess the credibility of the meta-analyzed studies and of the meta-analysis as whole in light of these possible biases?
4. Imagine you conduct a meta-analysis on a literature in which statistically significant results in either direction are much more likely to be published that non-significant results. Draw the funnel plot you would expect to see. Is the plot symmetric or asymmetric?
5. Why do you think small studies receive more weight in random-effects meta-analysis than in fixed-effects meta-analysis? Can you see why this is true mathematically based on the equations given above, and can you also explain the intuition in simple language?

READINGS

– A nice, free textbook with lots of good code examples: Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. (2022). Doing Meta-Analysis with R: A Hands-On Guide. Chapman & Hall/CRC Press. Available free online at https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/.

References

Allport, Gordon Willard, Kenneth Clark, and Thomas Pettigrew. 1954. *The Nature of Prejudice*. Addison-Wesley.

Boisjoly, Johanne, Greg J Duncan, Michael Kremer, Dan M Levy, and Jacque Eccles. 2006. "Empathy or Antipathy? The Impact of Diversity." *American Economic Review* 96 (5): 1890–1905.

Borenstein, Michael, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. 2021. *Introduction to Meta-Analysis*. John Wiley & Sons.

Brockwell, Sarah E, and Ian R Gordon. 2001. "A Comparison of Statistical Methods for Meta-Analysis." *Statistics in Medicine* 20 (6): 825–40.

Clunies-Ross, Graham, and Kris O'Meara. 1989. "Changing the Attitudes of Students Towards Peers with Disabilities." *Australian Psychologist* 24 (2): 273–84.

Coles, Nicholas A, Jeff T Larsen, Joyce Kuribayashi, and Ashley Kuelz. 2019. "Does Blocking Facial Feedback via Botulinum Toxin Injections Decrease Depression? A Critical Review and Meta-Analysis." *Emotion Review* 11 (4): 294–309.

Cooper, Harris, and Erika A Patall. 2009. "The Relative Benefits of Meta-Analysis Conducted with Individual Participant Data Versus Aggregated Data." *Psychological Methods* 14 (2): 165.

DerSimonian, Rebecca, and Nan Laird. 1986. "Meta-Analysis in Clinical Trials." *Controlled Clinical Trials* 7 (3): 177–88.

Duval, Sue, and Richard Tweedie. 2000. "Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis." *Biometrics* 56 (2): 455–63. https://doi.org/10.1111/j.0006-341X.2000.00455.x.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. "Bias in Meta-Analysis Detected by a Simple, Graphical Test." *British Medical Journal* 315 (7109): 629–34. https://doi.org/10.1136/bmj.315.7109.629.

Franco, A, N Malhotra, and G Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science*.

Goh, Jin X, Judith A Hall, and Robert Rosenthal. 2016. "Mini Meta-Analysis of Your Own Studies: Some Arguments on Why and a Primer on How." *Social and Personality Psychology Compass* 10 (10): 535–49.

Goldstein, Noah J, Robert B Cialdini, and Vladas Griskevicius. 2008. "A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels." *Journal of Consumer Research* 35 (3): 472–82.

Grant, Maria J, and Andrew Booth. 2009. "A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies." *Health Information & Libraries Journal* 26 (2): 91–108.

Hedges, Larry V. 1984. "Estimation of Effect Size Under Nonrandom Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences." *Journal of Educational Statistics* 9 (1): 61–85. https://doi.org/10.3102/10769986009001061.

Hedges, Larry V, Elizabeth Tipton, and Matthew C Johnson. 2010. "Robust Variance Estimation in Meta-Regression with Dependent Effect Size Estimates." *Research Synthesis Methods* 1 (1): 39–65.

Iyengar, Satish, and Joel B Greenhouse. 1988. "Selection Models and the File Drawer Problem." *Statistical Science*, 109–17.

Lau, Joseph, John PA Ioannidis, Norma Terrin, Christopher H Schmid, and Ingram Olkin. 2006. "The Case of the Misleading Funnel Plot." *British Medical Journal* 333 (7568): 597–600. https://doi.org/10.1136/bmj.333.7568.597.

Lefebvre, Carol, Julie Glanville, Simon Briscoe, Anne Littlewood, Chris Marshall, Maria-Inti Metzendorf, Anna Noel-Storr, et al. 2019. "Searching for and Selecting Studies." In *Cochrane Handbook for Systematic Reviews of Interventions*, edited by J. Chandler J. P. T. Higgins J. Thomas and V. A. Welch, 67–107. Wiley-Blackwell. https://doi.org/https://doi.org/10.1002/9781119536604.ch4.

Maier, Maximilian, Tyler J VanderWeele, and Maya B Mathur. 2022. "Using Selection Models to Assess Sensitivity to Publication Bias: A Tutorial and Call for More Routine Use." *Campbell Systematic Reviews* 18 (3): e1256.

Mathur, Maya B, and Tyler J VanderWeele. 2019. "New Metrics for Meta-Analyses of Heterogeneous Effects." *Statistics in Medicine* 38 (8): 1336–42.

———. 2020a. "Robust Metrics and Sensitivity Analyses for Meta-Analyses of Heterogeneous Effects." *Epidemiology* 31 (3): 356–58.

———. 2020b. "Sensitivity Analysis for Publication Bias in Meta-Analyses." *Journal of the Royal Statistical Society: Series C* 5 (69): 1091–1119.

———. 2021. "Estimating Publication Bias in Meta-Analyses of Peer-Reviewed Studies: A Meta-Meta-Analysis Across Disciplines and Journal Tiers." *Research Synthesis Methods* 12 (2): 176–91.

———. 2022. "Methods to Address Confounding and Other Biases in Meta-Analyses: Review and Recommendations." *Annual Review of Public Health* 1 (43).

McShane, Blakeley B, Ulf Böckenholt, and Karsten T Hansen. 2016. "Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes." *Perspectives on Psychological Science* 11 (5): 730–49. https://doi.org/10.1177/1745691616662243.

McShane, Blakeley B, and David Gal. 2017. "Statistical Significance and the Dichotomization of Evidence." *Journal of the American Statistical Association* 112 (519): 885–95. https://doi.org/10.1080/01621459.2017.1289846.

Nelson, Nanette, Robert Rosenthal, and Ralph L Rosnow. 1986. "Interpretation of Significance Levels and Effect Sizes by Psychological Researchers." *American Psychologist* 41 (11): 1299.

Paluck, Elizabeth Levy, Seth A Green, and Donald P Green. 2019. "The Contact Hypothesis Re-Evaluated." *Behavioural Public Policy* 3 (2): 129–58.

Pustejovsky, James E, and Elizabeth Tipton. 2021. "Meta-Analysis with Robust Variance Estimation: Expanding the Range of Working Models." *Prevention Science*, 1–14.

Riley, Richard D, Julian PT Higgins, and Jonathan J Deeks. 2011. "Interpretation of Random Effects Meta-Analyses." *British Medical Journal* 342.

Scheibehenne, Benjamin, Tahira Jamil, and Eric-Jan Wagenmakers. 2016. "Bayesian Evidence Synthesis Can Reconcile Seemingly Inconsistent Results: The Case of Hotel Towel Reuse." *Psychol. Sci.* 27 (7): 1043–46.

Simonsohn, Uri, Leif D Nelson, and Joseph P Simmons. 2014. "P-Curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General* 143 (2): 534.

Sterne, Jonathan AC, Miguel A Hernán, Barnaby C Reeves, Jelena Savović, Nancy D Berkman, Meera Viswanathan, David Henry, et al. 2016. "ROBINS-i: A Tool for Assessing Risk of Bias in Non-Randomised Studies of Interventions." *British Medical Journal* 355.

Thompson, Simon G, and Julian PT Higgins. 2002. "How Should Meta-Regression Analyses Be Undertaken and Interpreted?" *Statistics in Medicine* 21 (11): 1559–73.

Tipton, Elizabeth. 2015. "Small Sample Adjustments for Robust Variance Estimation with Meta-Regression." *Psychological Methods* 20 (3): 375.

Tsuji, Sho, Alejandrina Cristia, Michael C Frank, and Christina Bergmann. 2020. "Addressing Publication Bias in Meta-Analysis." *Zeitschrift für Psychologie*.

Van Assen, Marcel ALM, Robbie van Aert, and Jelte M Wicherts. 2015. "Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies." *Psychological Methods* 20 (3): 293.

Vevea, Jack L, and Larry V Hedges. 1995. "A General Linear Model for Estimating Effect Size in the Presence of Publication Bias." *Psychometrika* 60 (3): 419–35.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in r with the Metafor Package." *Journal of Statistical Software* 36 (3): 1–48.

Wang, Chia-Chun, and Wen-Chung Lee. 2019. "A Simple Method to Estimate Prediction Intervals and Predictive Distributions: Summarizing Meta-Analyses Beyond Means and Confidence Intervals." *Research Synthesis Methods* 10 (2): 255–66.

17 Conclusion

You've made it to the end of Experimentology, our (sometimes opinionated) guide to how to run good psychology experiments. In this book we've tried to present a unified approach to the why and how of running experiments. This approach begins with the goal of doing experiments:

> Experiments are intended to make maximally unbiased, generalizable, and precise estimates of specific causal effects.

This formulation isn't exactly how experiments are talked about in the broader field, but we hope you've started to see some of the rationale behind this approach. In this final chapter, we will briefly discuss some aspects of our approach, as well how this approach connects with our four themes, TRANSPARENCY, MEASUREMENT PRECISION, BIAS REDUCTION, and GENERALIZABILITY. We'll end by mentioning some exciting new trends in the field that give us hope about the future of experimentology and psychology more broadly.

17.1 Summarizing our approach

The Experimentology approach is grounded in both an appreciation of the power of experiments to reveal important aspects about human psychology and also an understanding of the many ways that experiments can fail. In particular, the "replication crisis" (Chapter 3) has revealed that small samples, a focus on dichotomous statistical inference, and a lack of transparency around data analysis can lead to a literature that is often neither reproducible nor replicable. Our approach is designed to avoid these pitfalls.

We focus on MEASUREMENT PRECISION in service of measuring causal effects. The emphasis on causal effects stems from an acknowledgement of the key role of experiments in establishing causal inferences (Chapter 1) and the importance of causal relationships to theories (Chapter 2). In our statistical approach, we focus on estimation (Chapter 5) and modeling (Chapter 7), helping us to avoid some of the fallacies that come

along with dichotomous inference (Chapter 6). We choose measures to maximize reliability (Chapter 8). We prefer simple, within-participant experimental designs because they typically result in more precise estimates (Chapter 9). And we think meta-analytically about the overall evidence for a particular effect beyond our individual experiment (Chapter 16).

Further, we recognize the presence of many potential sources of bias in our estimates, leading us to focus on BIAS REDUCTION. In our measurements, we identify arguments for the validity of our measures, decreasing bias in estimation of the key constructs of interest (Chapter 8); in our designs we seek to minimize bias due to confounding or experimenter effects (Chapter 9). We also try to minimize the possibility of bias in our decisions about data collection (Chapter 12) and data analysis (Chapter 11). Finally, we recognize the possibility of bias in literatures as a whole and consider ways to compensate in our estimates (Chapter 16).

Finally, we consider GENERALIZABILITY throughout the process. We theorize with respect to a particular population (Chapter 2) and select our sample in order to maximize the generalizability of our findings to that target population (Chapter 10). In our statistical analysis, we take into account multiple dimensions of generalizability, including across participants and experimental stimulus items (Chapter 7). And in our reporting, we contextualize our findings with respect to limits on their generalizability (Chapter 14).

Woven throughout this narrative is the hope that embracing TRANSPARENCY throughout the experimental process will help you maximize your work. Not only is sharing your work openly an ethical responsibility (Chapter 4), it's also a great way to minimize errors while creating valuable products that both advance scientific progress and accelerate your own career (Chapter 13).

17.2 Forward the field

We have focused especially on reproducibility and replicability issues, but we've learned at various points in this book that there's a replication crisis (Open Science Collaboration 2015), a theory crisis (Oberauer and Lewandowsky 2019), and a generalizability crisis (Yarkoni 2020) in psychology. Based on all these crises, you might think that we are pessimistic about the future of psychology. Not so.

There have been tremendous changes in psychological methods since we started teaching Experimental Methods in 2012. When we began,

it was common for incoming graduate students to describe the rampant p-hacking they had been encouraged to do in their undergraduate labs. Now, students join the class aware of new practices like preregistration and cognizant of problems of generalizability and theory building. It takes a long time for a field to change, but we have seen tremendous progress at every level – from government policies requiring transparency in the sciences all the way down to individual researchers' adoption of tools and practices that increase the efficiency of their work and decrease the chances of error.

One of the most exciting trends has been the rise of meta-science, in which researchers use the tools of science to understand how to make science better (Tom E. Hardwicke et al. 2020). Reproducibility and replicability projects (reviewed in Chapter 3) can help us measure the robustness of the scientific literature. In addition, studies that evaluate the impacts of new policies (e.g., Tom E. Hardwicke et al. 2018) – can help stakeholders like journal editors and funders make informed choices about how to push the field towards more robust science.

In addition to changes that correct methodological issues, the last ten years have seen the rise of "big team science" efforts that advance the field in new ways (Coles et al. 2022). Collaborations such as the Psychological Science Accelerator (Moshontz et al. 2018) and ManyBabies (Frank et al. 2017) allow hundreds of researchers from around the world to come together to run shared projects. These projects are enabled by open science practices like data and code sharing, and they provide a way for researchers to learn best practices via participating. In addition, by including broader and more diverse samples they can help address challenges around generalizability (Klein et al. 2018).

Finally, the last ten years have seen huge progress in the use of statistical models both for understanding data (McElreath 2018) and for describing specific psychological mechanisms (Ma, Körding, and Goldreich 2022). In our own work we have used these models extensively and we believe that they provide an exciting toolkit for building quantitative theories that allow us to explain and to predict the human mind.

17.3 Final thoughts

Doing experiments is a craft, one that requires practice and attention. The first experiment you run will have limitations and issues. So will the 100th. But as you refine your skills, the quality of the studies you design will get better. Further, your own ability to judge others' experiments will improve as well, making you a more discerning consumer of empirical results. We hope you enjoy this journey!

References

Coles, Nicholas A, J Kiley Hamlin, Lauren L Sullivan, Timothy H Parker, and Drew Altschul. 2022. "Build up Big-Team Science." Nature Publishing Group.

Frank, Michael C, Elika Bergelson, Christina Bergmann, Alejandrina Cristia, Caroline Floccia, Judit Gervain, J Kiley Hamlin, et al. 2017. "A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building." *Infancy* 22 (4): 421–35.

Hardwicke, Tom E, Maya B Mathur, Kyle Earl MacDonald, Gustav Nilsonne, George Christopher Banks, Mallory Kidwell, Alicia Hofelich Mohr, et al. 2018. "Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the Journal Cognition."

Hardwicke, Tom E., Stylianos Serghiou, Perrine Janiaud, Valentin Danchev, Sophia Crüwell, Steven N. Goodman, and John P. A. Ioannidis. 2020. "Calibrating the Scientific Ecosystem Through Meta-Research." *Annual Review of Statistics and Its Application* 7 (1): 11–37. https://doi.org/10.1146/annurev-statistics-031219-041104.

Klein, Richard A, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, et al. 2018. "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1 (4): 443–90.

Ma, Wei Ji, K Körding, and Daniel Goldreich. 2022. *Bayesian Models of Perception and Action: An Introduction*. unpublished.

McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in r and Stan*. Chapman; Hall/CRC.

Moshontz, Hannah, Lorne Campbell, Charles R Ebersole, Hans IJzerman, Heather L Urry, Patrick S Forscher, Jon E Grahe, et al. 2018. "The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network." *Advances in Methods and Practices in Psychological Science* 1 (4): 501–15.

Oberauer, Klaus, and Stephan Lewandowsky. 2019. "Addressing the Theory Crisis in Psychology." *Psychonomic Bulletin & Review* 26 (5): 1596–1618.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251).

Yarkoni, Tal. 2020. "The Generalizability Crisis." *Behavioral and Brain Sciences* 45: 1–37.

A Instructor's guide

A.1 Introduction

This is an instructor's guide to conducting replication projects in courses. In addition to benefiting the field in ways that have been previously discussed by some of the authors of this book (e.g., Hawkins et al. (2018), Frank and Saxe (2012)), replication-based courses can additionally benefit students in these courses. In this guide, we will describe these benefits, explore different ways in which courses may be modified depending on student level and resources, and provide some guidelines and examples to help you set up the logistics of your course.

A.2 Why Teach a Project-Based Course?

Over the years, we have observed many ways in which our replication-based courses benefited students above and beyond a more traditional lecture and problem set-based course. Some of these benefits include:

- **Student interest**: Since each student will be free to replicate a study that is aligned with their research interests, this freedom facilitates a more direct application of course methods and lessons to a project that is interesting to each student.
- **Usefulness**: If this course is taught in the first year of the program (as recommended), students may use their replication project as a way to establish robustness of a phenomenon before building studies on top of it.
- **Realism**: Practice datasets that are typically provided for course exercises lack the complexity and messiness of real data. By conducting a replication project and dealing with real data, students learn to apply the tools provided in the course in a way that more closely demonstrates their usefulness beyond the course.
- **Intuition**: Presentations of replication outcomes across the class along with a discussion of what factors seemed to predict these outcomes helps students develop a better intuition when reading the literature for how likely studies are to replicate.

- **Perspective**: Frustrating experiences with ambiguity (whether regarding experimental methods, materials, or analyses) can motivate students to adopt best practices for their own future studies.

A project-based course may look very different depending on student level (undergraduate vs. graduate/post-doc level) and availability of resources at your institution for a course like this, namely in terms of TA support and course funding (for data collection). For most of this guide, we will assume that you have a similar setup to ours (i.e., teaching at the graduate/post-doc level and have course funding and TAs to support the course), but we have also spent some time considering ways to adjust the course to fit different student levels and availability of resources (see "Scenarios for different course layouts").

A.3 Logistics

A.3.1 Syllabus considerations

If it is your first time teaching this course, you may want to decide ahead of time whether your course will mainly focus on content, or whether you will cover *both* content and relevant practical skills. For instance, if the course is for undergraduate students, you may decide to focus mainly on content, whereas if the course is for graduate students, they may find it more useful if the course covers both content and practical skills they can use in their research.

Another important consideration is how long your course will be. Depending on whether your university operates on quarters or semesters, the pace of the course will differ. For Psych 251, since we are on the quarter system, we use the 10-week schedule shown below. However, we have also adapted this schedule to a 16-week course given that it better represents a majority of other institutions' academic calendars. At the end of this chapter, we give a set of sample class schedules.

A.3.2 Grading

Depending on your course format and teaching philosophy, you may have preferred grading criteria. As a point of reference, in Psych 251, we wanted to encompass both the assignments (problemsets and project components) as well as actual course attendance and participation. In addition, because the replication project is a central part of the course, we weighted the project components slightly more than the problem sets:

- 40%: Problem sets (four, at 10% each)

- 50%: Final project components, including presentations, data collection, analysis, and writeup
- 10%: Attendance and participation in class

A.3.3 Course budget

For our course, we usually receive around US$1,000 for course funding from the Psychology Department. In addition, when students from other departments are enrolled, we have been lucky to receive additional funding from those departments as well, to further support the course. Still, making sure that the course funds cover all students' projects is one of the most challenging parts of the course. Assuming you have a budget to work with, here are some lessons we've learned along the way regarding budgeting (and if you don't have any funding, please refer to the section titled "Course Funding" under "Scenarios for different course layouts"):

- Before students pick their study to replicate, provide them with an estimate of how many participant hours they will be able to receive for their project
- As soon as students pick a study for their replication project, help each student run a power analysis to confirm that replicating the study would be within the budget (TAs can help with this)
- If a student feels strongly about a study that does not fit within the budget, consider the following ways to adjust the study: 1) can the study be made shorter by cutting out unnecessary measures? 2) if it is a multi-trial study, can the number of trials be reduced? 3) would their advisors be willing to provide additional funding? 4) can the study be run on university participant pools?
- As mentioned above, if there are students from other departments who are enrolled in your course, one possibility to obtain more funding is to reach out to the heads of those departments to see whether they would be willing to help support your course.

Once all projects have been approved as within-budget, we encourage you to create a shared spreadsheet containing each student's name, so that they can fill in the details of their replication project. Ultimately, this will help ensure that students are paying fair wages to their participants and keep track of how the course funds are being divided up.

A.3.4 Course-related Institutional Review Board application

While it may be possible to apply for individual IRB approval for each student's project, we recommend applying for course-wide standard

IRB approval for all replication projects that are conducted in your class. Contacting your review board early in the planning stages of the course should clarify what options you have available.

One important thing to remember when students run their individual projects is that they should have the course-wide consent form at the beginning of their studies (and TAs should check this when they review the paradigms). For reference, this is the consent form that each student is required to post at the beginning of their study:

"By answering the following questions, you are participating in a study being performed by cognitive scientists in the Stanford Department of Psychology. If you have questions about this research, please contact us at stanfordpsych251@gmail.com. You must be at least 18 years old to participate. Your participation in this research is voluntary. You may decline to answer any or all of the following questions. You may decline further participation, at any time, without adverse consequences. Your anonymity is assured; the researchers who have requested your participation will not receive any personal information about you."

A.4 Scenarios for different course layouts

Now that we have covered the standard format of the course, we want to now turn our attention to ways in which this format can be tweaked in order to fit different needs and resources. We have organized this section into two main categories: student level, and course resources (such as TAs and course funding).

A.4.1 Student level

While Psych 251 at Stanford is geared towards graduate students (and is currently a required class for entering first-year graduate students in the Psychology Department), we also accept advanced undergraduate students as well as graduate students from other departments (e.g., Education, Human-Computer Interaction, Philosophy, Computer Science). On the first day of our course, we tell students that they should be comfortable with two of the three following topics:

1) Some knowledge of psychological experimentation & subject matter
2) Statistical programming: things like functions and variables
3) Basic statistics like ANOVA and t-test

If students are only comfortable with one of the three topics above, we warn them ahead of time that the course may demand more time from them than the average student.

Now, if you are planning on catering this course for undergraduate students, chances are that they have had less exposure to these topics overall, so there are multiple ways to calibrate the course accordingly:

1) **Prerequisites**: Require students to have completed courses that cover at least two of the three topics mentioned above (i.e., a psychology class, a class that covers statistical programming, a class that covers basic statistics, any two of the three).
2) **Pace**: unlike Psych 251, where the entire course only lasts 10 weeks, a class for undergraduates may benefit from a slower pace, allowing more time to cover the foundational principles before diving into the project. For instance, the course could be held over multiple academic semesters/quarters, with the project goal of Course #1 being choosing and planning the replication study, and the project goal of Course #2 being the execution and interpretation of the replication.
3) **Pair-Group-Based Projects**: In our course, each student is required to conduct their own replication project. However, this structure may be overwhelming for undergraduate students who may have less confidence taking on an entire replication project by themselves. One option that may alleviate this pressure is to have students conduct these projects as pairs or as small teams, so that they can collectively draw on each others' strengths. When assigning these pairs or teams, it may be especially helpful to try to ensure a relatively even balance of students who are confident in each of the three areas outlined above (psychology, statistical programming, basic statistics).

Now that we've offered a few suggestions to address different student levels, let's dive into the issue of course resources.

A.4.2 Course resources

We think there are two main ways in which your course may have different resources from our model: In terms of course assistance (i.e., teaching assistants), and in terms of course funding for student projects. We'll explore ways to work around each of these in this section:

Teaching assistants

As a point of comparison, in general, 2-3 teaching assistants are allocated to Psych 251, which enrolls about 36 students, which comes out to about 12-18 students per TA. Since a project-based course requires individual attention and feedback, we would recommend against a student-TA ratio that is much higher than that. That means that if you know you will have just one TA for the class, you should think about reducing the enrollment cap accordingly. But what if you have *no* TAs? With some adjustments, there are still ways you can make the course work sans-TA; we outline a few ideas below:

1) **Peer grading**: As an instructor with no TAs, the area that will require the biggest lift in terms of time and attention is grading. One way to overcome this is to introduce a peer-grading system, in which students grade each others' work. If you choose this route, two things that may encourage fair grading among your students is to 1) distribute a clear and specific rubric that reduces the amount of subjectivity in the grading process as much as possible, and 2) anonymize the assignments so that students do not know whose assignment they are grading. If possible, it may again be beneficial to assign grading pairs that consist of students that are relatively knowledgeable in different areas, so that they can provide feedback that address weak points in each others' work.

2) **Collective troubleshooting**: The second most time intensive area you will have to make up for is the amount of troubleshooting you may have to do for students who run into issues implement their projects, anywhere from getting GitHub and RMarkdown up and running on their devices, to trouble with data collection on Mechanical Turk. One way to encourage communal support among your students is to set up a central discussion board for the course (e.g., Piazza or a course channel on Slack) where students can publicly (but anonymously, if desired) post issues they are running into. Then, you can offer extra credit to students who help troubleshoot these issues, in order to further incentivize collective troubleshooting. There will likely still be issues that cannot be addressed by the students, but this system at least frees up your time to focus your attention on those that only *you* can address.

3) **Single class-wide project**: Finally, if the collective grading and troubleshooting methods outlined above do not cut down on enough time, you could consider walking through a single replication project as a class.[1] To make a single-project course work, you could have students nominate studies they would like to replicate as a class, and then have them vote on the final choice. Once the target study has been selected, every student

[1] This approach does cancel out some of the benefits of a project-based course we mentioned at the start – namely, the project will likely no longer fit each student's specific research interest, so there may be less benefit in terms of specific student interest and usefulness for their program of research, but the other two benefits of realism and intuition (especially if the project is discussed in the context of other replication findings) still stand.

can individually carry out all the steps of the project, including preregistering and writing up the analysis script. Then, setting up and running the data collection phase can happen during class, and once data has been collected, you can distribute it to the students for them to run it through their analysis script and interpret the result. Whether you choose to have students grade each others' work or whether you grade their work yourself, the fact that the project is standardized should cut down on a lot of the time you would otherwise spend learning about the details of every individual project.

Course funding

In addition to availability of TAs, another way in which your course may be different from ours is in terms of course funding. If you have little or not funding for your course (even after reaching out to relevant members of your department or institution), we suggest the following adjustments:

1) **Pair-Group-Based Projects**: Similarly to suggestion #3 for addressing different student levels, one option for limited course budgets is to have students conduct the replication projects as pairs or teams to reduce the cost of data collection. This structure may have the added benefit of encouraging students to problem-solve together. Alternatively, each student in the pairs or teams could complete each step of the replication individually (e.g., writing up the report, analyzing the data, interpreting the result), which would ensure that each student takes full responsibility for every step of the project. This structure may also provide opportunities for interesting discussions at the end of the course around analytic reproducibility, especially if students in the same teams (with the same dataset) differed in the conclusions they drew about the replication outcome.

2) **Funding from Advisors**: In some cases, students come to us with target studies that require more funding than we are able to allocate, but that they feel particularly invested in (e.g., because of how relevant the study is to their line of research). Once we rule out other ways of making the study fit our budget (e.g., dropping extra control conditions, running a subset of the study), we often ask students whether their advisor would be willing to fund the study. We have found that advisors are often willing to do this, especially if the replication could serve an important role in the development of the student's research program. Similarly, one way to reduce the burden on a limited course budget would be

to encourage all students to first ask their advisors about whether they would be willing to fund part or all of the data collection for the replication. While chances are that some advisors will be unwilling or unable to do this, there should still be a meaningful reduction in the number of projects the course will need to fund.

3) **Reproduce a Replication**: The suggestions above apply if you at least have *some* amount of course funding, but what if you have *no* funding at all? While there are obvious limitations to this solution, one suggestion is to have students reproduce past public replications. For instance, our course Github page[2], contains public repositories of all past replication projects that have been conducted in our course. Since the data for each replication project is available in these repositories, you could provide each of your students with a dataset and the original paper associated with it, and assign them to reproduce the results of the replication. Students should then be able to follow each step of the replication project described below (e.g., writing the report, identifying the key analysis, running the analysis). This format will only work if students do not view the original final replication reports that are posted publicly for their project, so it may be necessary to be clear about this at the beginning of the course.

[2] https://github.com/psych251

For those of you who are working with a different course format (whether in terms of student level or course resources), we hope these suggestions were useful. If you try out a new idea in your course that you found helpful, we would be thrilled if you shared them with us!

A.5 Sample course schedules

The sample syllabi laid out below are categorized along the following decisions: 1) Material: whether the course focuses on just content or both content and skills, and 2) Duration: whether the course is 10-weeks long or 16-weeks long.

For undergraduate instructors, we have labelled advanced topics in purple. We expect that these topics are best suited for advanced undergraduate students. As for content around statistics (e.g., Estimation, Inference), instructors should decide how much of this content to teach, depending on how prepared students have been in previous classes.

A.5.1 10 weeks

Table A.1: A sample 10-week syllabus with both skills and content materials.

Week	Day	Topic	Chapter	Appendix
1	M	Class Introduction	1	
1	W	Theories	2	
1	F	Version Control		B
2	M	Reproducible reports	14	C
2	W	Tidyverse Tutorial		D
2	F	Tidyverse Tutorial continued (with TAs)		
3	M	Measurement, Reliability, and Validity	8	
3	W	Design of Experiments	9	
3	F	Sampling	10	
4	M	Project Management	13	
4	W	Experiments 1: Simple survey experiments using Qualtrics		
4	F	Experiments 2: Project-specific Implementation (TAs)		
5	M	Estimation	5	
5	W	Inference	6	
5	F	Sample Size Planning		
6	M	Survey Design		
6	W	Midterm Presentations 1		
6	F	Midterm Presentations 2		
7	M	Preregistration	11	
7	W	Meta-analysis	16	
7	F	Open Science	3	
8	M	Visualization 1	15	E
8	W	Visualization 2		
8	F	Exploratory Data Analysis Workshop		
9	M	Sampling, Representativeness, and Generalizability	4	
9	W	Data and Participants Ethics	12	
9	F	Authorship and Research Ethics		
10	M	Open Discussion	17	
10	W	Final Project Presentations 1		
10	F	Final Project Presentations 2		

A.5.2 10 weeks, content only

Table A.2: A sample 10-week syllabus with only content materials.

Week	Day	Topic	Chapter
1	M	Class Introduction	1
1	W	Theories	2
1	F	Replication and reproducibility	3
2	M	Open Science	
2	W	Measurement	8
2	F	Design of experiments 1	9
3	M	Design of experiments 2	
3	W	Sampling	10
3	F	Experimental strategy	
4	M	Preregistration	11
4	W	Data collection	12
4	F	Visualization 1	15
5	M	Visualization 2	
5	W	MIDTERM EXAM	
5	F	Introduction to statistics	
6	M	Estimation 1	5
6	W	Estimation 2	
6	F	Inference 1	6
7	M	Inference 2	
7	W	Models 1	7
7	F	Models 2	
8	M	Meta-analysis	16
8	W	Project management	13
8	F	[Instructor-specific topics]	
9	M	Sampling, Representativeness, and Generalizability	4
9	W	Data and Participants Ethics	12
9	F	Authorship and Research Ethics	
10	M	Conclusion	17
10	W	Conclusion	
10	F	FINAL EXAM	

A.5.3 16 weeks

Table A.3: A sample 16-week syllabus with both skills and content materials.

Week	Day	Topic	Chapter	Appendix
1	1	Class Introduction	1	
1	2	Theories	2	
2	1	Version Control		B
2	2	Reproducible reports	14	C
3	1	Tidyverse Tutorial		D
3	2	Tidyverse Tutorial continued (with TAs)		
4	1	Measurement, Reliability, and Validity	8	
4	2	Design of Experiments	9	
5	1	Sampling	10	
5	2	Project Management	13	
6	1	Experiments 1: Simple survey experiments using Qualtrics		
6	2	Experiments 2: Project-specific Implementation (TAs)		
7	1	Estimation	5	
7	2	Inference	6	
8	1	Sample Size Planning		
8	2	Survey Design		
9	1	Midterm Presentations 1		
9	2	Midterm Presentations 2		
10	1	Preregistration	11	
10	2	Meta-analysis	16	
11	1	Open Science	3	
11	2	Visualization 1	15	E
12	1	Visualization 2		
12	2	Exploratory Data Analysis Workshop		
13	1	Sampling, Representativeness, and Generalizability	4	
13	2	Data and Participants Ethics	12	
14	1	Authorship and Research Ethics		
14	2	[Instructor-specific topics]		
15	1	Open Discussion	17	
15	2	Open Discussion		
16	1	Final Project Presentations 1		
16	2	Final Project Presentations 2		

A.5.4 16 weeks, content only

Table A.4: A sample 16-week syllabus with only content materials.

Week	Day	Topic	Chapter
1	1	Class Introduction	1
1	2	Theories	2
2	1	Replication and reproducibility	3
2	2	Open Science	
3	1	Measurement	8
3	2	Design of experiments 1	9
4	1	Design of experiments 2	
4	2	Sampling	10
5	1	Experimental strategy	
5	2	Preregistration	11
6	1	Data collection	12
6	2	Visualization 1	15
7	1	Visualization 2	
7	2	MIDTERM EXAM	
8	1	Introduction to statistics	
8	2	Estimation 1	5
9	1	Estimation 2	
9	2	Inference 1	6
10	1	Inference 2	
10	2	Models 1	7
11	1	Models 2	
11	2	Meta-analysis	16
12	1	Project management	13
12	2	[Instructor-specific topics]	
13	1	[Instructor-specific topics]	
13	2	Sampling, Representativeness, and Generalizability	4
14	1	Data and Participants Ethics	
14	2	Authorship and Research Ethics	
15	1	Ethics: Open Discussion	
15	2	Conclusion	17
16	1	Conclusion	
16	2	FINAL EXAM	

References

Frank, Michael C, and Rebecca Saxe. 2012. "Teaching Replication." *Perspectives on Psychological Science* 7: 595–99.

Hawkins, Robert D, Eric N Smith, Carolyn Au, Juan Miguel Arias, Rhia Catapano, Eric Hermann, Martin Keil, et al. 2018. "Improving the Replicability of Psychological Science Through Pedagogy." *Advances in Methods and Practices in Psychological Science* 1 (1): 7–18.

B Git and GitHub (online only)

This appendix appears only in the online version of this book at https://experimentology.io/B-git.

C R Markdown and Quarto (online only)

This appendix appears only in the online version of this book https://experimentology.io/C-rmarkdown.

D Tidyverse (online only)

This appendix appears only in the online version of this book https://experimentology.io/D-tidyverse.

E ggplot (online only)

This appendix appears only in the online version of this book https://experimentology.io/E-ggplot.