

Research Methods in Psychology

Understanding Science:
How Do They Know That?

Paul J. Reber, Ph.D.
Northwestern University



Table of Contents

Preface	<i>i</i>
Hands-on Research Methods	<i>viii</i>
Acknowledgments	<i>xiii</i>
1 Experimental Methods	1
Hands-on example: a two group study	
2 Psychological Measurement	21
Operational definitions: turning constructs into numbers	
3 Experimental Control	42
Managing extraneous variables	
4 Experimental Procedure	64
Designing how to carry out an experimental study	
5 Statistics 1	84
Evaluating data with a t-test	

6 Reporting in APA format	110
How to report the results of a study	
7 Within-Participants Design	134
Using conditions instead of groups	
8 Research Ethics	151
Fair treatment of research participants	
9 Factorial Design	181
Designs with more than one independent variable	
10 Interpretation of Factorial Data	198
Interpreting results from more complex designs	
11 Participant Sampling	216
The impact of participant recruiting processes	
12 Statistics 2: ANOVA	231
Analysis of Variance for complex designs	
13 Proposing Research	253
Planning your own research project	

14 Non-experimental Design	269
Research without a manipulated variable	
15 Surveys & Instruments	288
Use and construction of surveys in research	
16 Statistics 3	319
Correlation and Chi ² analysis approaches	
17 Qualitative Research	340
Observational and interview research	
18 Ethics 2: RCR	358
Responsible conduct of research	
19 Applied Research	378
Field and intervention research	
20 Development and Neuropsychology	394
Case studies and specialized methods across the lifespan	

Preface

Understanding the method by which research is done is a core part of most scientific study. Psychological science, the science of human behavior, has its own characteristic set of methodologies and challenges in drawing robust conclusions. There are two major goals of instruction in research methods to university students. First is to prepare for participation in the process of science through collaborative research done as advanced students. Second, to develop a basic understanding of how inferences about the world are drawn through scientific study with a grasp of the strengths and weaknesses of specific scientific methodologies.

It has become particularly clear in recent years that this second point is an important element of critical thinking about science that has been a particular challenge for the general populace. People seek to better understand the world around them and are exposed to a wide variety of scientific claims and results, but are significantly hampered by a lack of understanding of the methodologies used to draw those claims. Without being prepared to critically review and understand how robust or reliable these claims are, misinformation spreads rapidly and dangerously. Attempts to combat misinformation directly have the unfortunate side effect of weakening confidence in the scientific process in general and shifting attention back to anecdotes and information personally observed.

Teaching methods and the process by which science is done is tricky. People naturally seem to like to learn facts and findings, but the ideas about the meta for how these findings were obtained does not appear to elicit the same natural curiosity. This is something that teachers of science need to work to overcome in order to generally increase the overall scientific literacy of the populace. That the method is interesting itself is something that can even surprise experienced scientists. Some years back in conversation with Kathleen Grady, Ph.D. (the author's mother), she remarked on her own surprise at being captured by interesting aspects of methodology framed as asking the question "How do they know that?" when encountering some brand new, unexpected result.

We may be able to inspire better understanding of how science informs us of the world around us by both encouraging asking this question and providing the tools to try to answer it. This question is taking as the sub-title for this text.

The structure of this text reflects an attempt to create a Research Methods textbook that aligns with the teaching style we use at Northwestern University. In a single 9-week quarter, we use a very hands-on approach to experimental research methods that incorporates both teaching the basic elements of design and significant APA-style writing assignments. We find this approach very effective for preparing undergraduates to understand research basics and be ready for both upper-level research oriented classes and opportunities to work directly within department research labs.

However, this requires an unusual pacing of the class that does not align with most traditional research methods in Psychology textbooks. Rather than starting with a more gentle introduction to the importance of science, the philosophical ideas about drawing inferences from human behavior or even an overview of research ethics, Chapter 1 in this approach is plunging in to a basic research design through an active example. After many years of starting with Chapter 7 in traditionally structured texts, I decided to try to prepare a text that followed the pacing of our course design.

While our course pacing may be idiosyncratic to my institution, I have also come to believe that rapid engagement with hands-on examples may be an effective tool for overcoming the natural disinclination to learning about methodology. In the abstract, the rationale and statistical purpose of employing a two-group design with random assignment to conditions is a fairly dry and possibly boring idea. Perhaps we can inspire more engagement with the concepts by seeing the concepts in action and immediately facing the questions of what we learn from data obtained via this methodology.

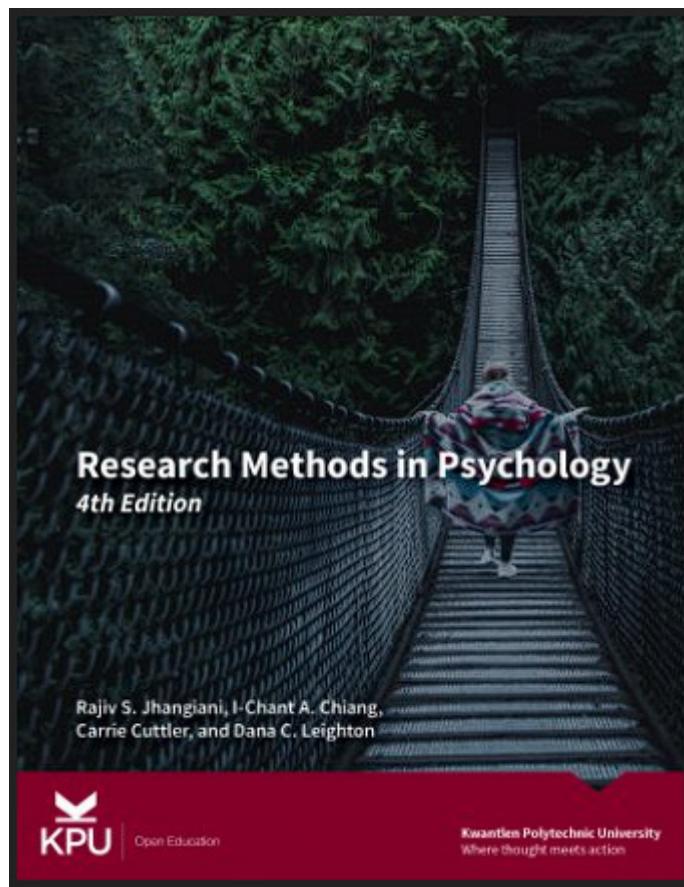
The content is aimed not to completely overlap with my classroom lecture slides content. I do not want the students to feel that classroom time is spent completely rehashing the text. I prefer to have different, novel examples illustrating the concepts and to use a very question-and-answer style in the classroom to maintain student engagement. I would like to work towards having that information available to students as well without minimizing the value of coming to class, but I'm not sure how to organize it.

We will start by reviewing methods of experimental psychology research. Typical textbooks for Research Methods start with a review of the scientific method, some history on psychological science, discussion of non-experimental methods and then the process of the design, implementation, and analysis of experimental research. Because this class is designed to be hands-on with active involvement in the actual course of research, we are starting immediately with experimental methodology.

Somewhat initially daunted by the prospect of preparing an entire textbook of content for the class, I started this project based on an open-source textbook made freely available by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, and Dana C. Leighton. This text was invaluable in motivating this process and the plan is to make this content similarly open-source and freely available.

You may notice some residual redundancy in the text, especially in areas where conceptual ideas are explained related to specific content for the Reber/NU class presentation and then explained again as presented by the original authors of the text. In some places these are left deliberately to help

build a better understanding of complex or non-intuitive ideas by multiple explanations from slight different perspectives.



Research Methods in Psychology

4th edition

RAJIV S. JHANGIANI; I-CHANT A. CHIANG; CARRIE CUTTLER;
AND DANA C. LEIGHTON

KWANTLEN POLYTECHNIC UNIVERSITY
SURREY, B.C.

Jhangiani et al. (2022) License

Research Methods in Psychology by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

This adaptation constitutes the fourth edition of this textbook, and builds upon the second Canadian edition by Rajiv S. Jhangiani (Kwantlen Polytechnic University) and I-Chant A. Chiang (Quest University Canada), the second American edition by Dana C. Leighton (Texas A&M University-Texarkana), and the third American edition by Carrie Cuttler (Washington State University) and feedback from several peer reviewers coordinated by the Rebus Community. This edition is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Jhangiani et al. (2022) Preface

Psychology, like most other sciences, has its own set of tools to investigate the important research questions of its field. Unlike other sciences that are older and more mature, psychology is a relatively new field and, like an adolescent, is learning and changing rapidly. Psychology researchers are learning and changing along with the emerging science. This textbook introduces students to the fundamental principles of what it is like to think like a psychology researcher in the contemporary world of psychology research.

Historically, psychology developed practices and methods based on the established physical sciences. Unlike physical sciences, psychology had to grapple with the inherent variation among its subjects: people. To better account for this, we developed some practices and statistical methods that we (naïvely) considered to be foolproof. Over time we established a foundation of

research findings that we considered solid.

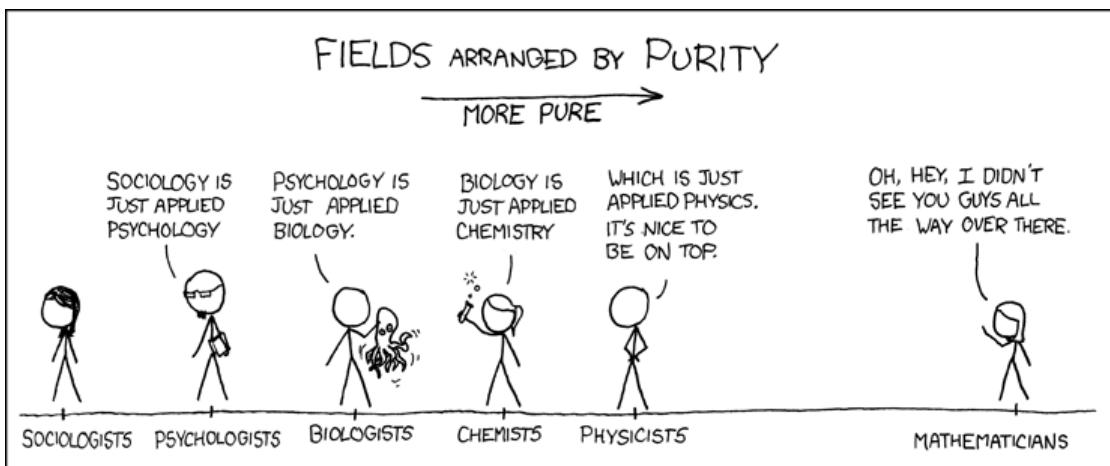
In recent years, psychology's conversation has shifted to an introspective one, looking inward and re-examining the knowledge that we considered foundational. We began to find that some of that unshakable foundation was not as strong as we thought; some of the bedrock findings in psychology were being questioned and failed to be upheld in fuller scrutiny. As many introspective conversations do, this one caused a crisis of faith.

Psychologists are now questioning if we really know what we thought we knew or if we simply got lucky. We are struggling to understand how what we choose to publish and not publish, what we choose to report and not report, and how we train our students as researchers is having an effect on what we call "knowledge" in psychology. We are beginning to question whether that knowledge represents real behaviour and mental processes in human beings, or simply represents the effects of our choice of methods. This has started a firestorm among psychology researchers, but it is one that needs to play out. For a book aimed at novice psychology undergraduates, it is tempting to gloss over these issues and proclaim that our "knowledge" is "truth." That would be a disservice to our students though, who need to be critical questioners of research. Instead of shying away from this controversy, this textbook invites the reader to step right into the middle of it.

With every step of the way, the research process in psychology is fraught with decisions, trade-offs, and uncertainty. We decide to study one variable and not another; we balance the costs of research against its benefits; we are uncertain whether our results will replicate. Every step is a decision that takes us in a different direction and closer to or further from the truth. Research is not an easy route to traverse, but we hope this textbook will be a hiking map that can at least inspire the direction students can take and provide some absolute routes to begin traveling.

As we wrote at the beginning of this preface, psychology is a young science. Like any adolescent, psychology is grappling with its identity as a science, learning to use better tools, understanding the importance of transparency,

and is having more open conversations to improve its understanding of human behaviour. We will grow up and mature together. It is an exciting time to be part of that growth as psychology becomes a more mature science.



Hands-on Research Methods

Overview

A key element reflected in this approach to teaching research methods is that it is effectively learned by active participation in the scientific process. Conceptual ideas and explanations of methodology will be covered in the text and these will be illustrated by concrete examples of implementing experimental design, collecting data, analyzing the results and preparing written reports on the outcome.

Students will complete three experiments over the course of the quarter, which is a substantial amount of work to be completed within the class schedule. A major inspiration for preparing this textbook is to organize the material relevant to experimental design to make the concepts available in parallel with the hands-on practice through these experiments. As a consequence, we start quickly with key design principles on the first day of class, skipping over the more typical introduction to scientific methods reviewing the history and philosophy of the field.

The three experiments will be used to illustrate design, statistical analysis, drawing inferences from the data and writing scientific reports in the standard APA style. The chapters are organized to make the relevant information

needed as available through this process.

Experiment 1

On the first day of class, students participate together completing a short memory experiment. This design and these data are used to illustrate basic experimental design, data analysis and to start the process of learning to write a formal, APA-style scientific report.

Chapters 1-4 introduce the basic concepts and terminology for experimental psychological research.

Chapter 5 explains how to analyze the Experiment 1 data.

Chapter 6 introduces APA-style scientific reporting to support students writing a brief overview of Experiment 1 as the first major writing assignment.

Chapter 7 extends simple research designs to include within-participant manipulations.

Chapter 8 introduces research ethics in Psychological science. The first midterm exam is typically given after this class.

Experiment 2

Chapter 9 and 10 introduce factorial design as a more complex variation of experimental research methods. We focus on 2x2 designs as the simplest version of this more complex design.

Experiment 2 is developed collaboratively with the students as a 2x2 design that extends Experiment 1. The students are responsible for recruiting participants and collecting data.

Chapter 11 discusses the issues of sampling, generalizability and how this affects inferences from data.

Chapter 12 supports the tools needed to carry out the analysis of Experiment 2 by the students.

The second major writing assignment includes both Experiment 1 and 2 and is completed at this point.

Student-led research

Students then begin their final projects for the class, starting with submitting a proposal to carry out their own, short research project.

Chapter 13 goes over the development of a research proposal to support the final project process.

In parallel with the final project process, the remaining chapters cover material typically included in Research Methods instruction that does not immediately support the three hands-on research projects.

Chapter 14 covers non-experimental design. Chapter 15 provides an overview of survey research and related instruments. Chapter 16 rounds out basic statistical tools including correlation and chi-squared tests. Chapter 17 introduces field research. Chapter 18 extends ethics in research to Responsible Conduct of Research. Chapter 19 introduces quasi-experimental design. Chapter 20 reviews special methodological considerations for Developmental and Neuropsychological research.

Advanced Research Methods

An alternate approach to teaching research methods focuses on the contrast between experimental design and non-experimental design, highlighting the differences in supporting causal inference of designs where the experimental independent variables can be manipulated. Here we focus primarily on the simpler experimental methods but acknowledge that there is a great deal of Psychological science that depends on methods termed non-experimental or correlational. These approaches depend on a set of methodological and sophisticated analytical tools that are outside the scope of this text, but could be seen as subsequent areas of study for interested students.

Acknowledgments

The idea to organize my notes and ideas about how to teach research methods into a textbook has been developing over the 25 years that I have been teaching this class at Northwestern University. I would like to acknowledge and thank the students in those classes, the graduate students who have helped out as Teaching Assistants over that time and the handful of faculty I have compared notes with about the structure, content and pacing of the class.

I would like to particularly acknowledge the students from Fall quarter 2022 and their teaching assistant, **Ouxun Jiang**, who were the first class to see the very first draft of the textbook.

The current classes in Winter quarter 2024 will be the first two sections to see the content attempted to be reformatted in a manner that roughly resembles a textbook. My teaching assistants **Zixin Zhang & Grace Coram** have been helpful with feedback, format suggestions and typo catching even before class starts. In addition several students provided feedback on the text, spotting typographical or other errors: **Dominic Bayer, Philip Bell, Lucy Munro, Yinghan Du and Madison Rowzat**. I also greatly appreciate the patience of all the other students in both sections of the class who had to work with a textbook that was in progress and often only available a few days before we covered the content in class.

I believe the current version is substantially improved from the prior version. I think there is a lot of room for additional improvement that I hope to continue to add over the next few years. I have a fairly substantial list of revision ideas for the next version, which will be used in the Fall of 2024 and might see some adjustments to the content and chapter re-organization.

On a more personal note, my parents both have Ph.D.'s in Psychology and have had long careers doing psychological science. My mother, Kathleen E. Grady provided my "tag line" here of *how do they know that?* from her own brief interaction with writing about methodology. Her career was in applied health research at the Massachusetts Institute Behavioral Medicine, which she founded. At one point in her career, a long-time friend asked her to help with materials for a textbook on research methods, which gave her the opportunity to write and think about a lot of these topics. That text was aimed at a different student group and class pacing and so did not contribute directly to this text other than the inspiration to take on the task.

My father, Arthur S. Reber, was a Professor of Psychology at Brooklyn College and the City University of New York. Outside of his research in learning and consciousness, his passion was words. During his career he wrote the Oxford Dictionary of Psychology, providing definitions for the terms of our field in addition to several books on learning and consciousness. The dictionary was a lifelong passion project for him and something he cared deeply about. Printed dictionaries are somewhat out of style in the modern age of online information and internet search, but I do hope to poach some of his detailed and thoughtfully written term definitions where I can in this text.

1 Experimental Methods

Hands-on Approach

- Participate in a short psychology experiment using the QR code or the link below. When you have finished you will get a Completion Code to enter as the answer to the first assignment for the class.
- *Note: the experiment and questions/discussion below will be covered on the first day of class.*
Review the Q&A below if you want a refresher for that discussion.
- Answer the following questions about that study. The following series of questions is based on the experiment but assumes some prior experience with psychological science. Since Research Methods typically follows and builds on classes *Introduction to Psychology* and basic *Statistics*, we therefore assume some familiarity with basic terms and ideas. Here we aim to reinforce understanding of these core ideas within the framework of what a simple experimental design looks like from the above hands-on example.



Or use the following link:
<https://tinyurl.com/Reber205>

For the following questions it is a useful exercise to cover the answers and try to answer the questions yourself before reading on. This will help you assess how much of the basic terminology and experimental approach you are already comfortable with. The terms will be defined in this chapter for general reference. The goal here will be to use the main terminology frequently enough that it simply becomes part of your understood vocabulary without need to look definitions up later on. The bolded terms below are ones to start becoming comfortable with.

What was this experiment about?

The general temptation for the answer to this question is to give a lot of detail about your experience with the experiment and guesses about how this relates to the underlying hypothesis. However, after just going through the experiment, you actually do not know what the experiment is about because you have not seen enough of the design. This is a typical experience for a participant in an experiment that has an **independent variable** that is manipulated **between-participants**. You only experienced one of the conditions, so the underlying **hypothesis** is not visible to you.

However, when we consider and evaluate research with examples as short

summaries or drawn from published papers, we will always start with this question and the answer we are looking for in this very basic question is the highest-level **construct** that gives the overall domain of the experiment. Here, that is simply “memory.”

As we will see, designing an experiment in psychology generally starts with something we are trying to learn about. In psychology, that will be a concept like memory, perception,

Key Terms

The bolded terms in the answers are key concepts in experimental design that will be used daily in class and throughout the text. A glossary of definitions is provided below for general reference.

anxiety, relationships, language, identity, etc. One of the specific challenges of experimental methods in psychology, as opposed to other areas of science (chemistry, physics, biology), is that while we intuitively understand each of those concepts, there is a significant amount of effort needed to turn that idea into things that can be used in research. That process is called identifying the **operational definition** of the **construct**, which is essentially, how are we going to capture that idea in a controlled study.

Answering the next questions will require being familiar with some technical terms that you may have encountered in prerequisite classes. If you are unfamiliar with the terms, they are defined below for your reference.

What was the independent variable?

To answer this question, you need some additional information. There were two different conditions used in this experiment. Half of the time, participants are given instructions to rate how much they like each word, on a 1-5 scale from “very much” to “not at all.” The other half of the participants get instructions to count how many vowels there are in each word and also make a response on a 1-5 scale.

The **independent variable (IV)** is the conditions created by the experimenter and applied to the participants. Here it is the instructions given for how to read and engage with the list of words. A more interesting question is what **construct** is this **independent variable** an **operational definition** of? What is the construct that the experimenter is manipulating in this study? The answer is “depth of encoding” which refers to how much engagement the participants have with the meaning of the words in the study list. Understanding why this is an interesting factor to manipulate will require some background reading to become familiar with the theory (which we will get to later).

Here, “depth” is an **experimental operational definition**, which refers to turning this **construct** (concept) into conditions that can be applied to a

research experimental design. Rating liking creates a higher level of depth by encouraging semantic engagement with the words. Counting vowels creates comparatively lower depth by focusing the participant on surface features of the word instead of meaning. The experiment is about how these conditions affect memory, which raises the next question.

What was the dependent variable?

The **dependent variable (DV)** in this experiment is a measured operational definition of memory, as in, how much memory did participants have of the word list after engaging with the work list in either of the experimental conditions. A measured operational definition turns a concept/construct into a quantitative number used to measure outcome. Here, the answer will be a numeric measure of performance on the recognition test that came at the end of the experimental protocol.

After going through the initial interaction with 30 words in *the study phase*, you completed a short delay/distraction task based on answering trivia questions. Then you completed a recognition memory task in which you were presented with 60 works, the 30 you saw initially and 30 words that you did not see at the beginning. Note that you might be tempted to answer

the question of "what is the DV?" with "the number of studied words you responded *old* to on the test." Here that is not quite correct as answering *old* to all 60 words would not reflect good memory (because you called all the new words old). More accurate is to describe the DV as score on the recognition test, which we can count as the number of test items responded to correctly (old called old, new called new).

Measuring Memory

If you are familiar with memory research, you might be familiar with more sophisticated ways to measure memory. A simple percent correct measure is enough for our simple study but not for all memory research.

State a hypothesis relating the independent variable to the dependent variable.

This is the first question that engages with the psychological science of the research study. The first few questions are just identifying the key terms as a basis for figuring out what the study might tell us about human thought or behavior. Stating **hypotheses** about experimental variables is a deceptively tricky task. It requires that the stated hypothesis be testable or falsifiable, which is not the same as correct.

Any statement relating the levels of the IV to scores on the DV are correct answers to a prompt like this. The hypothesis relating the experimental variables is: rating liking of words will lead to higher scores on the recognition test than counting vowels. Stating the opposite, that counting vowels will lead to higher recognition scores compared with rating liking is also an equally valid hypothesis, although we will see that it is false. That is, it is not supported by the data.

For the purpose of this question, stating the hypothesis in terms of the constructs would not be correct here. At some level, the experiment is about the hypothesis that deeper encoding of items being studied leads to better memory later. This is a perfectly valid hypothesis but in our analysis process we first focus specifically on how the experimental design tests a hypothesis about the experimental IV affecting the experimental DV.

An important part of analyzing an experiment is to find problems or errors in methodology. When we design studies, we need to consider our design critically to see if any errors have crept into our approach. And when we review research reports that we encounter and ask the question "how do they know that?" we should be looking for potential problems with the conclusions.

By explicitly framing the question in terms of the variables as asked here, we focus our attention on how the constructs of *deep processing*, *shallow processing*, and *memory* are implemented in this specific design. For

example, memory here is operationally defined as a recognition test for the list of words. A statistically reliable result for this study allows us to make a confident statement about how this independent variable affected this dependent variable. However, extending the idea from this study to all other ways we might study memory is an additional step that we should consider carefully.

One of the important and unique aspects of psychological science is being aware of the difference between the experimental design and data, which are based on operational definitions, and the theoretical conclusions, which are based on constructs. In this design, the operational definitions led us to use lists of words as the things to be remembered and one specific approach to what we mean by *depth of encoding*. These might be important **limitations** to consider about our conclusions, for example, do they apply to non-word stimuli, or how does depth influence other kinds of ways to measure memory?

The data obtained will tell us about the relationships of the variables we used in the experiment, pending the appropriate use of a statistical test to evaluate the reliability of any effects observed. Following this, we hope to draw a theoretical inference about the constructs as the scientific conclusions about the study. Critically evaluating research requires being able to identify methodological issues that might limit those conclusions that arise at any step in the research process.

What statistical test would we use to establish a reliable relationship between our independent and dependent variables that would allow us to test our hypothesis?

In virtually all psychological science, we are going to collect or consider data collected from a group of participants in our study. The people in our study are considered the **sample**, who are drawn from the larger **population** of all people. We want to make a broader statement than simply that the people who happened to be in this study showed better memory after deep

encoding, we want to infer that deep encoding would likely improve memory for all people. Statistical analysis is the method for drawing that broader inference that deep encoding generally improves memory and future uses of deep encoding by anybody would most likely improve their memory for the studied material. This basic idea should be familiar from your prior study basic statistical methods from a prerequisite class. However, statistics will be used here in a potentially different manner than in prior classes. Here they will be the bridge from your numeric, quantitative data to statements about the conclusions and meaning of your study.

Since this is a simple two group design with participants randomly assigned to one condition or the other, the most appropriate statistical test would be a **two independent samples t-test**. While other more powerful approaches could certainly be used, it is generally most effective to use the simplest test that effectively communicates the main findings.

For simple experimental design, where participants are randomly assigned to one or two conditions of one or two independent variables, questions of reliability are generally simple and often relatively uninteresting. Our use of statistics here will therefore be streamlined. We will focus on identifying the correct test to use from a constrained set of options and provide a recipe to carry out the analysis within the program R/RStudio. The result of the analysis will be reported in standard format (based on the American Psychological Association; APA) as part of the

Psychology is a STEM field!

Research Methods is about using the scientific method to understand human behavior, attitudes, cognitive processes, social interactions, personality and mental health. It is fundamentally quantitative even though advanced math skills are not strictly necessary for basic design

process of writing up the results of a study. While a strong foundational grasp of the underlying mathematics is always helpful, we will primarily focus on how statistics are used to test research hypotheses and how to report these in a result that is complete and comprehensible to other scientists.

In carrying out a research project, statistics are used to establish the **reliability** of the effect of your IV on your DV. As we will see over the next several chapters, this is a separate question of the **validity** of your conclusions drawn from the study.

In general, the review of basic experimental methodology will focus more on validity of experimental design and data than reliability. Details of the statistical approach become more important as experimental design becomes more complex. Here as we review how to design research studies, we will focus on procedures aimed to obtain reliability that are assessed through statistics. When we review descriptions of published research findings, we know that the peer-review applied to these findings before publication generally establishes statistical reliability. By asking How do they know that? and applying an understanding of experimental design, we will identify

questions of validity and alternate interpretations of the findings that might differ from the experimenters' stated hypothesis.

Psychology is the science of human behavior.

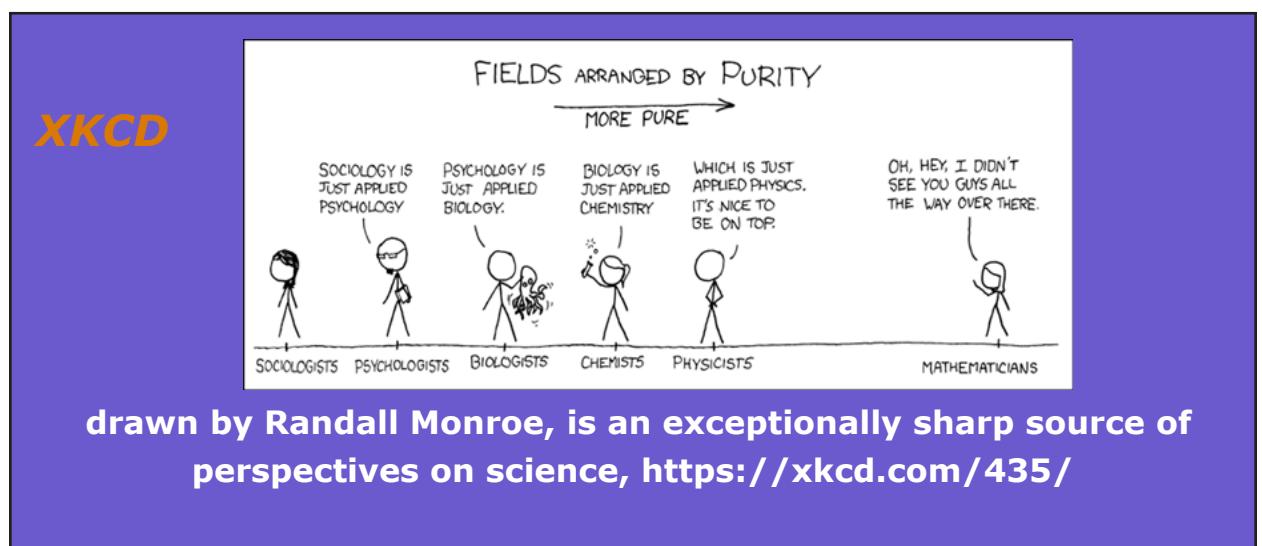
As a human, you have pretty good intuitions about how humans behave. Personal knowledge and experience can be a good starting point for the tricky problem of coming up with the operational definitions for psychological constructs.

Much of your ability to identify weaknesses in scientific methodology will come from your understanding of human behavior as a human. In this class, we will augment this with some practice applying critical thinking skills systematically to these questions.

Experiment 1

Our first experiment is based on some fairly old ideas in memory research but which hold up well for a simple demonstration experiment. The underlying ideas are described in Craik & Lockhart (1972), which lays out a **framework theory** for thinking about memory. Craik & Tulving (1975) reports a series of experiments that establish that manipulations designed to vary the *level of processing* or *depth of processing* have robust and reliable effects on measures of memory. While the core terminology and theoretical framing presented in these older papers is slightly out of date by more modern theories of memory function, the procedure still serves as an excellent example of a simple design that consistently produces a measurable effect.

Experiment 1 will be used to illustrate the typical path from theory through experimental design, data collection and analysis. We start with constructs like memory and a hypothesis, does deeper engagement with material lead to better memory? These are then turned into an experimental operational definition (liking and vowel counting) and measured operational definition (recognition test). Data are collected and will be analyzed. The statistical test will be used to allow us to support (or not) a statement about whether the IV reliably affected the DV. From there we will draw a final conclusion about how we think the original concepts are related and whether the data support the original hypothesis (or not).



Experiment 1 Design

Random Assignment to Conditions either Deep or Shallow, which implements the study IV

JOURNAL

For each word, rate how much you like the word on a 1 to 5 scale.
5 means you like the word very much and 1 means you dislike the word very much. Indicate your rating with keyboard or mouse.

Words will appear here
Follow the instructions above
Make your responses below
With the 1-5 keys, mouse or touchscreen.
There may be a short delay before the next word
Press Enter when ready to start

dislike a lot	dislike a little	neutral	like a little	like a lot
1	2	3	4	5

QUARTER

For each word, count the number of vowels in each word
Enter the number with keyboard or mouse

Words will appear here
Follow the instructions above
Make your responses below
With the 1-5 keys, mouse or touchscreen.
There may be a short delay before the next word
Press Enter when ready to start

1	2	3	4	5
---	---	---	---	---

Although the instructions differ, every word is shown for 4s to maintain matched viewing time

Both Groups See the Same 30 Words

**POCKET, PAINT,
PRISON, QUARTER,
CITIZEN, VEHICLE,
ROUGH, BRAIN, TEMPLE,
PRINCE, MEDICINE,
FILLING, GUARD,
JOURNAL, ENGINE,
PALACE, GRAVE,
BRANCH, CONCRETE,
DANCER, SALARY,
BASEMENT, MATCH,
NATIVE, STABLE, FENCE,
SWIMMING, QUEEN,
OCEAN, FACTORY**

After performing the rating task, the rest of the study procedure is the same for both groups

Lord Byron and William Blake wrote primarily in what genre of poetry?

1. Modernist
2. Renaissance
3. Romantic
4. Victorian

All done with trivia

Next is a recognition test
Words will be shown here
If you think you saw this word in the first part before trivia
Choose old below, otherwise new
Press Enter when ready to start

old					new
1	2	3	4	5	

Everybody completes 3 minutes of trivia between seeing the words and the recognition test

The study DV is obtained from the recognition test. In addition to the 30 studied words, 30 unstudied (new) words are shown. All participants get the same test

Recognition memory test				
QUARTER				
old				new
1	2	3	4	5

Recognition memory test				
ROUGH				
old				new
1	2	3	4	5

30 Unstudied Words For Memory Test

CLOTH, DRIVER, LIQUID, SOLDIER, GALLERY, CHAIN, LUMBER, STEEL, MISSILE, SMOKE, AUTHOR, OWNER, PORCH, FISHING, SMELL, SHADOW, TRACK, GUEST, DISPLAY, MUSCLE, SPEAKER, WEDDING, WORKER, GUIDE, BRUSH, HIGHWAY, NAVAL, CARBON, PARTNER, PASSAGE

The recognition test is scored as percent correct of responses out of 60 possible. Calling a word seen in the first phase “Old” is correct. Calling a word not seen “New” is also correct.

Materials

A set of 60 words was used for the study and test stimuli. Words were selected to have a written frequency of 30-80 per million and to be 5-8 letters in length.

Each participant obtains a score on the test and the average performance across the two groups is compared with a two independent samples t-test.

Some key experimental design terms

- **Experimental research:** The experimenter manipulates an independent variable and measures a dependent variable to test if the manipulation has an effect.
- **Construct:** The high-level concepts we aim to do research about. Typically, these things we that have an intuitive understanding of but need to be translated into specific experiment elements.
- **Operational definition:** Turning an intuitive but imprecise concept into something that can be measured quantitatively, or controlled categorically.
- **Measured operational definition:** A quantitative measure of a construct, essentially turning an idea into something that can be characterized as a number. For example, Experiment 1 operationally defines “memory” as percent correct on the test, a quantitative measure of the amount of memory obtained. A similar process might turn other constructs like anxiety, impulsiveness, attention into numbers that could be used as dependent variables in experimental design.
- **Experimental operational definition:** A controlled method of implementing a specific definition of a construct into levels or categories that can be manipulated by an experimenter in order to create the independent variable(s) for an experiment protocol.
- **Independent variable (IV):** Often referred to by the acronym IV, this is the element manipulated by the experimenter to see if or how it affects the measure being collected in an experimental design. Controlled manipulation of the IV is the defining feature of experimental research.
- **Dependent variable (DV):** Frequently referred to by the acronym DV, this is the measurement collected by the experimenter. The core idea in experimental research is to see how the scores on the DV change across the manipulation of the IV. If they do, we can conclude that the IV affected the DV.
- **Experimental Hypothesis:** A statement about the relationship between experimental variables that can be tested and importantly, falsified. If there are no data that would render a statement false, then it is not a falsifiable statement and is typically a description rather than a hypothesis. Typically the hypothesis is that the IV affects the DV, and we use statistics to reject the **null hypothesis**

(that the IV does not affect the DV). Note that hypotheses can be stated about the specific IV and DV used in an experiment but also stated separately about the constructs from which the IV and DV were operationally defined. Experimental data gives us confidence to make statements about the specific IV affecting the implemented DV but the goal of research is to draw inferences about the relationship among the constructs.

- **Limitations:** Concerns that conclusions about the underlying constructs might not be true in all cases and conditions other than the specific operational definitions used in the experimental design. Generally these are not issues with the fundamental **validity** of the experiment (Chapter 3), but questions about how widely the results can be applied. Identifying what limitations should be considered often requires some knowledge of the underlying theoretical ideas for a research study and can also indicate directions for future research. Using Experiment 1 as an example, we have data about memory for word lists measured with a recognition test a few minutes later. We might wonder if deeper encoding similarly affects memory for pictures, or if the effect might change with another measure of memory like recall. Studies examining those questions would reflect different operational definitions of memory, using different DV's and/or different operational definitions of deeper encoding as IV.
- **Statistical reliability:** We will evaluate whether the IV has a robust effect on the DV using standard statistical tools. Our focus here will be selecting the correct tool and reporting the use of the tool accurately. Statistics are often presented as a simple binary outcome: did the IV affect the DV reliably, can we reject the null hypothesis, was the probability of the null less than the criterion of .05 (these three statements are essentially synonymous). However, we will see that Psychological Science is moving towards a model of reporting **effect size** rather than relying on these binary descriptions. The effect size is helpful both with understanding the reliability of the statistics and also communicating the results. For Experiment 1, we might want to be able to say not just that deep encoding improved memory, but how much did this study approach increase our measure of memory?

Experimental vs Non-Experimental Research

A useful approach for understanding the definition of something complex, like experimental research, is to define what isn't experimental research. In non-experimental research, we also look for a relationship between an independent variable and a dependent variable, but the independent variable is not manipulated or controlled by the experimenter. For example, we could look for a correlation between your GPA and the score on the memory test in the demonstration experiment.

Non-experimental research is a powerful tool for psychological science as well as fields such as epidemiology, economics and sociology. However, the methods of the design of research studies and tools for analysis of data for non-experimental methods are quite different. The current approach focuses on experimental methods first, followed by some discussion of contrasting these methodologies for general reference in Chapters 9-11.

Experimental research has a significant advantage in drawing conclusions about how a manipulated variable (IV) affects a measured variable (DV). If we manage the challenge of adequate experimental control (Chapter 3-4) we can be fairly confident that changes in our DV were caused by our manipulation of the IV. However, experimental design is limited by needing conditions where we can create effective and accurate operational definitions of the constructs we want to study so that we can implement a protocol for a well-controlled laboratory experiment. There are a lot of important and interesting questions in Psychology that rely on data collected from the world in imperfectly controlled conditions.

Non-experimental research typically fights against the "correlation is not causation" problem and frequently uses more advanced quantitative analytic tools to improve our ability to draw causation from these data.

Experimental research uses simpler methodology and simpler analytic tools, making it an effective introduction to the design of psychological research.

Experimental Analysis

The following questions will be asked regularly about example designs and findings from psychological research. These will train your intuition to identify strengths and weaknesses of designs from short research descriptions. Later we will see how to read and write formal research reports following APA guidelines. Most of the research that you encounter will be in more informal context, but you can still ask the question: **How do they know that?**

- What is the experiment about?
- What is the dependent variable?
- What is the independent variable?
- What is the hypothesis or finding about how the IV affects the DV?
- What statistical test is used to establish a reliable effect?
- What is the conclusion drawn by the researcher?
- Do we see any problems with this inference?

Trying to identify the hypothesis and potential problems with the inference are the hardest but most important questions from this list. If there was a tried-and-true approach to always identify inference errors, professional researchers would never make mistakes about their findings (spoiler alert: they do).

The first three questions depend on the operational definitions used by the researchers and how well they capture the intent of the research. When there is a mismatch, this often reflects differences in how people understand common phrases. For example, we might want to test a hypothesis related to an adage like “time flies when you are having fun.” One of the first challenges we would face is how to define the constructs of “time flies” and “having fun.” Different researchers would likely define these ideas in different ways and rather than saying that some operational definitions are right or wrong, it is important to understand that the different definitions reflect different design ideas. Experiments with different definitions might be quite properly

constructed, but the conclusions drawn from carrying out the study might end up being very different.

Chapter 2 will discuss operational definitions as an example of Measurement Theory. Mistakes in operational definition are one important source of error in experimental design. These can lead to studies where the results are quite robust, the IV clearly strongly affects the DV, yet the main conclusion of the study is inaccurately stated because the variables are ineffective operational definitions of the constructs they were intended to capture.

The question of what statistical test is appropriate for the research is necessarily more technical. As noted above, this class assumes background in basic statistics. In Chapters 5 and 10, we will review the process of selecting and carrying out the appropriate statistical tests for common experimental designs. The focus here is knowing which analysis to use, how to carry out the basic analysis procedure and most importantly, accurately state the inferences the analysis supports.

Understanding the hypothesis and conclusions that are tied to the IV and DV, the specific operational definitions used in an experimental design is the key to ensuring you understand how to read and interpret scientific findings. Being an effective reader of science and understanding what is confidently learned from the data obtained in a psychological study is a major goal of this class and text.

Experimental Analysis Practice Examples

Practicing experimental analysis and learning the common types of research design will give you critical thinking tools to help strengthen your understanding of science. We will practice via example throughout class meetings with a daily example to evaluate and analyze.

Example 1

Time flies when you're having fun, but what is it about pleasant experiences that makes time seem to go by faster? In one experiment inspired by prior work (Gable & Poole, 2012), researchers tested the hypothesis that approach motivation causes perceptual shortening of time during pleasant experiences. That is, it isn't just positive affect (fun), time goes quickly when you are specifically motivated to obtain a reward. Thus, they predicted that time spent viewing pictures of "delicious desserts" would appear to go by particularly quickly if you expected to get to eat one of the desserts after the experiment.

Participants were randomly assigned to either be told they would get to eat a dessert after the experiment or not. Then they each looked at 36 pictures of desserts each presented for a 12s and rated a scale of 1 (time dragged) to 7 (time flew), how long the picture had been presented.

Go through the Experiment Analysis questions for this example

What is the experiment about?	The subjective experience of time passing
What is the dependent variable?	The numerical scale rating from 1 to 7 of whether time dragged or time flew
What is the independent variable?	Told they would get a dessert after the study or not
What is the hypothesis or finding about how the IV affects the DV?	Participants told they would get dessert would score higher on the DV reflecting a feeling that time flew
What statistical test would be used to establish a reliable effect?	Two independent samples t-test

If the data were consistent with the hypothesis such that scores on the time-passing rating scale were higher for the participants who expected a reward, the researchers would like to conclude that expecting reward makes time feel like it is passing more quickly.

We should always consider limitations of the broad level conclusion. We might note that the task is particularly dull but also intrinsically linked to the reward (both are related to eating dessert). We might also note that the conclusion does not argue against the idea that time flies when you are having fun, but only suggests time might also fly when you expect dessert.

Example 2

Martin hypothesizes that self-esteem affects snacking behavior. He thinks that low self esteem will lead to increased opportunistic eating. He conceives of the following experiment. A group of 50 participants is recruited. All are given the opportunity to play a game of chance. They are all told that the odds are in their favor and that 90% of the people who play win the game. However, they are really assigned randomly to two groups: half win and half lose. The winners are congratulated and the losers are told, "Wow, that's really unlucky. You must be a really unlucky person. Do you lose a lot of games like this?" Afterwards, all participants are then left alone in a room with a full bowl of peanuts for 15 minutes. The average weight of peanuts eaten during this period is compared for the 2 groups.

What is the experiment about?	Snacking behavior, self-esteem
What is the dependent variable?	Weight of peanuts eaten
What is the independent variable?	Whether the participants were made to feel that they were lucky or not
What is the hypothesis or finding about how the IV affects the DV?	Being told they were unlucky would lead to lower self-esteem and increase the number of peanuts eaten
What statistical test would be used to establish a reliable effect?	Two independent samples t-test

If the data were consistent with the hypothesis, the group randomly assigned to lose and be told they were unlucky would have consumed more of the peanuts left with the participants. This result could be statistically reliable

but we might still have concerns about the broader conclusions. We would want to be confident that the experimental manipulation really did affect self-esteem. An **alternate explanation** for the results would be that feeling unlucky leads to greater snacking, without involving perceptions of the self that incorporate self-esteem. The existence of this alternate account for explaining the result does not mean the conclusion is wrong, it simply means that there is more than one way of understanding the data from the experiment and we do not yet know which is correct. These situations are often good opportunities for future research with novel operational definitions of the underlying construct. Note that such an **alternate hypothesis** for the data do not imply the results were not reliable, but that there is a question or limitation about the validity of the conclusion about the constructs. To highlight this different, it is best to separately consider the results of the experiment both in terms of the actual variables (IV, DV) and then the inference in terms of the intended constructs.

We've dived into experimental design and analysis very rapidly here and introduced a fairly large vocabulary of critical terms and concepts very quickly. If that seems daunting, don't worry! We will be going back over the concepts in detail to ensure a solid foundation of methodology design principles across a range of common approaches and research areas. If that seems too easy because design is straightforward, don't worry! While simple designs are easy, it gets complicated fast. If it were really easy, trained and professional researchers wouldn't make mistakes in their research conclusions (spoiler alert: they do).

Exercises

Read Craik & Lockhart (1972) to orient you to the background theory behind our hypothesis for Experiment 1.

It is worth noting that this is a fairly old paper that reflects the theoretical understanding at that time. The “levels of processing” theory is presented as an alternative to “multistore models.” In modern memory research, elements of both theoretical ideas turn out to be true and the two approaches are not seen as inconsistent with each other.

The description and data of the multistore models reflects studies done prior to 1972. It is a useful overview, but if you are interested in the general topic of studies of memory, be aware that is a historical overview from a very long time ago. Characterization of the new ideas related to ‘levels of processing’ comes after this review in the paper.

Answer the following questions from the reading:

1. What is ‘depth of processing’ and why might it lead to better memory?
2. In our study, how would our definition of ‘deep encoding’ connect to this theoretical idea?
3. In our study, how does our definition of ‘shallow encoding’ provide a control comparison?
4. From the prior work cited (e.g., p 677), give an example of how researchers have implemented a different procedure to create shallow encoding.
5. Give another example of a procedure to create deep encoding from the briefly reviewed prior work.

2 Psychological Measurement

Researchers Tara MacDonald and Alanna Martineau were interested in the effect of female university students' moods on their intentions to have unprotected sexual intercourse (MacDonald & Martineau, 2002). In a carefully designed empirical study, they found that being in a negative mood increased intentions to have unprotected sex—but only for students who were low in self-esteem. Although there are many challenges involved in conducting a study like this, one of the primary ones is the measurement of the relevant variables. In this study, the researchers needed to know whether each of their participants had high or low self-esteem, which of course required measuring their self-esteem. They also needed to be sure that their attempt to put people into a negative mood (by having them think negative thoughts) was successful, which required measuring their moods. Finally, they needed to see whether self-esteem and mood were related to participants' intentions to have unprotected sexual intercourse, which required measuring these intentions.

To students who are just getting started in psychological research, the challenge of measuring such variables might seem insurmountable. Is it really possible to measure things as intangible as self-esteem, mood, or an intention to do something? The answer is a resounding yes, and in this chapter, we look closely at the nature of the variables that psychologists study and how they can be measured.

Do You Feel You Are a Person of Worth?

The Rosenberg Self-Esteem Scale (Rosenberg, 1989) is a common measure of self-esteem and the one that MacDonald and Martineau used in their study. The goal of this scale is to take the construct “self-esteem” and turn this into a number that reflects a quantitative measure of a participant’s subjective rating of this idea.

To obtain this measure, participants are asked to respond to each of the 10 items that follow with a rating on a 4-point scale: Strongly Agree, Agree, Disagree, Strongly Disagree.

1. I feel that I’m a person of worth, at least on an equal plane with others.
2. I feel that I have a number of good qualities.
3. All in all, I am inclined to feel that I am a failure.
4. I am able to do things as well as most other people.
5. I feel I do not have much to be proud of.
6. I take a positive attitude toward myself.
7. On the whole, I am satisfied with myself.
8. I wish I could have more respect for myself.
9. I certainly feel useless at times.
10. At times I think I am no good at all.

The responses are then used to calculate a total score based on the responses to each item. A response of Strongly Agree is counted as 3 points, Agree is 2 points, Disagree is 1 point and 0 for Strongly Disagree. Items 1, 2, 4, 6 and 7 are scored this straightforward way. Notice that items 3, 5, 8, 9 and 10 have statements that are conceptually backwards, that is, agreeing reflects less self esteem. For these items we reverse the scoring before calculating the total across all the items. The final number is a value that is higher for participants who have greater self-esteem and we have turned this relatively abstract construct into a quantitative value we can use for scientific research.

In the previous chapter, we introduced the idea of using quantitative measures as variables for experimental design. In this chapter we will focus on the measurement process of creating or identifying these quantitative measures to use in those designs. For a measure like this example, the score could be used in design as a dependent variable. For example, participants could be asked to complete the scale after manipulating an independent variable that was thought to have a temporary effect on self-esteem. We could also use this measure as an independent variable where it would be a special type of independent variable, called a **participant variable**, where we would test a hypothesis about participants with relatively higher or lower self-esteem on some other dependent variable (as in the cited study above).

For a measure like this, we often use a technique called a median split to sort our participants into groups with higher or lower scores.

Participant variables are fairly commonly used and act like independent variables in experimental design and drawing inferences, but should be noted that these cannot be manipulated by the experimenter. They often reflect intrinsic characteristics of the participants that are hypothesized to affect the dependent variable of interest.

Surveys

Surveys are a familiar methodology by which we turn concepts into numbers. While they look deceptively simple to construct, there is a lot of work that goes into establishing that a specific survey is an effective measure of the intended construct.

Developing a new, robust scale that reliably measures a construct is beyond the scope of what can be covered in basic research methods. A overview of this process is provided in Chapter 15. For student research, use of an existing scale from published research is strongly recommended.

Learning Objectives

1. Define measurement and give several examples of measurement in psychology.
2. Explain what a psychological construct is and give several examples.
3. Distinguish conceptual from operational definitions, give examples of each, and create simple operational definitions.
4. Distinguish the four levels of measurement, give examples of each, and explain why this distinction is important.

What Is Measurement?

Measurement is the assignment of scores to individuals so that the scores represent some characteristic of the individuals. This very general definition is consistent with the kinds of measurement that everyone is familiar with—for example, weighing oneself by stepping onto a bathroom scale, or checking the internal temperature of a roasting turkey using a meat thermometer. It is also consistent with measurement in the other sciences. In physics, for example, one might measure the potential energy of an object in Earth's gravitational field by finding its mass and height (which of course requires measuring those variables) and then multiplying them together along with the gravitational acceleration of Earth (9.8 m/s^2). The result of this procedure is a score that represents the object's potential energy.

This general definition of measurement is consistent with measurement in psychology too. Psychological measurement is often referred to as psychometrics. Imagine, for example, that a cognitive psychologist wants to measure a person's working memory capacity—their ability to hold in mind and think about several pieces of information all at the same time. To do this, she might use a backward digit span task, in which she reads a list of two digits to the person and asks them to repeat them in reverse order. She then repeats this several times, increasing the length of the list by one digit each time, until the person makes an error. The length of the longest list for which the person responds correctly is the score and represents their working

memory capacity. Or imagine a clinical psychologist who is interested in how depressed a person is. He administers the Beck Depression Inventory, which is a 21-item self-report questionnaire in which the person rates the extent to which they have felt sad, lost energy, and experienced other symptoms of depression over the past 2 weeks. The sum of these 21 ratings is the score and represents the person's current level of depression.

The important point here is that measurement requires some systematic procedure for assigning scores to individuals or objects so that those scores represent the characteristic of interest.

Psychological Constructs

Many variables studied by psychologists are straightforward and simple to measure. These include age, height, weight, and birth order. You can ask people how old they are and be reasonably sure that they know and will tell you. Although people might not know or want to tell you how much they weigh, you can have them step onto a bathroom scale. Other variables studied by psychologists—perhaps the majority—are not so straightforward or simple to measure. We cannot accurately assess people's level of intelligence by looking at them, and we certainly cannot put their self-esteem on a bathroom scale. These kinds of variables are called constructs (pronounced CON-structs) and include personality traits (e.g., extraversion), emotional states (e.g., fear), attitudes (e.g., toward taxes), and abilities (e.g., athleticism).

Psychological constructs cannot be observed directly. One reason is that they often represent tendencies to think, feel, or act in certain ways. For example, to say that a particular university student is highly extroverted does not necessarily mean that she is behaving in an extroverted way right now. In fact, she might be sitting quietly by herself, reading a book. Instead, it means that she has a general tendency to behave in extroverted ways (e.g., being outgoing, enjoying social interactions) across a variety of situations. Another reason psychological constructs cannot be observed directly is that they often

involve internal processes. Fear, for example, involves the activation of certain central and peripheral nervous system structures, along with certain kinds of thoughts, feelings, and behaviors—none of which is necessarily obvious to an outside observer. Notice also that neither extroversion nor fear “reduces to” any particular thought, feeling, act, or physiological structure or process. Instead, each is a kind of summary of a complex set of behaviors and internal processes.

Ethics

By diving straight into experimental design, we have taken on the idea of measurement without establishing a foundation for how researchers need to consider ethical aspects of psychological science.

There are many measures of individuals that are invasive of the *privacy* of the participant in research. These must be administered with great care for the rights of participants in human research. Questions about constructs related to mental health are often relevant to scientific hypotheses in psychological science. These questions must be administered within the framework of ethical research and with oversight from the Institutional Review Board.

Chapter 8 will review ethical research procedures in detail.

Conceptually Defining the Construct

Having a clear and complete conceptual definition of a construct is a prerequisite for good measurement. For one thing, it allows you to make sound decisions about exactly how to measure the construct. If you had only a vague idea that you wanted to measure people's "memory," for example, you would have no way to choose whether you should have them remember a list of vocabulary words, a set of photographs, a newly learned skill, an experience from long ago, or have them remember to perform a task at a later time. Because psychologists now conceptualize memory as a set of semi-independent systems, you would have to be more precise about what you mean by "memory." If you are interested in long-term episodic memory (memory for previous experiences), then having participants remember a list of words that they learned last week would make sense, but having them try to remember to execute a task in the future would not. In general, there is no substitute for reading the research literature on a construct and paying close attention to how others have defined it.

Example: Personality and The Big Five

The Big Five is a set of five broad dimensions that capture much of the variation in human personality. Each of the Big Five can even be defined in terms of six more specific constructs called "facets" (Costa & McCrae, 1992): Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism.

The conceptual definition of a psychological construct describes the behaviors and internal processes that make up that construct, along with how it relates to other variables. For example, a conceptual definition of neuroticism (another one of the Big Five) would be that it is people's tendency to experience negative emotions such as anxiety, anger, and sadness across a variety of situations. This definition might also include that it is hypothesized to have a strong genetic component, is required to remain fairly stable over time, and is positively correlated with other types of measurements, such as

the tendency to experience pain and other physical symptoms.

Students sometimes wonder why, when researchers want to understand a construct like self-esteem or neuroticism, they do not simply look it up in the dictionary. One reason is that many scientific constructs do not have counterparts in everyday language (e.g., working memory capacity). More important, researchers are in the business of developing definitions that are more detailed and precise—and that more accurately describe the way the world is—than the informal definitions in the dictionary. As we will see, they do this by proposing conceptual definitions, testing them empirically, and revising them as necessary. Sometimes they throw them out altogether. This is why the research literature often includes different conceptual definitions of the same construct. In some cases, an older conceptual definition has been replaced by a newer one that fits and works better. In others, researchers are still in the process of deciding which of various conceptual definitions is the best.

Operational Definitions

Once you have a conceptual definition of the construct you are interested in studying, it is time to operationally define the construct. Recall an operational definition is a definition of the variable in terms of precisely how it is to be measured. Since most variables are relatively abstract concepts that cannot be directly observed (e.g., stress), and observation is at the heart of the scientific method, conceptual definitions must be transformed into something that can be directly observed and measured. Most variables can be operationally defined in many different ways. For example, stress can be operationally defined as people's scores on a stress scale such as the Perceived Stress Scale (Cohen, Kamarck, & Mermelstein, 1983), cortisol concentrations in their saliva, or the number of stressful life events they have recently experienced. As described below, operationally defining your variable(s) of interest may involve using an existing measure or creating your own measure.

An operational definition is a definition of a variable in terms of precisely how it is to be measured. These measures generally fall into one of three broad categories. Self-report measures are those in which participants report on their own thoughts, feelings, and actions, as with the Rosenberg Self-Esteem Scale (Rosenberg, 1965). Behavioral measures are those in which some other aspect of participants' behavior is observed and recorded. This is an extremely broad category that includes the observation of people's behavior both in highly structured laboratory tasks and in more natural settings. A good example of the former would be measuring working memory capacity using the backward digit span task. A good example of the latter is a famous operational definition of physical aggression from researcher Albert Bandura and his colleagues (Bandura, Ross, & Ross, 1961). They let each of several children play for 20 minutes in a room that contained a clown-shaped punching bag called a Bobo doll. They filmed each child and counted the number of acts of physical aggression the child committed. These included hitting the doll with a mallet, punching it, and kicking it. Their operational definition, then, was the number of these specifically defined acts that the child committed during the 20-minute period. Finally, physiological measures are those that involve recording any of a wide variety of physiological processes, including heart rate and blood pressure, galvanic skin response, hormone levels, and electrical activity and blood flow in the brain.

For any given variable or construct, there will be multiple operational definitions. Stress is a good example. A rough conceptual definition is that stress is an adaptive response to a perceived danger or threat that involves physiological, cognitive, affective, and behavioral components. But researchers have operationally defined it in several ways. The Social Readjustment Rating Scale (Holmes & Rahe, 1967) is a self-report questionnaire on which people identify stressful events that they have experienced in the past year and assigns points for each one depending on its severity. For example, a man who has been divorced (73 points), changed jobs (36 points), and had a change in sleeping habits (16 points) in the past year would have a total score of 125. The Hassles and Uplifts Scale (Delongis, Coyne, Dakof, Folkman & Lazarus, 1982) is similar but focuses on

everyday stressors like misplacing things and being concerned about one's weight. The Perceived Stress Scale (Cohen, Kamarck, & Mermelstein, 1983) is another self-report measure that focuses on people's feelings of stress (e.g., "How often have you felt nervous and stressed?"). Researchers have also operationally defined stress in terms of several physiological variables including blood pressure and levels of the stress hormone cortisol.

When psychologists use multiple operational definitions of the same construct—either within a study or across studies—they are using converging operations. The idea is that the various operational definitions are “converging” or coming together on the same construct. When scores based on several different operational definitions are closely related to each other and produce similar patterns of results, this constitutes good evidence that the construct is being measured effectively and that it is useful. The various measures of stress, for example, are all correlated with each other and have all been shown to be correlated with other variables such as immune system functioning (also measured in a variety of ways) (Segerstrom & Miller, 2004). This is what allows researchers eventually to draw useful general conclusions, such as “stress is negatively correlated with immune system functioning,” as opposed to more specific and less useful ones, such as “people’s scores on the Perceived Stress Scale are negatively correlated with their white blood counts.”

Experiment 1

For example, in the in-class experiment, we measured ‘memory’ by score on a recognition test where you saw a list of words and for each one responded whether you had seen it before. This produces a numerical measure of memory in the number of correct answers. However, it is fair to also say that there are lot of other ways to think about memory. Memory can refer to being able to recount the events of an experience you had yesterday. Another common way to measure memory is via tests of recall, e.g., asking participants to report all the words they had seen during the original study phase. This would also produce a quantitative measure of memory for

the word list. In more advanced memory research, there are theoretical questions about how recognition and recall memory may be influenced by underlying mechanisms that might be specific to those processes. Recalling words seems to depend on something like “searching” our memories that might not be part of the process of deciding if you recognize a word seen before.

It would also be fair to say that any measure of memory for a list of arbitrary, unrelated words fails to capture important ideas that people are interested in that relate to the concept of “memory.” One of the most common complaints about memory is memory failures, such as the challenging issue of remembering somebody’s name after you meet them. People will also have the experience of walking into a room and forgetting why you went into the room, which is also described as a failure of memory. Understanding factors that affect memory for lists of words may inform our understanding of these kinds of memory failures, but the distance from the operational definition employed in our experiment to those applications should be noted in considering the meaning of our findings.

All forms of science employ measurement, but the idea of the distance from the operational definition to the underlying concept is somewhat unique to psychological science. In other areas like biology, chemistry, or physics it is more commonly the case that there is less debate about what is being measured exactly. Because psychology is the science of people, we have the advantage of intuition and a basic understanding of the high-level concepts. We all know what words like ‘memory’ or ‘anxiety’ mean. However, when we design experiments or read about others’ experimental work, we need to identify more precise definitions that turns these conceptual ideas into numbers. This also highlights the complexity of a word like “memory” and the associated challenge of indicating exactly what aspect of memory is being incorporated into the operational definition. This complexity is also why much modern psychological research uses increasingly specific and precise terminology to capture sub-areas of interest. For example, if you are interested in research aimed at understanding the phenomenon of

forgetting why you walked into a room, you will want to look for research on “prospective memory,” which is built around operational definitions based on memory for intentions to carry out actions and when that process surprisingly fails.

The process of establishing operational definitions applies to the process of setting up both the independent and dependent variables for a study. Many of the terms used to describe the key ideas in “measurement” apply more obviously to the dependent variable. For our basic experimental design, we expect the dependent variable to be a measured operational definition, which is a quantitative number that changes in a direction that can be conceptually connected to the construct. For our Experiment 1, more words recognized is clearly associated with more memory. It is also fine to consider measures that move the other direction, such as a measure like reaction time (speed to make a response) which tends to go down as a reflection of more knowledge. In communication about research, it is necessary to be clear about the details of the type and direction used for measurement.

Levels of Measurement

The psychologist S. S. Stevens suggested that scores can be assigned to individuals in a way that communicates more or less quantitative information about the variable of interest (Stevens, 1946). For example, the officials at a 100-m race could simply rank order the runners as they crossed the finish line (first, second, etc.), or they could time each runner to the nearest tenth of a second using a stopwatch (11.5 s, 12.1 s, etc.). In either case, they would be measuring the runners’ times by systematically assigning scores to represent those times. But while the rank ordering procedure communicates the fact that the second-place runner took longer to finish than the first-place finisher, the stopwatch procedure also communicates how much longer the second-place finisher took. Stevens actually suggested four different levels of measurement (which he called “scales of measurement”) that correspond to four types of information that can be communicated by a set of scores, and

the statistical procedures that can be used with the information.

The **nominal** level of measurement is used for categorical variables and involves assigning scores that are category labels. Category labels communicate whether any two individuals are the same or different in terms of the variable being measured. For example, if you ask your participants about their marital status, you are engaged in nominal-level measurement. Or if you ask your participants to indicate which of several ethnicities they identify themselves with, you are again engaged in nominal-level measurement. The essential point about nominal scales is that they do not imply any ordering among the responses. For example, when classifying people according to their favorite color, there is no sense in which green is placed “ahead of” blue. Responses are merely categorized. Nominal scales thus embody the lowest level of measurement.

The remaining three levels of measurement are used for quantitative variables. The **ordinal** level of measurement involves assigning scores so that they represent the rank order of the individuals. Ranks communicate not only whether any two individuals are the same or different in terms of the variable being measured but also whether one individual is higher or lower on that variable. For example, a researcher wishing to measure consumers’ satisfaction with their microwave ovens might ask them to specify their feelings as either “very dissatisfied,” “somewhat dissatisfied,” “somewhat satisfied,” or “very satisfied.” The items in this scale are ordered, ranging from least to most satisfied. This is what distinguishes ordinal from nominal scales. Unlike nominal scales, ordinal scales allow comparisons of the degree to which two individuals rate the variable. For example, our satisfaction ordering makes it meaningful to assert that one person is more satisfied than another with their microwave ovens. Such an assertion reflects the first person’s use of a verbal label that comes later in the list than the label chosen by the second person.

On the other hand, ordinal scales fail to capture important information that will be present in the other levels of measurement we examine. In particular, the difference between two levels of an ordinal scale cannot be assumed to

be the same as the difference between two other levels (just like you cannot assume that the gap between the runners in first and second place is equal to the gap between the runners in second and third place). In our satisfaction scale, for example, the difference between the responses “very dissatisfied” and “somewhat dissatisfied” is probably not equivalent to the difference between “somewhat dissatisfied” and “somewhat satisfied.” Nothing in our measurement procedure allows us to determine whether the two differences reflect the same difference in psychological satisfaction. Statisticians express this point by saying that the differences between adjacent scale values do not necessarily represent equal intervals on the underlying scale giving rise to the measurements. (In our case, the underlying scale is the true feeling of satisfaction, which we are trying to measure.)

The **interval** level of measurement involves assigning scores using numerical scales in which intervals have the same interpretation throughout. As an example, consider either the Fahrenheit or Celsius temperature scales. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees. This is because each 10-degree interval has the same physical meaning (in terms of the kinetic energy of molecules).

Interval scales are not perfect, however. In particular, they do not have a true zero point even if one of the scaled values happens to carry the name “zero.” The Fahrenheit scale illustrates the issue. Zero degrees Fahrenheit does not represent the complete absence of temperature (the absence of any molecular kinetic energy). In reality, the label “zero” is applied to its temperature for quite accidental reasons connected to the history of temperature measurement. Since an interval scale has no true zero point, it does not make sense to compute ratios of temperatures. For example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees; no interesting physical property is preserved across the two ratios. After all, if the “zero” label were applied at the temperature that Fahrenheit happens to label as 10 degrees, the two ratios would instead be 30 to 10 and 90 to 40, no longer the same! For this

reason, it does not make sense to say that 80 degrees is “twice as hot” as 40 degrees. Such a claim would depend on an arbitrary decision about where to “start” the temperature scale, namely, what temperature to call zero (whereas the claim is intended to make a more fundamental assertion about the underlying physical reality).

In psychology, the intelligence quotient (IQ) is often considered to be measured at the interval level. While it is technically possible to receive a score of 0 on an IQ test, such a score would not indicate the complete absence of IQ. Moreover, a person with an IQ score of 140 does not have twice the IQ of a person with a score of 70. However, the difference between IQ scores of 80 and 100 is the same as the difference between IQ scores of 120 and 140.

Finally, the **ratio** level of measurement involves assigning scores in such a way that there is a true zero point that represents the complete absence of the quantity. Height measured in meters and weight measured in kilograms are good examples. So are counts of discrete objects or events such as the number of siblings one has or the number of questions a student answers correctly on an exam. You can think of a ratio scale as the three earlier scales rolled up in one. Like a nominal scale, it provides a name or category for each object (the numbers serve as labels). Like an ordinal scale, the objects are ordered (in terms of the ordering of the numbers). Like an interval scale, the same difference at two places on the scale has the same meaning. However, in addition, the same ratio at two places on the scale also carries the same meaning (see Table 4.1).

The Fahrenheit scale for temperature has an arbitrary zero point and is therefore not a ratio scale. However, zero on the Kelvin scale is absolute zero. This makes the Kelvin scale a ratio scale. For example, if one temperature is twice as high as another as measured on the Kelvin scale, then it has twice the kinetic energy of the other temperature.

Another example of a ratio scale is the amount of money you have in your pocket right now (25 cents, 50 cents, etc.). Money is measured on a ratio

scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this actually implies the absence of money. Since money has a true zero point, it makes sense to say that someone with 50 cents has twice as much money as someone with 25 cents.

Levels of measurement are important for at least two reasons. First, they emphasize the generality of the concept of measurement. Although people do not normally think of categorizing or ranking individuals as measurement, in fact, they are as long as they are done so that they represent some characteristic of the individuals. Second, the levels of measurement can serve as a rough guide to the statistical procedures that can be used with the data and the conclusions that can be drawn from them. With nominal-

level measurement, for example, the only available measure of central tendency is the mode. With ordinal-level measurement, the median or mode can be used as indicators of central tendency. Interval and ratio-level measurement are typically considered the most desirable because they permit for any indicators of central tendency to be computed (i.e., mean, median, or mode). Also, ratio-level measurement is the only level that allows meaningful statements about ratios of scores. Once again, one cannot say that someone with an IQ of 140 is twice as intelligent as someone with an IQ of 70 because IQ is measured at the interval level, but one can say that someone with six siblings has twice as many as someone with three because number of siblings is measured at the ratio level.

Data Analysis

Measures that vary in the type and levels will also determine what kind of statistical approach is correct to use for a specific design.

Chapter 5, 12 and 16 will review different statistical methods for cases where there are IV's and DV's that have different levels of measurement.

The most common type of experimental design is to use an interval or ratio measure for the DV and a nominal measure for the IV. Experiment 1 is designed this way.

Reliability and Validity of Operational Definitions

Developing a novel measure of a construct that consistently and accurately numerically captures a complex construct is a complex and time-consuming task. We will discuss the general methodology for this later (Chapter 17, Surveys and Instrument Design) since this process is more often engaged with as part of non-experimental research than experimental research and is also generally outside the scope of this introductory class on psychological science. However, drawing inferences about experimental data will require considering how well the operational definition captures the underlying construct. Misalignment between the operational definition and the construct can lead to problems with inferences about the construct or can limit the applicability of findings to contexts outside the laboratory.

In the context of measurement, reliability refers to how consistently the measure obtains an accurate assessment of the underlying construct. For example, in personality research, characteristics such as 'conscientiousness' are expected to be stable individual traits over time. That means that subsequent attempts to measure the trait should generally produce the same number. However, data collected from human participants is virtually never perfectly stable for a wide variety of reasons. Participants might have external or internal distractions while engaged with a measure, or might have state-level effects (e.g., tiredness or hunger) that unexpectedly influence the score obtained. Everything that influences our measure that is unrelated to the construct creates measurement error, which shows up in our experimental data as a contribution to the observed variance in performance. We will discuss methodological techniques for managing measurement error as best we can in Chapters 3 and 4, but even with best practices, there will always be some component of "noise" in our data (also important for our statistical approach, Chapter 5).

Another key aspect of an effective measured operational definition is its validity in capturing the underlying construct. Robust techniques for establishing validity of a novel measure are complex (Chapter 17) but a

simpler key version of the issue is seen as the face validity of a measure. Face validity is one that can often be evaluated intuitively and is simply a question of whether the measure actually relates to the underlying construct. If we were to claim that our Experiment 1 recognition memory measure is a measure of how likely you are to forget why you walked into the kitchen, we would lack face validity and this level of inference about our data should not be trusted. In contrast, if we claimed that our measure was relevant for understanding how students could build better memory for studying material in the classroom, we would have better face validity (but not perfect and examples of where there might be a disconnect is left as an exercise for the reader).

Intelligence

To use a fairly controversial example, the IQ scale is an operational definition of the concept of *intelligence*, but there is no broad consensus of what exactly *intelligence* is. The IQ scale clearly measures something that has robust correlations with measures like academic success. However, whether there is a single underlying construct that is *intelligence* continues to be hotly debated. One alternative idea is that there are multiple *types of intelligence* that might be best measured separately. Obtaining data that argues for single or multiple types of intelligence turns out to be extremely challenging

Key Takeaways and Exercises

- Measurement is the assignment of scores to individuals so that the scores represent some characteristic of the individuals. Psychological measurement can be achieved in a wide variety of ways, including self-report, behavioral, and physiological measures.
- Psychological constructs such as intelligence, self-esteem, and depression are variables that are not directly observable because they represent behavioral tendencies or complex patterns of behavior and internal processes. An important goal of scientific research is to conceptually define psychological constructs in ways that accurately describe them.
- For any conceptual definition of a construct, there will be many different operational definitions or ways of measuring it. The use of multiple operational definitions, or converging operations, is a common strategy in psychological research.
- Variables can be measured at four different levels—nominal, ordinal, interval, and ratio—that communicate increasing amounts of quantitative information. The level of measurement affects the kinds of statistics you can use and conclusions you can draw from your data.
- Psychological researchers do not simply assume that their measures work. Instead, they conduct research to show that they work. If they cannot show that they work, they stop using them.
- There are two distinct criteria by which researchers evaluate their measures: reliability and validity. Reliability is consistency across time (test-retest reliability), across items (internal consistency), and across researchers (interrater reliability). Validity is the extent to which the scores actually represent the variable they are intended to.
- Good measurement begins with a clear conceptual definition of the construct to be measured. This is accomplished both by clear and detailed thinking and by a review of the research literature.

Exercises

For practice thinking through the process of creating operational definitions, considering the following 3 common sayings. For each, provide an example of how you might operationally define (a) an independent variable, (b) a dependent variable, and (c) state the direction in which the IV is hypothesized to affect the DV.

1. People feel sadder in blue rooms than in pink rooms
2. It takes longer to recognize a person in a photograph seen upside down
3. Absence makes the heart grow fonder

Additional optional questions

- Practice: Complete the Rosenberg Self-Esteem Scale and compute your overall score.
- Practice: Think of three operational definitions for sexual jealousy, decisiveness, and social anxiety. Consider the possibility of self-report, behavioral, and physiological measures. Be as precise as you can.
- Practice: For each of the following variables, decide which level of measurement is being used.
 - A university instructor measures the time it takes her students to finish an exam by looking through the stack of exams at the end. She assigns the one on the bottom a score of 1, the one on top of that a 2, and so on.
 - A researcher accesses her participants' medical records and counts the number of times they have seen a doctor in the past year.
 - Participants in a research study are asked whether they are right-handed or left-handed.
- Discussion: Think back to the last college exam you took and think of the exam as a psychological measure. What construct do you think it was intended to measure? Comment on its face and content validity. What data could you collect to assess its reliability and criterion validity?

3 Experimental Control

We have described the process of setting up an experimental design as starting with the high-level constructs and then implementing operational definitions that allow us to create an experimental procedure that assesses the effect of an independent variable on a dependent variable. The previous chapter discussed some aspects of creating a measured operational definition that can be used as the dependent variable. In this chapter, we will consider issues and methods for creating effective independent variables that will allow us to draw strong conclusions from our research studies.

The goal of experimental design is to be able to draw strong, valid conclusions from the results of our studies. However, to illustrate the first step in design, we can take the perspective of reading a scientific report that we are led to through a social media link with the tag line

Listening to music improves scores on stressful classroom tests

The conclusion here is a causal statement, listening to music causes better test performance. That is an interesting and possibly useful statement and should make us immediately very curious about the design and the operational definitions used. We will want to know what the operational definition of the stressful classroom test is, but also what was used for music. One of the first questions we should ask is: **Compared to what?**

The conclusion statement attributes an effect to music, but as is often the case when we have only the conclusion, we do not know what music is being compared to. To really understand the result, we will likely have to find the original scientific report, read the procedure section and see how they describe their **control condition**. In this case, the control condition would result from the operational definition of the independent variable. We would expect to have a description of the condition in which participants were exposed to music and in that section also a description of the condition in which there was no music.

Returning to the perspective of designing a study, we might start with a similar hypothesis such as:

Music helps me study. Notice that this is also a causal statement that just like the tag line, leaves out a lot of information about the independent and dependent variable. However, since we are thinking of designing a study to test our hypothesis, we have to come up with the operational definitions for these variables. Doing it well means that if we obtain a reliable effect in data we collect in our study, we will be able to confidently assert that our hypothesis was correct. In this Chapter we will discuss principles of design that aim to guide us to doing it well.

First we will have to establish

The Mozart Effect

Studies of the effect of music on cognition are often referred to as related to a hypothesized Mozart effect. This has essentially become a colloquial term for an old idea that classical music has some benefits for general cognitive processes. That specific idea has not replicated effectively but more modern studies have found small but reliable effects of ambient music on specifically spatial cognition tasks and may also be helpfully calming in some stressful environments.

what we mean by *music* and also what we will use as the *non-music* contrast. *Music* could refer to a specific piece of classical music, or perhaps popular music preferred by the participants. The control condition could be silence, or could be white noise, or soothing natural sounds. There is not one right answer to how to pick a correct operational definition here. Most of the alternatives are different, equally interesting studies.

Once we select the music and non-music conditions, we will be planning to collect data from a group of participants. This is when we need to consider the question of experimental control. When we do our data collection is there anything else going on around the participants that might affect how we administer the conditions or measure our DV? Is there other noise around? Might we get interrupted? How consistent and reliable is the technique for playing the music?

Anything not part of our planned design that affects our study is referred to as an **extraneous variable**. Experimental control is fundamentally about identifying as many extraneous variables as possible and designing our study to minimize the effects of these.

These uncontrolled extraneous variables can affect the quality of our research process in two main ways. Random external influences that interfere with carrying out the research procedure will tend to produce **reliability** problems. Technically, these increase the variance in the average scores on our dependent variable which makes it more difficult to obtain statistical confidence that our IV affects the DV. For example, if there are interruptions occurring randomly for some of the participants in the study, their DV scores might be higher or lower because of the interruptions rather than our control over the music being played.

That is a problem for our research, but an even greater problem occurs when there is an extraneous variable that systematically varies with our IV. This creates an **experimental confound**, which is a serious problem with the **validity** of any conclusions we would like to draw from our research.

Confounds are going to be the most important problem we are concerned with

when evaluating the validity of an experiment because they essentially ruin the ability to get a confident conclusion. As an example from our hypothetical music design, if we used different audio volume for our music and white noise stimuli with the white noise being painfully loud, we would see an effect on studying but it would not be a positive effect of music, it would be a negative effect of loud distracting white noise. Technically, the audio volume here is an extraneous variable and since it varies with the IV by always being louder for the white noise, it has confounded our experiment.

More commonly, extraneous variables are related to things like the testing context, whether there are distractions around, details of the technology or even things like time of day. These aspects can affect our data collection but if they are random and occur for both conditions roughly equally, they do not confound our study. In the laboratory environment, where we have reasonable control over the environment, there are two main techniques to managing these extraneous variables: **constancy** and **counter-balancing**.

In this chapter, we will introduce the basics of setting up our IV in experimental design with consideration of all the extraneous variables we can identify. We will review the particular danger of confounds and the impact of these variables on the reliability and/or validity of experimental design. In Chapter 4, we will extend this discussion to the specific challenges of things that differ across the participants in our study and the basic approach to establishing rigor through a well-planned experimental procedure.

Learning Objectives

1. Understanding the independent variable in experimental design
2. Define what a control condition is, explain its purpose in research on treatment effectiveness, and describe some alternative types of control conditions.
3. How to construct two treatment conditions as the IV for a study and how this might be extended to more complex designs
4. Extraneous variables: factors that affect the DV that were not part of the experimental design
5. Managing extraneous variables via constancy: keeping things as consistent as possible across levels of the independent variable
6. Managing extraneous variables via counterbalancing: if factors cannot be kept constant, distribute them evenly across the independent variable
7. Confounds: variables that reduce the internal validity of an experiment

Treatment and Control Conditions

For many people, their introduction to psychological research is by seeing findings reported from a study designed around having a **treatment** condition and a **control condition**. These kinds of studies are essentially **intervention** studies where something is done to improve things or fix a problem. This kind of study is very common in health and medical research as well and we will use those domains for examples of design even though the main application will be studies of psychological constructs. We will return in Chapter 17 to questions specific to intervention research, particularly implementation and ethical issues.

The music example above can be seen as a kind of intervention study as a test of the hypothesis that *music helps with studying*. Music is the intervention in question and improved studying would be the benefit. The use of the example is to focus our attention on the control condition that must be present in the design in order for there to be reliable scientific evidence for

the statement. In our above example, we considered using white noise as the comparison to music. As a result, our more accurate statement of the results of testing the hypothesis would be *music helps with studying compared to white noise.*

The same style of thinking should be applied to other hypotheses like *daily meditation helps with focus in the classroom* or *exercise helps to sleep better*. Those statements also have hidden control conditions to which an intervention was compared. When reading research, we look in the sections of the scientific report where the procedure is described to find the operational definition of the control condition. When designing research, we consider the inference we are hoping to draw and the best comparison condition to establish our conclusion.

The rest of the basic design for this kind of study is very simple. We administer the treatment to one group of participants, usually half of the total group. We administer the control condition to everybody else. Then we assess the dependent variable and see if the treatment affected those scores.

Of course, this approach being simple and effective depends on solving the measurement problems that were the main topic of Chapter 2 as well as

_____ with this one simple trick!

In the public sphere, interactions with findings from science often come packaged as statements like this where the blank could be filled in with lose weight, make more friends, or get better grades. If there is any science behind the statement it will be based on research that uses an intervention design. If it is good science, there will be an appropriate control condition, good sample size, no confounding variables, statistical reliability and a consistent rigorous study procedure. If not, then it is just click bait.

the experimental control questions which are the focus here of Chapter 3.

Before we generalize from treatment and control to more general designs with two levels of an independent variable, it is worth noting one specific design element associated with these kinds of designs, the **placebo** effect. Intervention research often needs to explicitly discuss the possibility of the control condition inadvertently affecting the dependent variable through the expectations of the participants in the research study. We will return to this idea in Chapter 4 under the general consideration of **demand characteristics** embedded in experimental design and standard approaches to controlling these.

Independent Variable with Two Levels

The basic framework we are using to examine research methodology in psychological science is to quantify how manipulation of an **independent variable** affects measurement of a **dependent variable**. Using treatment studies as an example, we see that at a minimum, our independent variable needs to have two conditions which we will more generally refer to as **levels** of the variable.

Our Experiment 1 illustrates the more general approach. To assess how deeper semantic processing of words leads to better memory for those words, two levels of depth of processing are compared. The deep condition, where words were rated by the participants as how much they liked them, is designed to have more depth than the comparison shallow encoding condition of counting vowels. Rather than thinking of the design as based on a treatment and control, we describe the independent variable as being an operational definition of depth of processing and use a procedure that has two different levels of depth.

Choosing an effective control condition is not always a straightforward process as the earlier music example showed. Research in behavioral health measures aimed at interventions such as physical activity to improve

health outcomes is an area famous for the difficulty in constructing control conditions. As a simple example, a straightforward intervention is to ask participants to add exercise to their daily routine to improve cardiovascular health. When the control condition is simply not adding exercise, it is important to also understand what activities are potentially being replaced by exercise. If the control group is not just sedentary but also engaged in unhealthy eating behaviors that might have been replaced by exercise, the intervention may not work in the manner hypothesized. That is, it may improve health by reducing unhealthy eating instead of via a direct effect of exercise on the body. Since we usually want to understand why our study worked, we might prefer using two levels of the independent variable instead. Here that might be comparing a high level of exercise, like running for 30 minutes, with a lower level of exercise, like stretching for 30 minutes.

To be clear, the use of terms like treatment or levels of the independent variable is merely a matter of terminology. There is no conceptual or design difference between descriptions that prefer one set of terms versus the other. There is even an additional synonym for the manipulated variables of an experiment where these are referred to as **factors** in design. Which of these terms gets used reflects customs in different sub-areas of science. The term factors is typically used in descriptions of more complex designs than we will start with in this chapter. When there are multiple independent variables being manipulated by the experimenter, these are described as **factorial design**, a topic discussed in great depth in Chapters 10 & 12.

When more than two levels of a single independent variable are contrasted, this design may also be called a factorial design. An independent variable with three levels adds only a little complexity to questions of experimental control. However, it adds substantially to the problem of drawing inferences from the data, partly due to needing more complex tools for statistical analysis. In addition, three levels implies at least three key comparisons that will need to be evaluated and interpreted. If there are conditions A, B and C in a design, we have to contrast A versus B, B versus C and A versus C, each with their own statistical contrast and conclusions to be drawn about reliable

differences found. Our discussions of design will focus on just two levels of the independent variable to keep this aspect of design and interpretation simple to start.

For the basics of experimental control, we will also temporarily hold aside the additional complexity brought on by designs in which participants each experience all levels of the independent variable. These designs, called **within-participant**, can also have an independent variable with two levels but all participants get both levels. Experiment 1 could have been administered this way by asking participants to rate some words for liking and count vowels for other words. In Chapter 7, we will discuss the many strengths of this approach to research and techniques for managing the additional challenges created by considering the possibility of order effects in administering the independent variable.

If everything else about experimental control goes as planned, differences in the scores on the dependent variable can only have arisen from the different experience of the participants across the levels of the independent variable. That is what allows us to make causal statements that the different levels of the independent variable caused different scores on the dependent variable. However getting everything to go as planned is not as simple as it looks and requires a research procedure designed with the necessary experimental control.

Extraneous variables

Virtually everything we need to worry about with respect to planning an effective experimental design aimed to produce reliable outcomes and valid conclusions boils down to identifying and managing **extraneous variables**.

Extraneous variables reflect anything going on around or during the experiment that could affect scores on the dependent variable that were not part of the experimenter's design. In most simple experimental designs, we will be planning to collect data on samples of participants roughly including

20 to 60 people. The conditions under which these participants are in the experiment will vary across those data collection points. They could vary in time of day, location, distractions around them, etc. If participants are completing an experiment through an online system, they could be interacting with the system from a variety of devices, different internet connectivity strengths, different interfaces like touchscreen or keyboard.

We will rely heavily on intuition and our knowledge of psychology as people to identify as many relevant extraneous variable as possible for the designs we consider. As you develop specific knowledge of research within a specific area of psychology, you will learn about the variables that commonly have to be considered in that domain. It can be tricky to find these in research areas that we are less familiar with. Since this class is about general design, we will tend to not focus in detail on domain-specific details. As an example, unless you are familiar with memory research with word lists, you are likely unaware that characteristics of words themselves influence memory for those words. Words that are very uncommon, termed *low frequency*, tend to be much more memorable for designs like the one we used in Experiment 1. We will see that the specific words in that study were selected with awareness of this extraneous variable and designed to minimize the potential impact of this aspect of the stimuli.

There are always a large number of possible extraneous variables implicit in any experimental design. Our general approach to evaluating designs will be based on reviewing the described procedure and then brainstorming as many of these as we can come up with. Then evaluate which are likely to be affecting measures in the study, likely discarding most of the candidates we consider. If we identify a plausible variable that seems likely to have a substantial impact on the dependent variable measure, we will go back to the described procedure and verify that this element was handled effectively. If not, there may be a problem with the conclusions drawn from the study.

We use the same process for developing a new experimental design except that as experimenters, we are responsible for putting together the procedure for the planned design and handling these variables. We will discuss two

basic approaches to keeping these variables from weakening our research methodology here. It may be worth mentioning here that a very useful approach to the problem of identifying all the possible extraneous variables in a proposed research design is to follow many of the procedure elements in a published scientific study. Successful studies have generally demonstrated how to avoid the problems of extraneous variables. Later, when we discuss creating research proposals for potential class projects, the recommended technique will be to take an existing, working design and add one new thing to it to help increase the probability of a successful result.

In this chapter we are focusing on extraneous variables associated with the environment, procedure and stimuli. There are also clearly a lot of aspects related to the participants in the study that can affect the dependent variable. For our Experiment 1, we might hypothesize that some participants are better at memory for words, some were better able to pay attention during study, some were stressed or had not slept well the night before. These **participant variables**, which often reflect **individual differences** between people, are managed slightly differently and will be discussed in Chapter 4. Conceptually, these concerns influence the accuracy of experimental design in the same manner as other extraneous variables.

We will consider the impact of these extraneous variables as they affect the **reliability** and **validity** of our experimental design. Note that these terms here are being used in their formal, technical sense. Reliability specifically refers to the statistical evaluation of experimental results. That is, the data indicate that we have sufficient confidence the results could not have happened by chance. This is equivalent to saying we believe the result will **replicate** if the experiment were carried out again. Validity specifically refers to whether the inference is accurate that the independent variable affected the dependent variable. It might seem like these two attributes would tend to go together, but they do not always. We can have an experimental result that is reliable, but not valid. And likewise, a valid hypothesis for which the data is not reliable in a study.

Errors arising from design mistakes

One set of terminology used in describing experimental errors is to describe errors as either **Type 1** or **Type 2**. It is more important to understand the implications of these errors and how to spot them than be specifically familiar with the terminology, although the short hand terms are frequently helpful.

A **Type 1** error is a **false positive** claim where we believe the independent variable has causally influenced the dependent variable, but the claim is wrong. This happens when there is a failure of **validity**, often also described as an error of **internal validity**. This error is a serious problem for science since it asserts an incorrect claim, which if believed can cause damage when people rely on the claim to influence their behavior subsequently. Much of what we do to establish a rigorous experimental design is aimed to minimize the possibility of this kind of error. In consideration of extraneous variables, the key problem we aim to avoid is having a **confound** in our study, a **confounded variable** with our independent variable (see below).

A **Type 2** error is a **false negative** in which our study does not reliably support a claim about the independent variable affecting the dependent variable. This generally reflects a failure in **reliability**, which we will interpret mainly on statistical grounds. This happens when an experiment does not appear to work, that is, across the two levels of the independent variable, the differences in the measure of the dependent variable were not large enough support a claim of a finding. This is an error when it turns out later that the independent variable normally does affect the dependent variable but the particular study carried out did not observe the typical effect. That can happen simply due to poor luck related to happening to observe higher levels of variability in performance than usual. It can also happen when there are a lot of uncontrolled extraneous variables that were not managed properly in designing the experiment.

In general, we prefer a Type 2 error to a Type 1 error since a lack of statistical reliability will often lead to results not being published and therefore there is no danger of people relying on an inaccurate scientific finding. Note

that this is different from concluding that the independent variable never affects the dependent variable, which would be a true negative. As we will see in Chapter 5, standard statistical models for scientific inference are not well designed to test this kind of null hypothesis. Difficulty in establishing no effect is part of why a Type 1 error is difficult to correct, as it requires correcting a finding with a null finding. Alternate statistical approaches have been proposed to extend our inference models, but none have been broadly adopted and these are outside the scope of our methodological discussions here.

Most of the time, extraneous variables do not affect the internal validity of a study. They create noise in measures that can lead to a failure to reject the null hypothesis statistically so that the experiment does not produce a result supporting a conclusion. When that happens, you do not know initially if you have experienced a Type 2 error, or simply that your hypothesis is wrong. The only thing you know for sure is that your experiment did not work. However, a critically important aspect of experimental design is to be as careful as possible to avoid the possibility of an experimental confound in your design. These can lead to the much more problematic Type 1 error, a false claim.

Confound

For an extraneous variable to be a confound, it has to vary with the independent variable. That is, the extraneous variable changes so that it exactly matches the independent variable. When this happens, we can no longer be confident that the intended independent variable was the cause of any changes we observe in the dependent variable. If a reliable effect is observed, it could have been caused by the extraneous variable. It is also possible that it was caused by the independent variable, but once there is a confound, there is no way to know what caused the effect.

We can illustrate the problem with our music example above. First we select operational definitions for music and the comparison, non-music, condition,

which we suggested could be white noise. Implementing these, we would plan to play audio clips and realize we had to set the volume these clips are played at. The volume here is a potential extraneous variable as it will affect perceptions of the intended independent variable and then influence the dependent variable scores. The worst thing you could do would be to have the volume for the music and the white noise be very different. If the white noise was played at a loud, obnoxious volume and the music at a pleasant, moderate volume, the extraneous variable volume would be confounded with music. If the scores on the dependent variable differed between conditions, we would have no way to know if it was the music or the loudness that caused the effect. Our experiment is confounded and we would describe any attempt to state an effect of music to have a problem with internal validity. If we attempted to claim from these data that music improved studying, we would be at risk of making a false claim, a Type 1 error.

One of the first elements of your design that need to be considered for potentially confounding extraneous variables are the stimuli to be used in the experiment. If we are using pictures in our study, we want all the pictures to be as similar as possible across conditions. If we ask participants to read stories about altruistic behavior, we want to be sure everything else about the stories they read is as similar as possible. Ideally, the only thing that differs is the exact variable we want to use as our independent variable.

We can also consider the broader contextual elements of our data collection. If there are multiple research team members engaged in data collection, they should not each run different conditions. If you are collecting at different times of day, avoid collection one condition in the morning and the other in the afternoon.

If it seems like it should be easy to avoid these kinds of problems, most of the time it is! Below we will provide two terms to describe very simple and straightforward ways to plan for these kinds of extraneous variables. It is necessary to go through the process of identifying the potentially confounding variables and controlling them since not doing so is essentially catastrophic for being able to draw any conclusions from your research study.

Internal Validity

The term **internal validity** is used to characterize an experimental design that will be able to test the underlying hypothesis. Any major problem that impairs the ability to draw a conclusion from the experimental data is a problem with the internal validity of the study. In addition to confounded variables, one way this can happen is if there is a mistake in the operational definitions. If they do not accurately reflect the underlying construct, the main inference about the constructs cannot be drawn from the data. Internal validity challenges are closely related to the problem of Type 1 errors.

This issue is distinct from **external validity**, which reflects the degree to which the conclusions can be applied to participants outside the research lab, in the real world. External validity generally depends on the methods of sampling participants, that is, how they are found and recruited into the study. This issue will be discussed in depth in Chapter 13, but as a preview, you can consider the concern being raised about the general dependence of psychological research on behavior measured from undergraduate students at major American universities. The question is whether the results obtained from university participants correctly predict the behavior of the broader population and whether we need to consider broader sampling or limiting the expected breadth of our conclusions.

If you suspect there might be a problem with external validity, that does not mean there is a problem of internal validity. For example, in our Experiment 1, the participants are university students. This not a problem with internal validity, nor a confound in the design. There might be a **limitation** to our conclusions that would cause us to raise the question of whether better memory from deep encoding only occurs with this specific type of population. In Chapter 4, we will consider when characteristics of the participants can actually affect internal validity but this is a separate issue from external validity.

Non-confounding Extraneous Variables

Most of the time, the many variables that vary across the course of collecting data for an experiment are not confounded with the independent variable. Things like time of day, or time of year, the weather, the details of the testing room, the social skills of the researchers supervising the data collection. If these differences are happening at random, they will affect the dependent variable, but will occur roughly equally often across both levels of the independent variable. Mainly what this does is increase the observed variance in the data, posing a challenge to the reliability of our results by making it more difficult to observe a robust difference in the dependent variable scores.

Any measure derived from human participants is going to have variance in performance associated with it. This is embedded in our statistical model for determining reliable effects of the independent variable. The difference in the average scores for the participants in each condition must be sufficiently large compared to the variance for us to have confidence it did not happen by chance.

Conceptually, variance in measured scores results in part from measurement error, which reflects the important idea that no quantitative operational definition is ever perfect. Another important component of this variance is the random noise caused by these extraneous variables. If there are many of these and they have large impacts on the measurement of the dependent variable, then extraneous variables can create **reliability** problems for an experiment design.

If the extra variance leads us to be unable to conclude that there was an effect of the independent variable on the dependent variable, then we may have run into a **Type 2 error**. It appears our experiment did not work, but we do not know for sure if our hypothesis was actually false. As noted above, this kind of error is less costly than a false positive created by a confound, but it still means the time and effort put into carrying out the study was not put to best use.

Control of Extraneous Variables

The principles for implementing best practices for reducing the effect of extraneous variables are simple in theory. Once the variables have been identified, keep as many as possible constant across conditions following the principle of **constancy**. Anything that cannot be kept constant but can be controlled, **counterbalance** across conditions so that it occurs equally often across levels of the independent variable. These two basic techniques remove the possibility of extraneous variables being confounds and maintain the internal validity of the study.

Constancy

As much as possible in any experimental design, keep things constant across the levels of the independent variable. This is the preferred technique for extraneous variable as it provides the best opportunity to observe a reliable effect of the independent variable. If the only aspect of the study that differs across conditions is the independent variable, then we can be very confident that changes in the dependent variable were caused by the manipulation.

Looking at the structure of Experiment 1 in Chapter 1, it should be clear how much of the presentation of that experiment was designed with constancy in mind. Everything possible about the overall look and feel of the interaction with the word lists was kept the same, except for the instructions about how to interact with the words. Maybe it would not have mattered if the word font, or font size differed across conditions, but if the readability of the words affected memory, then we would have had a major confound in the design.

It can be surprising to students who get to participant in psychological science research how meticulously detailed data collection procedures typically are. Many studies have carefully written scripts describing how the research team interacts with participants through the data collection process. This is done to keep interactions as constant as possible across conditions as well as to manage **demand characteristics** of psychological research, which will be

discussed in Chapter 4.

The importance of this level of experimental control is also seen when research is reported through. In Chapter 6, we will see that the Methods section of an APA-formatted research report includes a lot of this meticulous detail so that the reader can evaluate whether sufficient experimental control was used to justify the conclusions of the study.

The ability to impose a high level of control is a hallmark of laboratory experimental research. Studies done in well-controlled conditions will have the highest level of internal validity. Later, we will consider and contrast approaches used in field research where it is necessary to give up a lot of this control in order to increase our external validity and confidence that the findings apply in situations outside the laboratory.

Counterbalancing

For any factors that cannot be kept constant, distribute how these are implemented equally across conditions. For example, if participants are being run throughout the day, collect data from both of the experimental conditions equally early and late in the day to avoid confounds due to circadian (time of day) effects. If it is necessary to have multiple experimenters, make sure they each contribute to data collection in each condition. If the stimuli are presented in different orders to participants, make sure the orders are used equally across the conditions of the study.

Counterbalancing is focused on making sure the extraneous variables do not confound the study. It allows for the potential that these variables may contribute to measurement noise and increased variance in the dependent variable. When the full control of constancy cannot be achieved, we always prefer the risk of a Type 2 error, where our experiment does not achieve a reliable result, to the risk of a Type 1 error, where we draw an incorrect conclusion due to a confound.

Practically, implementing a counterbalancing procedure can be as simple as

alternating the administration of experimental conditions. In Experiment 1, students may be completing this study at the beginning of a class session. Early arrivers to class will have a slightly different experience than late arrivers. Students who arrive to class slightly later will potentially be completing the study under a sense of time pressure and aware that the rest of the class is waiting for them. Since we cannot control when students arrive to class, the experiment is implemented to alternate experimental conditions for each person who starts the study. If one student receives the *deep encoding* instructions, the next student receives *shallow* instructions. This distributes the variable of arrival time across the experimental conditions and removes the possibility of this being confounded with the independent variable.

In some cases, we may not be able to control variables enough even to counterbalance them carefully. In Experiment 1, participants may have been completing the study on a laptop, or on their personal phone which often have very different size screens. The online administration of the study was not done in a manner that allowed control of this variable so instead, we relied on the technique of **random assignment** to distribute this variable across our conditions.

There will always be a number of variables that are outside our control that will end up randomly in one condition or another. An important class of these variables in psychological science is differences among our participants. For Experiment 1, some people might be better at memory for word lists than others. In Chapter 4, we will discuss these kinds of **participant variables** and see that random assignment is the primary technique for experimental control. This approach does not reduce the internal validity of our experiment. In fact, our statistical procedures are designed around exactly the idea that the observed difference between conditions is larger than the difference that would happen if it was solely due to random assignment of these extraneous variables across conditions.

Summary

When planning a research study, or reading about a completed study, the standard method to try to identify potential confounds is to try to think of as many extraneous variables as possible that might affect the dependent variable. There will generally be quite a few, but most or all of these will not vary with the independent variable so we do not have to worry about them reducing the internal validity of the study by creating a confound.

Figuring out all the relevant extraneous variables can be challenging and benefits from knowledge and experience with related research findings. Practically speaking, there are generally a set of variables related to the stimuli used in the experiment and testing conditions that can be managed in order to both avoid confounds and minimize noise in the measure of the dependent variable.

One of the challenges to psychological research is that we can never be sure we have found all the possible extraneous variables. In fact, one major avenue of scientific discovery is finding new factors that affect our dependent variable that was not previously expected. Psychological science generally advances as a series of studies that build on each other using slightly different operational definitions and methods of experimental control. This both builds confidence in the main conclusions drawn about the underlying constructs and allows for gradual identification of other factors that affect psychological processes.

Key Takeaways

- An **extraneous variable** is any variable other than the independent and dependent variables.
- A **confound** is an extraneous variable that varies systematically with the **independent variable**.
- Studies are high in **internal validity** to the extent that the way they are

conducted supports the conclusion that the independent variable caused any observed differences in the dependent variable.

- Experimental methods are high in internal validity when confounds are avoided because the manipulation of the independent variable lets us infer that is what caused the observed difference in the dependent variable.
- A **Type 1 error** is a false positive claim where a researcher mistakenly thinks the independent variable reliably influences the dependent variable but it does not.
- A **Type 2 error** is a false negative claim where a researcher thinks the independent variable does not affect the dependent variable but it does in truth.
- Extraneous variables that do not confound the study increase variance in observed performance increasing the probability of a Type 2 error.
- **Constancy** is a method for reducing the effect of extraneous variables by as much as possible keeping everything the same across levels of the independent variable
- **Counterbalancing** is the approach to use for anything that cannot be kept constant. Distribute extraneous variable conditions across the levels of the independent variable to keep this element from being confounded with the independent variable and reducing validity of the experiment.

Exercises

Question 1: Laughter is the best medicine

Imagine you have just read an article in the newspaper describing a scientific study in which researchers found that people who laugh a lot tend to have lower blood pressure, stronger immune systems, feel less stressed out.

Considering the problem of extraneous variables and potential confounds, give an alternate hypothesis for how this relationship might be observed without supporting the authors' conclusion. Note that this requires a statement consistent with the data, not consistent with the conclusion.

Outline an experimental approach to this question that would more directly test the hypothesis. Provide an example of an operational definition of the IV, the DV and what you would expect to find if laughter positively affects health.

Question 2: Briefly answer the following questions about experimental control from our Experiment 1:

- Why have both groups read the same words?
- Why have 1-5 scales for responding for both conditions?
- Why require the word to be on screen for minimum 3 s?
- Does it matter if the trivia questions use words from the study list?

4 Experimental Procedure

In each of the examples given at the beginning of the first three chapters, the process of trying to identify all the extraneous variables should encounter the idea that people themselves differ importantly on each of the various measures. Not everybody is the same when it comes to memory for words, self-esteem, or reactions to music. Individual differences on these aspects of psychology are considered **participant variables** for the purpose of experimental design. Classic examples to consider are the age and gender of the participants, which may affect some psychological constructs.

It is important to acknowledge that our participants will differ but it would be impossible to try to identify every possibly way they differ in order to apply our experimental control tools of constancy and counter-balancing to these variables. There is a large area within psychological science, **personality research**, aimed at identifying, understanding and characterizing aspects of how humans differ from each other. Psychological science has a core philosophical tension between the attempt to draw broad conclusions that we believe are true for all humans while also acknowledging that many of these statements will apply differently to some individuals.

Since we cannot control participant variables, we mainly aim to keep them from being confounded with the independent variable. The technique for this is to **randomly assign** participants to conditions. This implements counter-

balancing as long as we do not somehow get unlucky that participants similar on some variable all end up in the same condition. Making this kind of unlucky circumstance unlikely is one of the main reasons we carry out psychological research on sizable groups of participants rather than a few individuals. Mathematically, once we have groups in the range of 30 participants or so, it is exceedingly unlikely that random assignment would lead to unbalanced participant variables. Such a low probability event is even less likely when a study gets **replicated** a second time with a new group of participants.

In recent years, there has been some concern in psychological science about important findings being difficult to replicate with new studies. The fear is that some claims are actually Type 1 errors and need to be discarded and underlying theoretical ideas revised. However, it is difficult to be sure of this as a failure to replicate might itself be a Type 2 error reflecting some aspect of weak experimental control in the attempted replication. However, the most interesting cases of failures to replicate is when they reveal new variables that affect the constructs and these often come from participant variables.

As an example, you may be familiar with the idea of stereotype threat in which exposure to a group-based statement of cultural expectations of poor performance can actually create poorer performance in that group. Stereotype threat research can be carried out with a simple design with two levels of an independent variable. In one condition, participants are exposed to the stereotype threat content and in the other condition, participants are exposed to control content that does not mention the stereotype. The dependent variable is measured performance on a related test. When a group participants score lower on the test, we can infer that exposure to the stereotype caused lower scores. Note that these results do not reinforce the stereotype because the group of participants not exposed are not affected, even though they are from the same group.

However, not every study of stereotype threat has been found to produce a reliable effect, leading to questions about the robustness of the phenomena. In Aronson et al. (1999) some insight into this variability was provided in

a study that examined stereotype threat on math performance but further asked participants how important math was to them, a participant variable they defined as *math-identification*. For students who self-reported that math was extremely important to their identity, stereotype threat was found to impair performance. Their study was also notable in that they used threat stimulus applied to white males who were exposed to the stereotype that *Asians perform better on math tests*, which additionally showed the influence of stereotype threat on a group normally seen as privileged. Students who reported that math was important to them were negatively affected by the stereotype, replicating the main hypothesis. However, students who reported low math-identification did not score lower on the test, indicating that they were not affected by exposure to the stereotype.

This finding implied that our understanding of stereotype threat needs to incorporate the idea of individual identity and how these interact. This insight likely emerged from researchers puzzling over a failure to replicate the simpler study and developing a theory that there was an unconsidered participant variable involved. Phenomena like stereotype threat are an important part of understanding how inequality can be inadvertently embedded in an educational context. It is therefore important that when the phenomena is found to vary across experiments that it not be discarded, but explored further for better understanding. Additional research can then build on these findings to test and understand ways to combat harmful effects of stereotypes on students in the classroom.

This example illustrates a difficult aspect of coping with extraneous variables in experimental design. The important factors are often not known in advance of the research. It can take a lot of experience and expertise in the specific research domain to learn where design problems might emerge from before new research can be done to extend the known theory about the main constructs for the study.

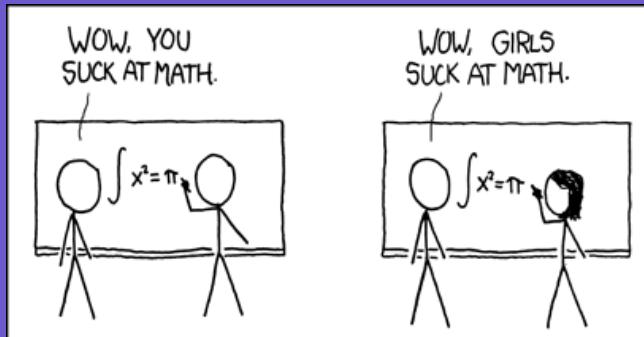
Stereotypes and bias are also well-known to be challenging to study in laboratory conditions because of effects related to **demand characteristics**, which we will also discuss below. These have to do with participants

modifying their behavior due to the awareness that they are participating in a research study. Participants knowingly aware they are part of a study of stereotype bias may monitor their behavior more carefully in order to avoid any accidental implicit shows of bias. That may pose a difficult challenge for constructing a research protocol high in internal validity.

Challenges even in simple experimental design reminds us that psychological science is not a list of facts to be memorized, although it may in some cases first be taught this way. It should be seen as a process of successive understanding as scientific studies build on each other to gradually improve our overall understanding of human behavior, cognition and emotion.

Bias in studying bias

Methodologies for studying stereotype threat are often excellent examples for simple experimental design, but teaching them seems to run the risk of actually causing the effects they describe. Stereotypes are examples of misunderstanding of correlation and causation. They may describe current relationships that could arise due to historical expectations, but lead to misleading conclusions about why. Observed relationships in the world are too frequently assumed to reflect inherent differences instead of effects caused by environmental factors.



Learning Objectives

1. **Participant variables** as extraneous variables that may affect the dependent variable, validity and reliability of psychological research.
2. **Random assignment** of participants to conditions to best distribute differences on participant variables.
3. **Demand characteristics** in experimental research and the effect of awareness by participants of being in a research study.
4. Using **single-blind** methodologies to limit some kinds of demand characteristics.
5. **Placebo effects** as an example of demand characteristics in health and intervention research
6. Avoiding bias in scoring the dependent variable through **double-blind** methodologies
7. Implementing the experimental procedure using care and consistency in following the protocol for best experimental control

Participant variables

People differ in ways that might affect their behavior in an experimental situation. They may vary in aspects that are measured on the dependent variable. They may react or interact with the levels of the independent variable differently. These differences act like the extraneous variables discussed in Chapter 3. If the participant variables are confounded with the independent variable, this causes a threat to the internal validity of an experiment. If they are randomly distributed across conditions, validity will not be challenged, but these differences will contribute to variance in the quantitative measures in the experiment and may challenge reliability.

We have to treat participant variables slightly differently than some other kinds of extraneous variables because we cannot control or change these

aspects of our participants. We have a lot of control over the stimuli in an experiment, the environment around the testing procedure, the manner in which the research team interacts with the research participants. For participant variables, we have one very effective technique: **random assignment** to conditions.

Random Assignment to Conditions

As long as participants are randomly assigned to conditions, individual differences should never confound the final result. It is tempting to worry that it is possible to get unlucky in our randomization and assign all the participants who are better at the task to the same condition. However, this is exactly what our statistical tools are designed to test. For all our statistical tools for deriving inferential statistics, the *p value* that we calculate is formally the probability that we accidentally observed the difference that occurred due to this random chance. The idea of accidentally seeing the difference is formally the same as the **null hypothesis** that there is no effect of the independent variable on the dependent variable. When we reject the null hypothesis, we explicitly consider and mathematically rule out the possibility that individual differences, or any other non-confounding extraneous variable accounted for our results.

It is important to note that for random assignment to work, it has to be carried out correctly and there needs to be an adequately large sample of participants recruited for the study. A good, simple rule-of-thumb is to try to have at least 30 participants in each of your experimental conditions, if possible. It isn't always possible to obtain that many volunteers, however, and 15-20 per condition also often works. Carefully estimating statistically adequate sample sizes is a process called **power analysis**, which can actually be a complex task. A full consideration of the factors that go into a power analysis is beyond the scope of this text, but we will touch on the key ideas in discussions of statistics in Chapter 5, sampling techniques in Chapter 11 and research proposals in Chapter 13.

Smaller sample sizes weaken the effectiveness of random assignment. In some specialized cases with restricted populations such as neuropsychological research or other kinds of case studies, it is not always possible to recruit large samples. In these cases, it may be necessary to use designs based on **matched participants**, where participant-based extraneous variables are assessed and explicitly balanced across the levels of the independent variable. This and related techniques were used in some older psychological science studies that pre-date the modern recommendations to use larger sample sizes. The challenge of matching procedures is the need to identify all possible participant-based extraneous variables and then have reliable measures of all of these prior to assigning conditions. It is generally much simpler just to randomly assign a large group of participants to conditions and trust that the statistical model will account for assignment luck.

Random assignment, properly carried out, will prevent individual differences from confounding an experiment. However, incorrectly following the randomization procedure can lead to embedding bias in a study. A very simple procedure for effective randomization is to alternate conditions for participants as they are recruited into the study. Later in this chapter we will discuss the importance for having a pre-planned randomization strategy, even a simple one, for avoiding any accidental bias in carrying out a research study.

Constancy in participant variables

In Chapter 3, we saw that our two main approaches to extraneous variables are counter-balancing and constancy. It is possible to apply something like constancy to participant variables by selectively recruiting participants keeping a variable constant. For example, many cognitive neuroscience studies of language limit participants to right-handed people, who generally have their language areas isolated in their left cerebral hemispheres. Left-handed people are more likely to have their language areas isolated in their right cerebral hemispheres or distributed across both hemispheres, which

can change the way they process language and thereby add noise to the data. We might also select participants if our hypothesis is specifically about a subgroup, such as reactions to emotional stimuli among people high in anxiety.

Selective recruiting can increase the reliability of a planned study without causing a reduction in internal validity, but it lowers the **external validity** of the study—in particular, the extent to which the results can be generalized beyond the people actually studied. Typically this results in conclusions that are limited by including the recruiting criteria. Cognitive neuroscience studies of language typically have to include the statement *in right-handed participants* to reflect that selection of participants.

Historically, a great deal of early health-based research was done with insufficient attention to maintaining appropriate diversity in participant recruiting. In some studies, recruiting was entirely based on white males, leaving large gaps in the scientific literature about how these health interventions affected everybody else. The attempt to justify this at the time was that this reduced variability in participants, increasing the power to detect whether a health improving intervention was clinically effective. However, it should be clear that this also raises significant ethical concerns that these research studies were not being designed to provide benefit widely across the population. This is rare in modern research but provides a reminder that we would prefer our science to provide findings that apply as broadly as possible. So restricting participants on their characteristics is something we would generally prefer to avoid.

Concerns about how effectively even modern psychological science really capture the diversity of the human population have drawn attention to the fact that a lot of psychological research depends entirely on university students as participants. In the USA, Canada and Western Europe, it has been observed that the undergraduate population may not reflect the broader population in society. This issue has been described as an over reliance on **WEIRD** participant samples, which is an acronym for Western, Educated, Industrialized, Rich and Democratic. As a result of this recruiting aspect,

commonalities in social or cultural expectations in these participants may be implicitly embedded in many psychological research reports. The main implication of this idea is that there may be unknown extraneous variables that vary across social and cultural groups that affect behavior in ways we have yet to explore in research. That does not invalidate research that depends on WEIRD populations, but may affect applications of the findings to broader, more diverse populations.

One technique for increasing potential diversity of research is to use methodologies for collecting data online. Research on how online methodologies affect recruiting diversity is ongoing but suggests that these samples are at least more diverse across ages.

Practically, psychological science reflects restrictions in recruiting in a way we will see in Chapter 11 is called **convenience sampling**. Participants in research are those who are available through the local environment, usually a university, or through online systems. Since it is not logically possible to use a perfectly diverse sample, we accept and should be aware of some limitations in the external validity of our conclusions. Within that constraint we most commonly maintain the internal validity of the research process by the simple expedient of random assignment of participants to conditions.

As we consider the practicalities of carrying out a research protocol, we should be aware of another class of extraneous variables that are difficult to fully control related to the fact that participants generally know that they are in a research study. In most experimental research, participants are fully aware that they have volunteered to participate in a psychological science project. This very likely has substantial effects on behavior for research on sensitive topics like stereotype bias but may also influence measures much more broadly.

Demand Characteristics

The general set of issues related to unexpected environmental effects on experimental methodology are termed **demand characteristics**. These reflect influence or bias accidentally imposed by details of the methodological procedure. A goal of experimental control is to prepare a rigorous research process that minimizes the risk of these effects distorting the experimental measures. Improperly controlled, demand characteristics can create confounds with the independent variable, leading to Type 1 errors.

The idea that participants might change their behavior simply because they are being watched is sometimes referred to as the Hawthorne Effect. This refers to an old, possibly apocryphal, study of industrial processes where every controlled change to the environment led to performance improvements. It was subsequently hypothesized that the workers in the studies simply put in more effort when they believed they were being observed for the research.

In controlled laboratory conditions, participants are generally going to be aware that they are participating in a research process where their behavior is observed and measured. As we will see in Chapter 8, Basic Ethics, standard ethical practice is to inform participants about the research they are participating in and obtain signed **informed consent** to participate. In some kinds of field research or non-experimental observational studies this may not be the case but for the current purposes, we consider experimental control when participants are aware of their participation.

Since participants will be aware of being in a study for all conditions of the independent variable, this aspect itself is maintained through constancy. However, we need to attend to details of the levels of the independent variable as they interact with the **expectations** of participants. When participants are in a condition that they expect to affect their behavior, they may exhibit changes due to those expectations instead of the actual effect of the independent variable.

The most common example of this is the well-known effect of **placebos** that occur in medical and clinical research. A placebo is a simulated treatment that lacks any active ingredient or element that should make it effective, and a placebo effect is a positive effect of such a treatment. Many folk remedies that seem to work—such as eating chicken soup for a cold or placing soap under the bed sheets to stop nighttime leg cramps—are probably nothing more than placebos. Placebo effects are not primarily driven by people's expectations that they will be effective. Many people are not surprised that placebos can have a positive effect on disorders that seem fundamentally psychological, including depression, anxiety, and insomnia. However, placebos can also have a positive effect on disorders that most people think of as fundamentally physiological in medical research. Placebo effects are interesting in their own right as they imply surprising interactions between psychological and physiological processes. However, they also pose a challenge in experimental control for researchers who want to determine whether a treatment works.

With two levels of an independent variable in an experimental design, we do not typically have to be specifically concerned about a placebo effect, but we do need to attend to the expectations of participants. In Experiment 1, participants were not told that the hypothesis for the experiment was that the deep encoding condition would lead to better memory. If they were told this, those participants may have more actively engaged with effortful study and expecting the hypothesis to be true might have influenced the results. This would create a significant validity problem for the study, potentially producing a Type 1 error.

As an aside, sometimes researchers become concerned about the opposite kind of expectation effect where they suspect the participants have become aware of the hypothesis and are deliberately producing behavior to ruin the experiment. This is both unlikely to occur and would also produce the less problematic Type 2 error if the hypothesis were correct. A concern like this more likely emerges from the fact that carrying out effective research can be very challenging and many well-conceived studies still do not work reliably.

The simplest way to avoid the basic expectations problem in a two-group

independent samples design is to not inform the participants about the hypothesis or the other condition of the study that they are not participating in. This approach is referred to as a **single-blind** procedure and was the way we implemented our Experiment 1 here. This is an extremely common method for designing psychological research that strengthens the internal validity of the experiment by eliminating concerns about demand characteristics.

In some cases more extreme versions of disguising the research study are used to avoid expectation effects. There are a variety of ways in which **deception** is used in research procedures to keep participants unaware of the hypothesis. These can range from telling the participant that they are waiting for the next part of the procedure but their behavior is being watched surreptitiously, having participants believe they are interacting with another participant but it is really a research team member (confederate) or deliberately misleading participants about performance on a test to manipulate their emotional state.

Deception in research reflects a significant challenge to ethical research practices. It is one of the more common research techniques that can only be employed with oversight and awareness of the scientific regulatory body, the **Institutional Review Board**. In Chapter 8, we will touch on the ethical implications and common practice for balancing scientific rigor with fair and appropriate treatment of human participants. As a general rule, we prefer participants to know what they are engaging in when participating in research. Yet in any kind of blind design, even without overt deception, we usually cannot explain everything in advance, which we will see is an example of tension between ideal ethics and ideal experimental design.

Inadvertent bias in research procedures

When carrying out a research protocol using a design that keeps information about the hypothesis away from participants, some care has to be taken by the research team in how they interact with participants. There will often

be a planned procedure for explaining some aspects of the study, but not all. Interactions with participants can even be scripted to help make sure the protocol is administered in the same way across all the participants.

Areas where bias can inadvertently creep into research procedures can come from expectations of the research team. Virtually all researchers want their experiments to succeed, which can lead to subtle effects like simply being more socially engaging and interacting with participants in the condition where participants are hypothesized will perform better. Many psychology experiments require a research team member to be in the room to observe behavior and in this case it is vitally important that interactions with participants are constant across conditions.

This kind of bias can even occur through assignment to conditions if consistent procedures are not implemented. If participants are assigned to conditions by the research team when they arrive to participate, more attentive and engaged participants might get assigned to the more challenging or interesting condition. This creates a confound of **sampling bias** that reduces the internal validity of the study. This kind of bias is a greater risk in health and clinical research with targeted populations than more general experimental psychology approaches. In these kinds of studies, it is particularly important to have a well-documented and carefully followed procedure for assignment to conditions, even if it is just as simple as alternating between conditions.

Some protocols require so much interaction between the research team and participants that it is impossible to be confident that all interactions will be free of any inadvertent bias. For these studies, it is necessary to use a **double-blind** procedure where the members of the research team are also unaware of what condition the participants are assigned to. These are logically complex to employ so they only get used when there is significant bias risk. An example from medical research are clinical studies of the effectiveness of a new drug. In those studies, the pharmacy prepares numbered doses that appear identical and no knowledge of which are the treatment or control is available to the research team during the research

protocol. After data collection is complete, the study is then unblinded and information about which dose was treatment or placebo is provided. Only then can the data be analyzed for treatment efficacy.

Fully double-blind procedures are rare in psychological science due to the difficulty of implementing these consistently. More common are simple procedures for assignment to conditions, combined with scripted interactions with participants before and during administration of the research protocol. However, it is not that uncommon to need to develop a special procedure for scoring some kinds of dependent variables when there is a subjective element to rating participants' behavior.

Avoiding bias in scoring

Because psychological science is often about factors that affect behavior, some studies use an operational definition that requires quantification of observed behavior. For example, we might be interested in evaluating the effect of an anxiety-reducing manipulation on public speaking performance. The dependent variable here would be an evaluation of how well the participants performed on a public speaking task, requiring a quantified judgment of that performance by the research team. Because the judges will need to make subjective decisions about the quality of performance, there would be a high risk of bias if they had full awareness of the participants' condition when making those ratings.

Whenever a subjective evaluation is part of scoring the dependent variable, it is common to use **independent raters** who carry out the scoring process without knowledge of the level of the independent variable. This requires having some members of the research team remain blind to condition but others may have full knowledge of the procedure. Keeping the raters unaware of the experimental condition avoids any bias influencing their rating by the experimental hypothesis. The raters are often trained with detailed instructions on how the scoring process is to be carried out and in some cases multiple raters are employed and scores across them combined.

In Chapter 2, we introduced the idea common to nearly all psychology studies that an abstract construct needs to be turned into a quantitative variable through the process of creating an operational definition. In some cases this can be a scale, like for self-esteem, or a performance measure like recognition memory accuracy. However, there are many areas within psychology where it takes a human being to provide an evaluation of behavior in order to carry out the operational definition. We might want to measure an aspect of emotional expression such as laughter, or rate the quality of partner interactions in a study of relationships. For any subjective judgment like this, we assume that experimenters who are aware of the design are also invested in the outcome of the study, and therefore are at risk for experimenter bias and should not be the source of the measure. In these cases, methodologically rigorous research relies on ratings provided without knowledge of experimental condition.

Design of Experiment 1

Our Experiment 1 reflects a handful of design decisions aimed to keep extraneous variables constant across the two conditions in the study: deep and shallow encoding. All participants rated the exact same set of 30 words, although the instructions for the rating varied as the independent variable. The words themselves were selected to be between 5 and 8 letters in length and to have a *written frequency occurrence* of 30-80 times per million. The characteristics of the words were kept similar to reduce variance in memory for the words chosen for the experiment.

Unless you have some experience in memory research using word lists, you might not have anticipated that the length or frequency of the stimulus would be important for the design. Knowing what potential extraneous variables are relevant to a specific study often requires some prior knowledge of research in that domain. Once the variables are identified, the technique for controlling them is straightforward: select words in a restricted range from a database of word frequency information.

In addition to the stimuli, note that the two scales used for rating the stimuli were also constructed to have 5 levels. Although it is unlikely that the specific number of levels on the scale will affect memory, it is good practice to keep as many design elements the same as possible across conditions.

In cases where the data collection for Experiment 1 are done in the classroom, we also gain the benefit of all the participants complete the study in the same conditions in terms of surrounding and time of day. When this experiment is completed by participants outside the classroom, there may be influences of outside distractions and attention that are outside of experimental control. Note that these would be examples of extraneous variables that increase variance, but do not confound the study because we have no reason to believe that either of the conditions of the independent variable would be more affected by distraction.

The design of Experiment 1 also includes 3 minutes of irrelevant trivia questions to be completed after performing the word rating and before the surprise recognition test. The time of the trivia task is kept constant across participants, but the number of questions answered and the content of the questions is not. The number and content of the questions experienced is allowed to vary randomly across all the participants in the study, potentially contributing to variance in the memory measure but not in a way that is confounded with the study conditions.

Practical considerations

Best practices for controlling extraneous variable in carrying out psychological research can lead to fairly elaborate and precise procedures for research personnel. As a consequence of this, it is very common for research procedures to be evaluated with a short period of **pilot testing** before starting formal data collection. Sometimes this can mean simply practicing carrying out the research procedures under observation of other researchers to ensure it is working as intended by the planned operational definitions. It can also mean running a small preliminary sample of participants to evaluate

the procedure and scripts. It should be very clear in the overall research plan when pilot testing is underway and when that process is complete and formal data collection for the planned study starts. Pilot testing data is not intended for inclusion in published research and may often depend on knowledgeable members of the research team (or collaborating teams). This can affect demand characteristics of those participants making their behavior or performance importantly different from the main intended recruited sample.

A common feature of pilot testing of procedures is to include a measure referred to as a **manipulation check**. This is a measure that will often look like a dependent variable but is not part of the research hypothesis. For example, in a mood manipulation study using music to create positive/negative moods, participants might be asked after listening to the music to rate their mood. If mood ratings were not consistent with the independent variable (music type), we would have concern about the operational definition being used. In some research publications, manipulation check data may be included and even analyzed statistically but note that no real hypothesis is being tested. A statistically reliable effect that the music manipulation affected self-rated mood only validates the operational definition of the IV and does not lead to any general conclusion.

Pilot and preliminary testing can also be used to examine the distributional characteristics of the dependent variable. As we will see in the next chapter, our ability to draw inferences from our data will depend on observing statistically reliable effects of the IV on the DV. Poorly controlled extraneous variables may lead to high levels of variability in performance, which will show up as high variance and may indicate a need to improve experimental control in design. Accurate estimates of variance often require large participant samples, though, so this cannot always be anticipated.

Pilot testing is often very useful to identify potential statistical problems with floor effects or ceiling effects in the DV. Ceiling and floor effects occur when the dependent variable measurement range is not properly anticipated in the experimental design. For example, a floor effect will occur when a task is too difficult for participants. If participants are given a problem-solving task with

the intention of the measure being the number of problems solved but nobody is able to solve any of the problems, everybody will score zero regardless of the IV manipulation (no reliable difference can be detected). Similarly, if all participants get all the answers correct, performance is at ceiling for all groups and again there is no possibility of observing a statistically reliable effect. Pilot testing is often used to verify that scores on the dependent variable will be within a range that allows for detectable influence from the independent variable so that we have some chance that our statistics will be effective.

Key Takeaways

- **Random assignment** to conditions in between-participants experiments is a fundamental element of experimental research. The purpose of this technique is to control extraneous variables so that they do not become confounding variables.
- Restricted participant sampling may reduce variance but should be used infrequently and carefully due to effects on generalizability of findings.
- **Demand characteristics** have surprisingly large effects on behavior. Simply knowing that they are in a research study may change behavior of participants.
- **Blind** or **single-blind** designs keep the participants from knowing the full experimental hypothesis and influencing their behavior and are very commonly used.
- **Double-blind** designs mean that some members of the research team are unaware of which condition participants are in. This is used to avoid any bias in scoring the dependent variable, especially if there is any element of scoring that requires subjective judgment.
- Rigorous systematic procedures for data collection are important and contribute to research success. Written detailed scripts for the research process help manage both the influence of extraneous variables and demand characteristics on psychological research.

Exercises

Question 1

Craik & Tulving (1975) reports a series of studies examining the effect of various approaches to deep and shallow processing on memory. Review this publication and answer the following questions about specific experiments reported there comparing the procedure to our Experiment 1.

- In their Experiment 1, how many levels of the IV were used? What was the DV measure of memory?
- Their Experiment 5 is carefully designed to address what confounding alternative hypothesis? To do so, what aspect of the IV is made as constant as possible?
- In what way was their Experiment 9 similar to our in-class experiment? Identify some methodological differences

Question 2

You are doing a study at a local school. Because of the way things are scheduled, you can have one small testing room in the morning and another much larger testing room in the afternoon. If you have two treatment conditions (A and B), how can you assign subjects to the testing rooms so that the type of room will not lead to confounding your experiment?

Question 3

Dr. L is planning a large scale learning experiment. He would like to have 100 rats in one treatment group and another 100 in the other group. Because he needs so many rats, he says, "Well, I can't test all these animals by myself. I'll ask Dr. P. to help me. He can run the animals in the one group while I test the animals in the other group." What is the potential problem with this approach and how would you improve the procedure to correct it?

5 Statistics 1

Statistics are necessary to establishing confidence in the results of our experiment so that we can draw conclusions about the influence of our independent variable on the dependent variable. Since we have created operational definitions of our constructs, we now have a method to quantitatively describe and communicate the observations we obtained in our study. Mathematically, our inferential statistical models tell us if the size of the difference observed exceeded what would happen by chance distribution of variance associated with the extraneous variables and other aspects of measurement error.

Our focus will be on practical use and application of statistics so that we are able to draw inferences from data. It is assumed the student here has a background in basic statistics and the foundational math and probability of carrying out statistical tests. Here we will use standard tools for helping us carry out a statistical test, how to read the output of these tools and most importantly, how to report the results in standard APA format. Scientific writing in psychology generally follows a format defined by the American Psychological Association (APA) and described in the Publication Manual. Explanation for how to prepare a full scientific report starts in the next chapter (Chapter 6) but in this chapter we will see how to go from basic quantitative description of the data to the pieces that are crucial for the

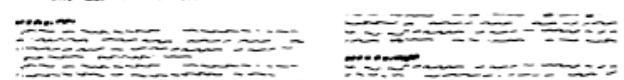
The Replicability Crisis in Psychology

The gold standard for confidence in a result from a scientific study is that it will *replicate*, that is, if you carry out the study again you will obtain the same result.

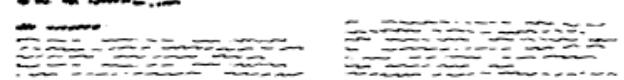
There has been some recent concerns about some well-known findings in psychological science for which new attempts to replicate have not been successful. The fear is that there are false positive, Type 1 errors in the published literature that are being relied on and taught as fact. However, there are other reasons why replication studies might not succeed. These studies could have used different participant groups or varied on other extraneous variables. If these factors affect the results, that reflects scientific progress in better understanding how these factors interact with the constructs. Attempted replication studies may also have encountered Type 2 errors.

Confidence in findings should come from a series of studies that embed replications rather than a single headline experiment. Successful and unsuccessful extensions and replications should be expected as part of the normal progress of scientific advancement.

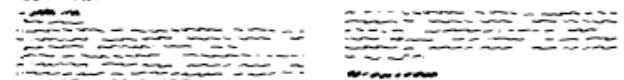
MANY COMMERCIAL ANTIBODY-BASED IMMUNOASSAYS ARE UNRELIABLE



PROBLEMS WITH THE p-VALUE AS AN INDICATOR OF SIGNIFICANCE



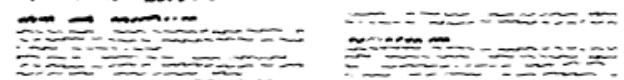
OVERFEEDING OF LABORATORY RODENTS COMPROMISES ANIMAL MODELS



REPLICATION STUDY FAILS TO REPRODUCE MANY PUBLISHED RESULTS



CONTROLLED TRIALS SHOW BUNSEN BURNERS MAKE THINGS COLDER



Results section in an APA format manuscript. That will focus on how to extract the information to be reported, which is a subset of the information provided by our statistical analysis tools.

We assume the reader has basic familiarity with the idea of **rejecting the null hypothesis** that two distributions do not differ and the standard criteria for reliability of $p < .05$. Here we will focus on practicalities of basic data analysis and drawing inferences from the data. These are conditioned on meeting this statistical criterion so that we can draw conclusions about what we have learned by a finding that our manipulated independent variable affects our measured dependent variable.

More recently within psychological science it has been pointed out that we should avoid an over-reliance on whether the statistical analysis meets the magical .05 criterion. In addition to this value, we should report and consider the **effect size** of our experiment, that is, exactly how much did the score of the dependent variable change across conditions. In some simple designs, like our Experiment 1, the effect size is simply the difference in the mean performance across conditions. Later when we discuss more complex designs, we will observe that it becomes more important to explicitly discuss how large our observed effects are in addition to the basic binary statement about meeting or not meeting the reliability criterion.

The importance of effect size also depends to some degree on the operational definition of the dependent variable. For Experiment 1, the recognition memory score is a fairly abstract measure of memory from study but the exact percent correct score does not immediately tell us how much better we might study in practice. For health research, or measures of a construct like self-esteem, exactly how much a manipulated independent variable affects that measure may be an important part of effectively communicating the results. The underlying goal of our statistical approach to data is to effectively communicate what we have learned from our research study.

Learning Objectives

- Using **descriptive statistics** to describe the results of the data obtained in the experiment in a way that connects to the design
- Describe the purpose of **inferential statistics** and carry out a t-test
- Distinguish between descriptive and inferential statistics
- **Reporting Results** in APA format
- Preparing a **data visualization** to support communication of results

Descriptive Statistics

Descriptive statistics are used to organize or summarize a set of data.

Examples include percentages, measures of central tendency (mean, median, mode), measures of dispersion (range, standard deviation, variance), and correlation coefficients.

Measures of central tendency are used to describe the typical, average and center of a distribution of scores. The **mode** is the most frequently occurring score in a distribution. The **median** is the midpoint of a distribution of scores. The **mean** is the average of a distribution of scores.

Measures of dispersion are also considered descriptive statistics. They are used to describe the degree of spread in a set of scores. So are all of the scores similar and clustered around the mean or is there a lot of variability in the scores? The **range** is a measure of dispersion that measures the distance between the highest and lowest scores in a distribution. The **standard deviation** is a more sophisticated measure of dispersion that measures the average distance of scores from the mean. The **variance** is just the standard deviation squared. So it also measures the distance of scores from the mean but in a different unit of measure.

Typically **means** and **standard deviations** are computed for experimental research studies in which an independent variable was manipulated to

produce two or more groups and a dependent variable was measured quantitatively. The means from each experimental group or condition are calculated separately and are compared to see if they differ.

A critical descriptive statistic for reporting results that is not always thoroughly discussed in prerequisite statistics classes is the **standard error**, which is related to standard deviation but reflects estimated error in the data-driven estimate of the conditions means. This value is critical to evaluating whether there are reliable differences between the conditions means and is a standard element of reporting the descriptive statistics in APA format writing. The formula for calculating standard error, SE, is:

$$SE = \frac{SD}{\sqrt{n}}$$

Data Aggregation and Handling

To organize and evaluate the data from a research study, it is strongly recommended to use a spreadsheet software program such as Google Sheets, Microsoft Excel or similar. We will provide practical examples here to illustrate basic data handling in Google Sheets with specific examples of how to calculate the descriptive statistics and visualize the data.

Standard practice for research data is to prepare a matrix in which the data for each participant is on a single row with values across columns in the sheet. An example is shown below.

A section of data from Experiment 1. There are four columns of information. The first column, A, is headed Code and contains the participant codes of the students who participated in this experiment. Each row shown is a different individual participant from this study. The second column, B, headed Condition is the study condition these participants were in. All of these participants were in the *deep encoding* condition. Column C, Percent Correct is their score on the recognition test, the *dependent variable* for this experiment. We have a little additional information here in Column D, Missed Trials, which is the number of times they did not answer on a test item.

	A	B	C	D
1	Code	Condition	Percent Correct	Missed trials
2	YED7S	deep	100	0
3	AZ2KR	deep	95	2
4	6SEV4	deep	50	4
5	LXRKN	deep	91.7	2
6	2Z5MN	deep	93.3	1
7	427GH	deep	96.7	1
8	8GQ9X	deep	95	3
9	8Y42F	deep	91.7	1
10	F3AQ6	deep	95	5
11	B62ST	deep	93.3	0
12	2AQGF	deep	95	0
13	4ZZE4	deep	96.7	1
14	54UHW	deep	93.3	0
15	ZTRVII	deep	98.3	0

In actual practice, getting experimental data into a format like this includes several steps. Data from individual participants might be entered by hand, for example from scoring written answers from in-person data collection. It may also be produced from an automated scoring program. Once all the data is accumulated in the same spreadsheet, we can calculate the descriptive statistics to summarize the performance of the groups of participants who were in each experimental condition.

Practically Calculating Descriptive Statistics

For simple designs, the set of needed descriptive statistics needed may only be the three basic ones: mean (average), standard deviation (SD) and standard error (SE). For a design with one independent variable with two levels or conditions, we will need to calculate these three numbers for each of our two groups. More complex designs may require more calculations for more groups or more variables. The point of calculating the descriptive statistics is to do what it sounds like, describe the results observed. Only after we have done this, should we move on to calculation of inferential statistics needed to draw inferences.

Calculating descriptive statistics using spreadsheet software is done by typing a formula to be used into an empty cell and directing the formula calculation to the data we are trying to describe. An overview of using formulas in spreadsheet software is shown below.

After calculating the mean, we will repeat the process with the same data range with the formula to calculate the standard deviation. The name of the formula to calculate mean is usually **=average()**. The name of the formula to calculate the standard deviation of a sample of data is **=stdev.s()**. You may remember that there is a small correction to the calculation of the standard deviation of a sample from the population. This formula includes that correction and provides the correct number.

Formulas

Continuing our example, we have chosen cell C28 as the place to calculate the average of the Percent Correct measure for this sample of participants.

In this cell, we entered the formula =average(and then selected the data from C2 to C26 for inclusion. The content of the resulting formula is shown at the top =average(C2..C26). Once entered, the value in the C28 cell is the calculation of this formula, which is the number 91.02.

This is the basic process we will use to obtain our descriptive statistics from this type of data organization.

The screenshot shows a Google Sheets spreadsheet titled "Sample Data". The data is organized into four columns: A (Code), B (Condition), C (Percent Correct), and D (Missed trials). Rows 1 through 29 contain data points. Row 28 is highlighted in blue and contains the formula =AVERAGE(C2:C26). The formula bar at the bottom also displays =AVERAGE(C2:C26).

	A	B	C	D
1	Code	Condition	Percent Correct	Missed trials
2	YED7S	deep	100	0
3	AZ2KR	deep	95	2
4	GSEV4	deep	50	4
5	LXRKN	deep	91.7	2
6	Z25MN	deep	93.3	1
7	427GH	deep	96.7	1
8	8GQ9X	deep	95	3
9	8Y42F	deep	91.7	1
10	F3AQ6	deep	95	5
11	B62ST	deep	93.3	0
12	2AQGF	deep	95	0
13	4ZZE4	deep	96.7	1
14	54UHW	deep	93.3	0
15	7TBVU	deep	98.3	0
16	AWC15	deep	83.3	6
17	BDVVR	deep	91.7	1
18	EYXFI	deep	96.7	0
19	HNQ6W	deep	86.7	2
20	IXA6U	deep	96.7	0
21	KDM9X	deep	85	4
22	MDFIV	deep	80	1
23	PEM6F	deep	95	0
24	WJCUW	deep	96.7	0
25	WVQ6H	deep	81.7	0
26	XPTXY	deep	97	0
27			C28	=AVERAGE(C2:C26)
28				
29				
30				
31				

There is no built in formula for calculating the standard error of the sample so we will need to calculate this from the above formula for SE. Since we have already calculated the standard deviation, SD, we need to get the number of participants, n. This can be done with the =count() formula. Then we construct a formula calculating the standard error as the standard deviation divided by the square root of n, sqrt(). Below on the left are the formulas entered and on the right is the values these formulas produce with the sample data.

28	Mean	=AVERAGE(C2:C26)
29	SD	=STDEV.S(C2:C26)
30	N	=COUNT(C2:C26)
31	SE	=C29/SQRT(C30)
--		

28	Mean	91.02
29	SD	10.06714955
30	N	25
31	SE	2.013429909

At this point we would have the necessary descriptive statistics for just one of our two conditions. To complete the calculation of the descriptive statistics for the experiment, it will be necessary to repeat the process for the mean, SD and SE for the other condition, the shallow encoding group of participants.

Various software packages that support calculation of the more complex inferential statistics will also often calculate and report descriptive statistics. However, reviewing and manually calculating these is a highly recommended step in data organization and handling. Data from human participants is often very noisy, with outlier participants who may have failed to comply with the provided instructions. Automatic scoring processes can produce occasional or even consistent errors. Looking over experimental data carefully will help identify mistakes or errors in data processing. Catching mistakes is critical to accurate reporting of our experimental results.

As we will see in Chapter 6, reporting the results of the experiment relies heavily on these descriptive statistics to summarize the observed data. Actual data tables from individual participants is almost never included in standard reporting. Therefore, the reader is relying on the researcher to fairly and accurately report the careful quantitative summaries of performance. Redundant calculations, once in the spreadsheet and once in the statistical software, help spot errors through discrepancies in values. Avoiding these is a critical part of a rigorous scientific methodology, especially when working with typically messy data from human participants.

Inferential Statistics

As you learned in your prerequisite course in Statistics, researchers sample from a population but ultimately they want to be able to generalize their results from the sample to a broader population. Researchers typically want to infer what the population is like based on the sample they studied. Inferential statistics are used for that purpose. Inferential statistics allow researchers to draw conclusions about a population based on data from a sample. Inferential statistics are crucial because the effects (i.e., the differences in the means

or the correlation coefficient) that researchers find in a study may be due simply to random chance variability or they may be due to a real effect (i.e., they may reflect a real relationship between variables or a real effect of an independent variable on a dependent variable).

Researchers use inferential statistics to determine whether their effects are statistically significant. A statistically significant effect is one that is unlikely due to random chance and therefore likely represents a real effect in the population. More specifically results that have less than a 5% chance of being due to random error are typically considered statistically significant. When an effect is statistically significant it is appropriate to generalize the results from the sample to the population. In contrast, if inferential statistics reveal that there is more than a 5% chance that an effect could be due to chance error alone then the researcher must conclude that their result is not statistically significant.

It is important to keep in mind that statistics are probabilistic in nature. They allow researchers to determine whether the chances are low that their results are due to random error, but they don't provide any absolute certainty. Hopefully, when we conclude that an effect is statistically significant it is a real effect that we would find if we tested the entire population. And hopefully when we conclude that an effect is

Statistical Tools

This chapter provides a recipe for carrying out a two independent samples t-test using the statistical software, R, within an interface, RStudio.

There are many other statistical packages that could also be used effectively. Each has their own specific output format that includes the key values that need to be reported. Developing familiarity with one package through practice with the output format will be valuable for students learning experimental processes.

not statistically significant there really is no effect and if we tested the entire population we would find no effect. And that 5% threshold is set at 5% to ensure that there is a high probability that we make a correct decision and that our determination of statistical significance is an accurate reflection of reality.

The *t*-Test

For the simple experimental designs we have considered so far, our hypotheses about how the independent variable affects the dependent variable depend on observing the difference between measures for the two groups. We calculate the mean observed scores for the dependent variable for each group then need to ask, is the difference between these two means enough to justify a conclusion about our study. The most common null hypothesis test for this type of statistical relationship is called the **t-test**, shorthand for the Student's t-test.

In this section, we will review the standard types of t-tests to provide a connection to your prior study of basic statistics. We will then focus on several simplifying practicalities for using these tools and do a hands-on exercise carrying out a t-test analysis in R. The basic types of t-test to review first are: the one-sample t-test, the dependent-samples t- test, and the independent-samples t- test.

One-Sample *t*-Test

The **one-sample t-test** is used to compare a **sample mean, M** , with a hypothetical **population mean, μ_0** , that provides some interesting standard of comparison. The **null hypothesis** is that the mean for the population (μ) is equal to the hypothetical population mean: $\mu = \mu_0$, sometimes described as "chance performance." The alternative hypothesis is that the mean for the population is different from the hypothetical population mean: $\mu \neq \mu_0$.

To decide between these two hypotheses, we need to find the probability of obtaining the sample mean (or one more extreme) if the null hypothesis were true. But finding this probability, or **p value**, requires first computing a test statistic called t. The test statistic is a statistic that is computed only to help find the p value. The formula for t is as follows:

$$t = \frac{(M - \mu_0)}{SE}$$

Here, M is the sample mean and μ_0 is the hypothetical population mean of interest. We can see the importance of the standard error statistic n calculating the key test statistic. that is calculated as described above.

When reporting the results of a t-test analysis, we will report the t-value, the associated p-value statistic and the degrees of freedom, df, for the analysis. Those additional numbers do not come from the basic t-statistic formula and below we will review how to extract these numbers from the output of a statistical program supporting experiment data analysis.

Without the p-value report, it is useful to know that a good rough rule of thumb for reliability in a t-test is that the t-statistic should be >2.0 for a result to be statistically reliable (exact probabilities depend on the number of participants). Conceptually this means that the difference between the mean and chance should be roughly twice the size of the SE. This rule of thumb provides an intuitive way to get a quick sense of how robust the experimental effects are within a dataset and helpse demonstrate why calculation of the SE is an important component of the descriptive statistics.

The Dependent-Samples t-Test

The **dependent-samples t-test**, also called the **paired-samples t-test**, is used to compare two means for the same sample tested at two different times or under two different conditions. This comparison is appropriate for pretest-posttest designs or within-participants experiments. We will discuss

this method of experimental design in Chapter 7.

The null hypothesis is that the means at the two times or under the two conditions are the same in the population. The alternative hypothesis is that they are not the same. This test can also be one-tailed if the researcher has good reason to expect the difference goes in a particular direction. For this analysis, it may help to think of the dependent-samples t-test as a special case of the one-sample t-test. The first step in the dependent-samples t-test is to reduce the two scores for each participant to a single difference score by taking the difference between them. This is a natural descriptive statistic to calculate for an intervention study where we measured the dependent variable before and after performing an intervention.

At this point, the dependent-samples t-test becomes a one-sample t-test on the difference scores. The hypothetical population mean (μ_0) of interest is 0 because this is what the mean difference score would be if there were no difference on average between the two times or two conditions. We can now think of the null hypothesis as being that the mean difference score in the population is 0 ($\mu_0 = 0$) and the alternative hypothesis as being that the mean difference score in the population is not 0 ($\mu_0 \neq 0$).

For our purposes here, the primary question is whether the experiment you are planning to do data analysis for is one that requires a dependent samples t-test or independent samples t-test. To this point, our standard simple experimental design is based on two groups of participants, who each received different levels of an independent variable before measuring a dependent variable. For this common case, the correct analysis is the independent samples t-test.

The Independent-Samples t-Test

The independent-samples t-test is used to compare the means of two separate samples (M_1 and M_2). The two samples might have been tested under different conditions in a between-participants experiment like our Experiment 1, or they could be pre-existing groups in a cross-sectional design (e.g., women and men, extraverts and introverts). The null hypothesis is that the means of the two populations are the same: $\mu_1 = \mu_2$. The alternative hypothesis is that they are not the same: $\mu_1 \neq \mu_2$. Again, the test can be one-tailed if the researcher has good reason to expect the difference goes in a particular direction.

The t statistic here is a bit more complicated because it must take into account two sample means, two standard deviations, and two sample sizes. The formula as taught in basic statistics classes is:

$$t = \frac{(M_1 - M_2)}{\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}}$$

Notice that this formula includes squared standard deviations (the variances) that appear inside the square root symbol, but these are calculated slightly differently than the SE of each group. Here the calculation of the variance is based on pooled variance, which reflects an assumption of equal variances across the experimental conditions.

You may have calculated a t statistic from sample data by hand previously in a course on statistics. In modern scientific practice this process would be too error-prone for us to confidently rely on the results. Instead, we use established, reliable software systems to calculate our t statistic, the associated p-value and the degrees of freedom for the analysis needed for reporting the results.

Degrees of Freedom (df)

To evaluate the probability with which we can reject the null hypothesis, we will need both the t-statistic and the **degrees of freedom (df)** for the analysis. For a two independent sample t-statistic under the assumption of equal variances, the df will always be the total number of participants minus 2. Conceptually, the more degrees of freedom there are, the more participants are included in the study and the more confident we can be of our conclusions. Because it is intrinsic to the statistical evaluation of data, the df must always be included in properly formatted reports of experimental results.

How many tails?

When first introduced to statistics, it is common to consider the question of whether the experimental hypothesis is **directional**. In theory, if we had a strong hypothesis that the experimental conditions would not just differ, but would differ in a specific direction, then we have a directional hypothesis. For Experiment 1, we expected that the deep encoding condition would lead to better memory than the shallow encoding condition. However, practically speaking, it is actually rare to have a full directional hypothesis. Even though we expected a specific kind of difference, if we had observed a difference in the other direction, we would have been very curious about that result and wanted to examine it further. Technically, if we approached our research with a strongly directional hypothesis, we would not further consider our data even if there were a large effect in the other direction.

We can make a simplifying assumption for the majority of experimental research in psychological science: **always perform a two-tailed test of statistical significance**. The two-tailed test is more rigorous in that it requires a larger difference between the groups to reach the reporting criterion for statistical significance. As a result, even if we use the two-tailed test when we could have justified a one-tailed test, we are slightly increasing

the possibility of a Type 2, false negative error, for the benefit of a big reduction in the possibility of the much worse Type 1, false positive, error.

It might be surprising to some readers that our belief about the directionality of the hypothesis affects the use of a statistical test. An important insight is that inferential statistics are not purely mathematical. The evaluation, interpretation and reporting of the math is quite strongly affected by the number and direction of the hypotheses underlying the research study. Later in Chapter 12, we will return to this idea in the context of **multiple parallel comparisons** in more complex analysis.

Assumption of equal variance

You will notice when carrying out a t-test that the math works slightly differently depending on whether you **assume equal variance** across the conditions of your experiment. A t-test can be carried out contrasting performance between conditions in which performance variability was very different. This is generally not a good idea. If one condition has more than twice the variability than the other, you should not assume equal variance but in most cases you should also not be running a t-test. If you did not hypothesis differing variance, you should review the experimental procedure to try to determine why this occurred. If you expected unequal variance, you probably should have been planning a more complex statistical approach outside the scope of the discussion here.

The guidelines here of strongly assuming equal variance and always use two-tailed tests reflect an approach based on simplifying the core approach to statistical inference consistent with most simple experimental design. Our goal in this class is to provide hands-on practice with the process and tools of psychological research in an accessible and tractable way. More complex analysis, including reconsideration of these assumptions, is common in psychological research and there are a wide range of more sophisticated tools for analysis that are applied in cutting-edge research. These advanced topics in research methods will build on the foundations presented here.

Running a t-test in R

Start by installing the R program and the RStudio suite (in 2 steps)

- Use the link below to go to R download page and choose the version that is compatible with your computer's operating system: <https://cran.r-project.org>
- Once R has downloaded, install it on your computer.
 - It requires permissions.
 - Accept the license.
 - Install all the default components.
 - Don't customize startup options.
 - Default additional tasks are fine.
- After R has been installed use this to download the RStudio version that is compatible with your operating system:
<https://rstudio.com/products/rstudio/download/#download>
 - If you are coming through the RStudio site, go to products, then RStudio Desktop. Use the Open Source Edition (Free).
 - Download will adjust to your OS. The Windows download is 171M, so be aware of bandwidth constraints and speed.
 - Current version is 1.3.
 - MS Windows complains to me that it isn't a Microsoft verified app. However, it is safe to install.
 - Once RStudio has downloaded, install it on your computer.
 - Note: You will not be able to install/run RStudio until R has been installed.

Use the RStudio program to start an analysis session

- Launch RStudio. You should see a screen with 4 panels. We will be primarily working with the left 2 panels.
- The top left panel will have lines of code, a 'script' for carrying out the steps

required for an analysis.

- The bottom left panel will have the output results of executing those steps, including error messages if something goes wrong.
- Use File -> Open and navigate to the folder on your computer where you've installed the files and associated data from our experiments
- Open the file provided for data analysis. This is an R script for testing your installation and re-running the t-test analysis from our Experiment 1 data for the Inclass experiment.
 - On a fresh install, this will produce a warning that there are required packages that are not installed. The option to install them is provided. You can also install them by working through the script analysis steps.
- Set the **working directory** to where your data are stored on your computer. If you have put the data file in the same folder as the analysis file, navigate to the Session menu, then to Set Working Directory and select the top option **To Source File location**.
- To run a single step of the analysis press the **Run** button that is in the upper right part of the top-left panel. This carries out the step in the script on which the cursor is currently. If you didn't do the installation of the 'psych' and 'ez' packages above, put the cursor on line 2 and Run. Then put the cursor on line 3 and Run.
- The installation process will also download and install a series of other packages needed (called dependencies). The process should only take a few minutes to run.
- Now move down to line 6, "library(psych)" and press Run. This loads a set of routines for data analysis for psychology experiment data that are helpful.
- The cursor moves down to the next line after each Run. Press it again to load the library on line 7, 8, and 9 ('psychTools', 'tidyR', and 'ez').
- With luck you are not getting error messages in the bottom left panel. If you are, something may have gone wrong with the above steps.
- The next step, line 12 will start loading our actual data. If everything is working you should see: "Data from the .csv file Inclass_Exp1_data_R.csv has been

loaded." In red in bottom left panel.

- Run on line 13 will cause the data table to be printed in another tab. It should look a lot like what the source data file looks like if you open it in Excel or another spreadsheet program.
- Run on line 17 to see the output of the describeBy function, which provides descriptive statistics for our data. You may notice that this needs to be unpacked a bit to find the key numbers, which are the Test_score values for each condition. Check that these numbers are identical to the descriptive statistics you calculated in your spreadsheet previously.
- Run on line 19 to carry out the two independent samples t-test for the data.

If everything works up to this point, then congratulations! You have just run your first formal analysis of experimental psychological data.

Writing up statistical reports

At this point, we are nearly done with the process of inferential statistics. The use of these statistics in experimental research is to support conclusions about our research study and the hypothesis about the underlying constructs. The final step in handling our inferential statistics is to format the output to follow the standard reporting format for an APA publication. This reporting format will include the t-statistic, degrees of freedom and the p-value that indicates the probability of the data occurring as observed under the null hypothesis. Any program to support calculation of statistical inference will provide that information, but every program tends to have its own unique way of formatting the output. Look through the output to identify those three key numbers and then format the results within the following basic frame:

$$t(df) = X.xx, p < 0.yy$$

In your report of the results, replace the **df** with the reported degrees of freedom, replace the **X.xx** with the reported t-statistic and replace the **yy** with the reported p-value.

In general, you will want to round the numbers to 2 significant digits no matter how many digits your output contains. Remember that the goal at this point is to be able to communicate the results of the experiment. With our rule of thumb that we expect $t>2$ for reliable results, we do not need to include a lot of additional numbers after the decimal point to make our case. Having a long list of irrelevant digits is actually harming our ability to effectively communicate the results.

Similarly, the point of reporting the p-value is to establish that it is $<.05$, the standard criterion for psychological science reporting. Very small p-values indicate that the results are reliable and that we may be extremely confident that the observed difference could not have occurred by chance. However, as a reminder, this has nothing to do with the more important validity of the experiment, which depends on the accuracy and rigor of the experimental design. We can establish reliability in our report with a reasonable number of

significant digits, usually two.

As we will see in Chapter 6, the information about the result of the calculation of the statistical test is included in the Results section of a report. Typically the information frame, with accurate numbers, is used to support a directional statement about the effect of the independent variable on the dependent variable. Making the statement directional is important for effective communication. For Experiment 1, it would be correct to state that the deep encoding condition led to higher scores on the recognition memory test. You should avoid non-directional statements like, the level of processing of words affected the recognition memory score. This form of statement is ambiguous in isolation and requires the reader to go back and recheck the descriptive statistics report to understand what the statistics are describing.

Because the direction of the results are the most important element of a report describing the outcome of a research study, it is good practice to support the descriptive statistics with a data visualization that shows the results graphically. Ideally effective presentation of complex data is beyond the scope of the text, but we will illustrate some simple principles with our Experiment 1 data. For this kind of simple design, simple line graphs or bar plots will be very effective at showing a reader what the results of the study were.

For this next section, a hands-on example of making a simple 2-bar plot will be shown. However, a critical element of this kind of scientific data visualization is the ability to show brackets on the data means that reflect the standard error of the sample. Unfortunately, Google Sheets does not currently provide this functionality. Here the process of adding error bars with Microsoft Excel is shown and it may be observed the adding the SE bars is the most complicated part of the process.

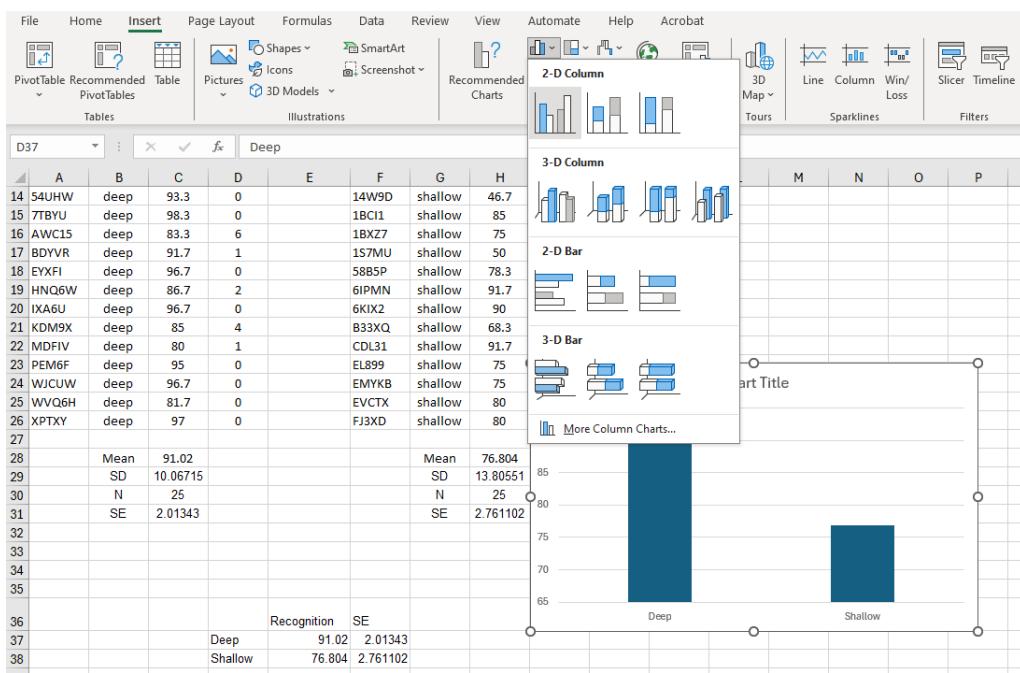
Data Visualization

A visualization of the experimental data that communicates the results of a research study is called a **Figure**. A well-made figure implicitly communicates both the descriptive and inferential statistics of the experimental data. As an example, we will make a Figure for our Experiment 1 data using MS Excel. Note that Excel calls data visualizations 'charts' but they are formally referred to as Figures in APA format.

For a simple 2-group design like our Experiment 1, a bar plot is a highly effective method for communicating results. Later we will consider the use of line graphs for illustrating more complex designs. In those more complex designs, we can compare the value of graphing data as bars or lines. There is no hard rule on which is better and the choice of how to present data is up to the author. The goal is to use a data visualization figure to help the reader understand the results, so it should be foremost in your plan for visualization design to ensure that it communicates effectively. Here we will illustrate some basic elements to include to accomplish this.

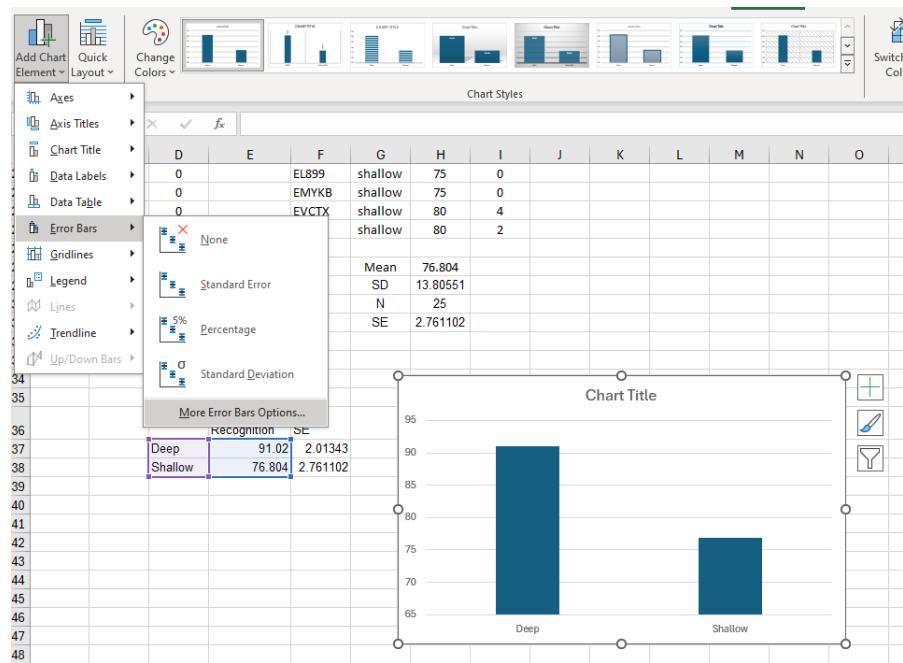
The first step towards creating a figure is generally to create a separate, labeled table of the key numbers that will contribute to the graph. The numbers we need for the graph will be the mean and the SE for each of the two conditions in our study: deep and shallow. For simplicity, we want these organized into a separate data table. In the figure below, you can see a labeled 2x2 table where the mean recognition memory percent correct scores are copied from the descriptive statistics to cells where they are adjacent at the bottom of the image.

In this image, those two cells and their labels are selected with the mouse and then in the Insert tab, the upper left option of Bar Charts is selected and then the upper left option of 2-D Columns is selected. As you do this, Excel already renders an initial image of what the starting bar plot is going to look like overlaid on the spreadsheet.

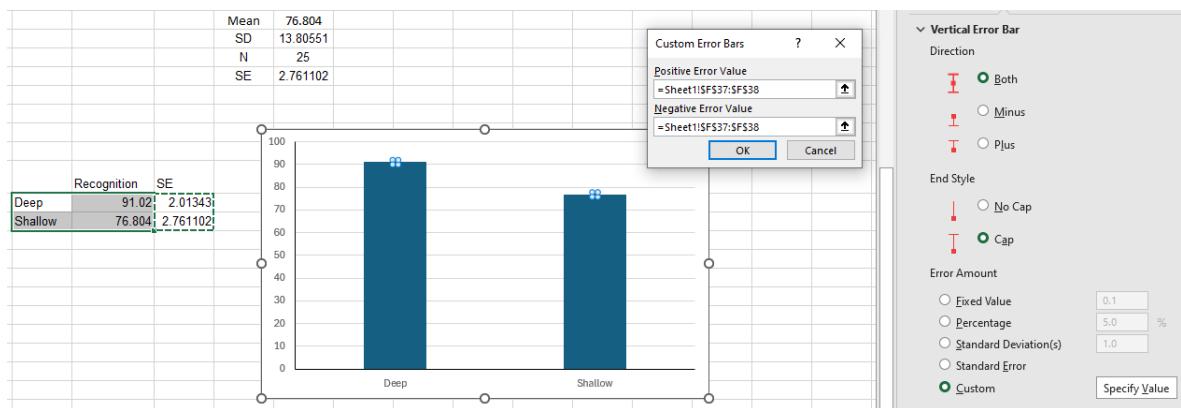


From this initial draft, we need to do some editing to the layout of this chart to make it effective and in approximately standard format. First, the Chart Title can be cut as we generally do not include titles on figures in manuscripts. Titles are used to help describe data in presentation formats, but APA reporting format requires that Figures be accompanied by a figure caption which is where the description of the illustration should be included. In addition, both the x- and y- axis should be drawn in black to ensure visibility of the axes. You will want to label the y-axis by adding the Chart Element, Axis Title -> Primary Vertical and then change the text label to Recognition Score. You may also optionally choose to remove the horizontal lines (these are chart elements called gridlines accessed through the Chart Design menu) or even change the color of the bars. You

Once the basic layout is set, the last element to be added is brackets reflecting the SE of mean. We kept these numbers near our means above, but note that we did not select those numbers when making the graph (if you have 4 bars in your graph, you may have selected them accidentally).



To add error bars correctly the size of the calculated SE, click on the graph and specifically one of the two bars. Then in the Add Chart Element menu, select the Error Bars option and the bottom choice, More Error Bars Options from there. In the Format Error Bars pane, choose Custom for your error bar size (bottom option) and select Specify Value. We will need to specify both the positive and negative sizes of the error bars, above and below the mean. For our data, these are the same sizes. For both the Positive Error Value and Negative Error Value choose the range where the SE values have been copied (F37:F38 above). Because they are next to each other, we can select both values and these will be correctly applied to both bars. The figure below shows roughly what this will look like. When done. select Ok.



This feature of setting the error bars flexibly to specific values for each group allows for correct presentation of both the means of the observed data, shown in the height of the bars, and the variance, shown in the SE bars. This puts many of the key descriptive statistics into the Figure visually. In addition, a useful trick is to look at the range implied by the error bars. For an independent samples experiment, if the error bars do not overlap (touch), then you most likely have a reliable difference between the groups. That means that the figure is also carrying some implicit information about the inferential statistics. You should always check or carefully include the actual statistical test in the reported text, but a well-made figure acts as a very effective overview of the results.

Figures need to be accompanied by a Figure Caption that provides some general explanation for how to read the figure. In general in the caption, you will want to explain the axes, that the dependent variable is shown on the y-axis, the conditions of the study as they are labeled, state the direction of the results and note that “brackets reflect the standard error of the means.”

Review and discussion of how to properly report your results in APA standard format will continue in Chapter 6.

Exercises

Report the results of Experiment 1 using sample or provided data from class the class following the guidelines in this chapter.

This should include:

- Descriptive statistics for both groups
- Inferential statistics about the difference in performance across groups
- A data visualization of the results. This should be a Figure, which includes a graph results, properly labeled, including standard error bars and a caption.

Note: Save these for inclusion in the report of Experiment 1 to be described in Chapter 6

6 Reporting in APA format

In this chapter, we look at how to write an APA-style empirical research report, an article that presents the results of one or more new studies. Recall that the standard sections of an empirical research report provide a kind of outline. Here we consider each of these sections in detail, including what information it contains, how that information is formatted and organized, and tips for writing each section.

The overall goal of scientific writing is to communicate the results of a research project to a wider audience. Once a successful study has been conducted, the researchers now know a new scientific fact about psychology and how the human mind works. The primary goal is now to disseminate this information to others. A secondary goal is to explain everything about how the study was run so that it can be evaluated with a more critical eye to look for challenges to the internal validity of the study. The challenge of effective scientific writing is to balance these two goals well without neglecting either of them.

Learning Objectives

1. Identify the major sections of an APA-style research report and the basic contents of each section.
2. Plan and write an effective APA-style research report.
3. The **Abstract**, a leading summary of about 250 words
4. The role and general content of the **Introduction** section
5. Explaining the **Methods**, including **Participants**, **Materials** and **Procedure**
6. Reporting the **Results**
7. The **Discussion**, where the **Results** are explained
8. The **References** used to list publications cited to support the report

The audience

Writing effectively in any context requires an understanding of the audience for whom the writing is intended. In a laboratory carrying out active research, this question is often raised around different kinds of journals a research report might be published in. Some journals aim for a broad audience (e.g., Psychological Science) and the description of the research needs to be kept accessible to the whole range of psychological researchers. Some journals are more specialized (e.g., Memory & Cognition) and the writing can assume some more familiarity with some parts of the background theory. Audience is an even more important consideration when planning other kinds of research presentations such as conference posters, colloquium talks or short “brown bag” research presentations.

For classroom projects, a common challenge is to avoid writing for the lecturer or the teaching assistant in the class. While classroom writing is generally evaluated by teachers, the style of writing to that audience often leads the writing style in the wrong direction. For example, students are

Writing with style

Writing to APA format is somewhat more formal than other types of prose. It is generally a good idea to avoid use of the first person, although it is not specifically prohibited. If there is an opportunity where first person is more effective than to use passive voice, it will generally be the first person plural – virtually no scientific work is ever done alone in modern science.

Scientific writing in general tends to be very compact, concise, and precise. Some sections, like the Results of a study, will be very short but must be written and proofread extremely carefully. Unlike other types of writing, the best scientific writing often takes very few words overall but can take much longer to prepare than more verbose forms.

aware that other people in the class are fully aware of all the methodological details of the research and then omit proper descriptions of the procedure. Better is to aim to describe your work to students outside the class. With that audience in mind, the overall writing style is generally more effective.

The major sections of a research report are included in the order listed below. However, there is no requirement to write them in that order. A common approach to building a manuscript is to start with the description of the results and any data visualizations used to present these, first preparing the Results and Figures sections. Then write the Methods section, which precedes that section in the manuscript. As noted below, the goal of the Methods section is to provide enough detail for a reader to replicate the research project and the important findings of the Results section highlights the key methodological techniques to explain. Then, having those sections drafted, write the Discussion section and summarize the major conclusions of the

report. Only then, when the major points to communicate about the findings are clear, write the Introduction to the whole paper with a clear idea in mind of the audience for whom the manuscript is intended. Lastly, after the rest of the paper is drafted, write the Abstract.

Writing the sections in the above order is not a requirement but illustrates an important aspect of scientific writing distinct from many other forms of writing in that it is not prepared with a top-to-bottom overall flow. After the sections are written, it is critical to read and proofread the paper in order from Abstract to Conclusions, but effective scientific writing has a formality in the report that is very different from other kinds of writing. In contrast, scientific journalism is written with a very different style that is more focused on flow of the writing and the story behind the research. Journalism is not written to meet the second major goal of scientific writing: to provide enough detail to allow for a critical review of the validity of the research. The APA format is designed to meet that criterion and completeness of description is required. As an example, the audience for a media report of a scientific finding would not need the statistical reports written to include the type of test, degrees of freedom and p-value. All of those details are critical for the scientific audience of a report written to APA format.

Sections of a Research Report

The overall organization of a research report will be broken down as into sections with the main goals of each of these and the main elements that should be included in each section. Descriptions of these sections will be supported by examples from Sample One Experiment Paper.

Title Page

An APA-style research report begins with a title page. The title is centered in the upper half of the page, with each important word capitalized. The title should clearly and concisely (in about 12 words or fewer) communicate the primary variables and research questions. This sometimes requires a main title followed by a subtitle that elaborates on the main title, in which case the main title and subtitle are separated by a colon. Here are some titles from recent issues of professional journals published by the American Psychological Association.

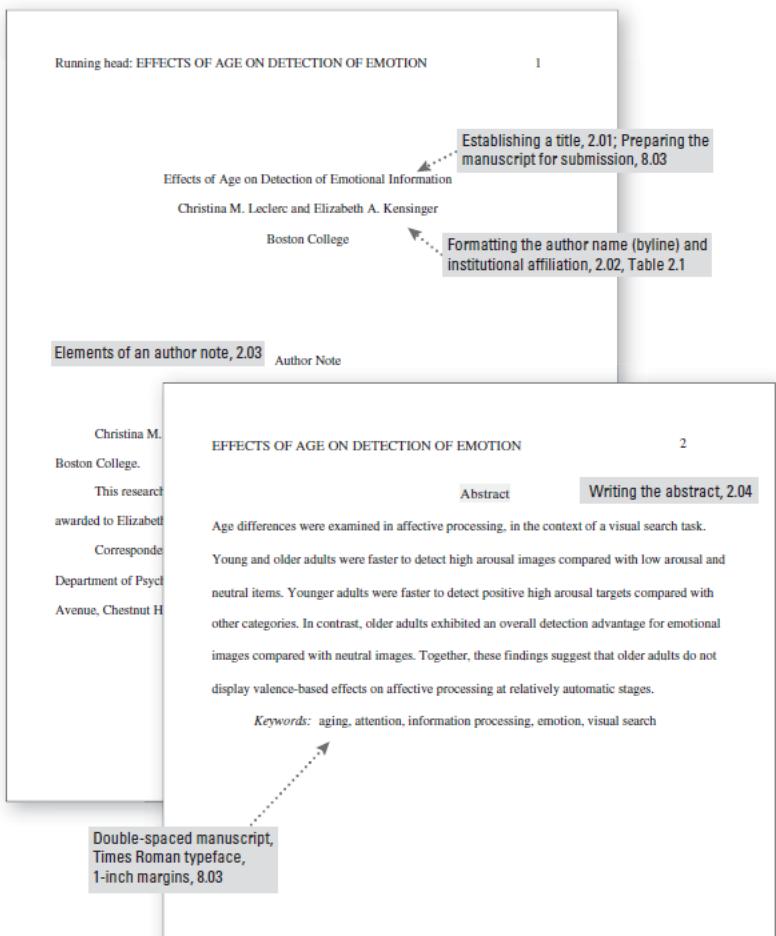
- Sex Differences in Coping Styles and Implications for Depressed Mood
- Effects of Aging and Divided Attention on Memory for Items and Their Contexts
- Computer-Assisted Cognitive Behavioral Therapy for Child Anxiety: Results of a Randomized Clinical Trial
- Virtual Driving and Risk Taking: Do Racing Games Increase Risk-Taking Cognitions, Affect, and Behavior?

Below the title are the authors' names and, on the next line, their institutional affiliation—the university or other institution where the authors worked when they conducted the research. As we have already seen, the authors are listed in an order that reflects their contribution to the research. When multiple authors have made equal contributions to the research, they often list their names alphabetically or in a randomly determined order.

For articles that are being submitted for publication, the title page also includes an author note that lists the authors' full institutional affiliations,

any acknowledgments the authors wish to make to agencies that funded the research or to colleagues who commented on it and contact information for the authors. For student papers that are not being submitted for publication—including theses—author notes are generally not as detailed but should include at least the lead authors email address.

Figure 2.1. Sample One-Experiment Paper (The numbers refer to numbered sections in the *Publication Manual*.)



Manuscript format

Note that the sample manuscript does not look like a publication. The text is double spaced and the content spread across pages. This is the expected format for manuscripts being evaluated and considered for publication. This is the style to write for here.

Abstract

The abstract is a summary of the study. It is the second page of the manuscript and is headed with the word Abstract. The first line is not indented. The abstract presents the research question, a summary of the method, the basic results, and the most important conclusions. Because the abstract is usually limited to about 200 words, it can be a challenge to write a good one.

As a general heuristic, the abstract will have sentences that attempt to encapsulate each of the sections of the main report. The first sentence or two is generally a summary of the key ideas from the Introduction. For example, one sentence to introduce the key constructs that the experiment is about and one sentence to describe the hypothesis. For a report of a single experimental study, there will be one sentence summarizing methods and one sentence summarizing the results. In a typical 200-300 abstract, that leaves one sentence to summarize the main conclusions drawn from the findings.

Compactly summarizing the major sections of the report is not an easy task. It is often a good idea to write the Abstract last when preparing a research report (note that it still goes on page 2). Once you have written the other sections effectively, that shows how to prepare concise one to two sentence versions of the major sections for inclusion in the abstract.

Introduction

The content, structure and length of the Introduction section will vary across different journals that publish research reports that describe psychological science studies. Here we will focus on the style popular in journals that have come to focus on relatively shorter-form publications. Shorter research reports typically have much shorter Introduction and Discussion sections and focus more on the details of the Methods and Results. In most cases, only one or two experiments are described, although in some cases with a series of very similar designs you might see as many as five experiments presented

extremely compactly.

The shorter form publication was originally aimed at speeding the process of bringing science from the laboratory to a final published form to disseminate the findings. A shorter manuscript can be written more quickly and more rapidly reviewed by experts in the field through the peer-review process (see below). Shorter articles can also be more rapidly digested and understood by researchers interested in the topic of the reported studies, if it is well-written. However, it should be noted that preparing a highly effective, comprehensive, concise and precise report of research can still be a time-consuming process. Unlike some other forms of writing, the resulting length is not commensurate with the time invested in the writing process.

Even in a short Introduction section, the goal of this part of the report is to introduce the main theoretical question and the experimental hypothesis. This will often rely on connections to prior published research on which the current report will build. The operational definitions of the main constructs have to be explained and tied to the design being used so the reader can see how the design will test the hypothesis and what we will learn from the results.

Although the rationale for a study can be framed as a theoretical question, there is generally no suspense in scientific writing. The answer to the main question is usually provided in the abstract or the title, so a research report is not constructed as a build up to a big reveal. Scientific writing is about accurate description of the facts of the study, the methods employed and observations collected. This will make it generally somewhat dry compared to other writing forms, but is done to emphasize the need for accuracy and precision so that we can rely on the inferences drawn from the research study.

The introduction begins on the third page of the manuscript. The heading at the top of this page is the full title of the manuscript, with each important word capitalized as on the title page. The introduction first introduces the research question and explains why it is interesting, and provides the reader

with the key constructs. In our earlier experimental analysis, when we ask the first question “what is this research about?” the answer should be provided here clearly and early on.

Defining the key terms and constructs will often be supported by citing previously published research that has used similar ideas. Building on the theory and definitions used by other researchers strengthens the presentation of scientific results. Any report has to have new elements, of course, but the main source of novelty in scientific writing is the data and the new methodological elements. Psychological science is about better understanding of constructs like *memory*, which are intuitive, but the goal of research is to better characterize what exactly memory is, how it works, what factors affect its operation.

Citations to previous research are used to ground the results being presented in prior peer-reviewed research. This is always done by listing the authors by last name and the year, e.g., Craik & Tulving (1975). First names of the authors, their affiliations and the title of the published work are never included in the manuscript text. These go exclusively in the References section at the end of the report. This is an important distinction with more journalistic styles of writing but an important stylistic aspect of APA format to do correctly. When referencing prior work, the preferred method is cite the authors’ work and paraphrase their findings. Avoid long quotes or frequent short quotes.

The closing of the introduction—typically the final paragraph or two—usually includes two important elements. The first is a clear statement of the main research question and hypothesis. This statement tends to be more formal and precise than in the opening and is often expressed in terms of operational definitions of the key variables. Very often this is a restating of the core question or hypothesis framed in terms of the variables instead of the constructs. This will lead into the second element, which is a brief overview of the method and some comment on its appropriateness. With this, the introduction leads smoothly into the next major section of the article—the method section.

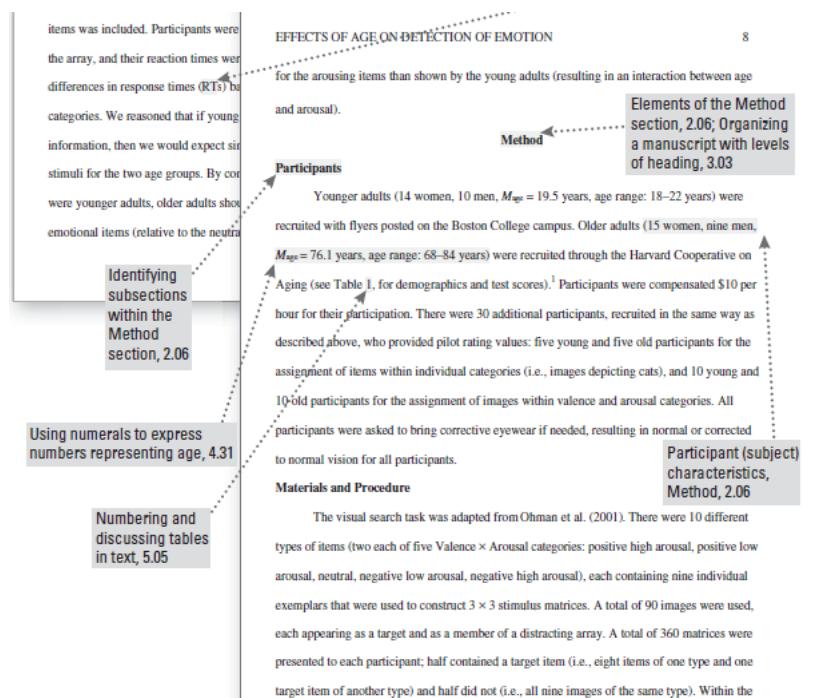
Method

The method section is where you describe how you conducted your study. An important principle for writing a method section is that it should be clear and detailed enough that other researchers could replicate the study by following your “recipe.” This means that it must describe all the important elements of the study—basic demographic characteristics of the participants, how they were recruited, whether they were randomly assigned to conditions, how the variables were manipulated or measured, how counterbalancing was accomplished, and so on. At the same time, it should avoid irrelevant details such as the fact that the study was conducted in Classroom 37B of the Industrial Technology Building or that the questionnaire was double-sided and completed using pencils.

The method section begins immediately after the introduction ends with the heading **Method** centered on the page. Immediately after this is the subheading **Participants**, left justified and in italics. The participants subsection indicates how many participants there were, the number of women and men, some indication of their age, other demographics that may be relevant to the study, and how they were recruited, including any incentives given for participation. The participants section is a necessary subsection of the Methods section and should always go first.

After the participants section, additional subsections may be titled **Materials**, **Design** and/or **Procedure**. The **Materials** section gives the characteristics of any stimuli used in the study, e.g., the words studied, pictures shown, questionnaires or surveys employed. If any special equipment was used for data collection, it would go here. In cases where the research used established questionnaires or procedures that do not require specific stimuli, then the Materials section may be skipped or combined into a **Materials and Procedure** section.

The **Design** section tends to be used with complex experimental designs or in health or clinical research to characterize exactly how the recruiting and assignments to conditions was carried out. It is less common in the kinds of



relatively simple experiments we primarily focus on here.

The next section explain how the experimental protocol was carried out and is necessary for the report. It is generally called the **Procedure** section, although can be called **Design and Procedure**. All the details of the experimental procedure need to be fully described here. The guiding principle is that a reader should be able to carry out a replication of the procedure from the description.

The description will always be in terms of the operational definitions of the constructs since these define how the main ideas were implemented. The procedure is how the study was carried out. It often works well to describe the procedure in terms of what the participants did rather than what the researchers did. For example, the participants gave their informed consent, read a set of instructions, completed a block of four practice trials, completed a block of 20 test trials, completed two questionnaires, and were debriefed and excused. All of that information should be included in the Procedure section.

Results

The **Results** section is where you present the main results of the study, including the results of the statistical analyses. Although it does not include the raw data—individual participants' responses or scores—researchers should save their raw data and make them available to other researchers who request them. Many journals encourage the open sharing of raw data online, and some now require open data and materials before publication.

Although there are no standard subsections, it is still important for the results section to be logically organized. Typically, it begins with certain preliminary issues. One is whether any participants or responses were excluded from the analyses and why. The rationale for excluding data should be described clearly so that other researchers can decide whether it is appropriate. A second preliminary issue is how multiple responses were combined to produce the primary variables in the analyses. For example, if participants rated the attractiveness of 20 stimulus people, you might have to explain that you began by computing the mean attractiveness rating for each participant. Or if they recalled as many items as they could from study list of 20 words, did you count the number correctly recalled, compute the percentage correctly recalled, or perhaps compute the number correct minus the number incorrect? A final preliminary issue is whether the manipulation was successful. This is where you would report the results of any manipulation checks.

The results section should then tackle the primary research questions, one at a time. Again, there should be a clear organization. For studies with complex designs such as multiple dependent and independent variables, the Results section should be organized carefully and with a clear plan. One approach would be to answer the most general questions and then proceed to answer more specific ones. Another would be to answer the main question first and then to answer secondary ones.

For relatively simple studies, the Results section can be written in a very compact way. Report the descriptive statistics about the participants

performance by group/condition. Then report the inferential statistics supporting a claim about the difference between the groups. Remember to include the direction of the difference – do not just say that group performance was reliably different, tell the reader which group performed better.

However the Results section does not generally include any additional explanation of the meaning of the data with respect to the hypotheses. Interpretation and evaluation of the findings will go in the next section, **Discussion**. For more complex designs, some explanation of how the analysis connects to the hypothesis may be necessary to communicate the results effectively.

For simple designs, like Experiment 1, the Results section can be very compact. The main elements to include are the descriptive statistics, the inferential statistics and statement of the finding. This can be accomplished in just a few sentences in many cases.

Supporting the numerical report, it is often very helpful to have a visualization of the results to support the presentation of the quantitative data in the text. A bar or line graph can very clearly communicate the findings to the reader. These are included as Figures in the manuscript. If you include a Figure, make sure you reference the Figure in the appropriate place in the Results section where you have the numbers that relate to the graph. Figures should be planned to illustrate the results and include labels on the axes that show the units of the dependent variable (typically on the y-axis) and the grouping conditions of the independent variable (typically on the x-axis). For complex designs, they may also include a Legend in the figure and/or use color to clarify the grouping.

All Figures must have a Figure Caption included with the figure to help the reader understand the content. The caption explains the axes, characterizes the results and explains visualization details such as the use of SE bars to communicate the observed variance. In the official APA guidelines, Figures and Captions are included as separate pages at the end of the manuscript.

It is also acceptable to include the figures “in line” on pages near the Results section. Note that if you include the figures “in line” make sure the formatting is readable, e.g., keep the figure and caption together on a page (sometimes this requires making these a separate page to keep document software from moving things around).

Tables of data can also be helpful ways of communicating performance across a complex set of conditions. Tables should only be used to present average performance (and variance measures) across conditions and should virtually never include all the individual participant data for a study. Tables are generally not needed in simple designs where the scores on the DV can be easily described in the text in a sentence format. In general, you should choose between a table or a figure to help the reader understand the data and it is very rare that redundant presentation of both formats is helpful.

Discussion

The **Discussion** is the last major section of the research report. Discussions usually consist of the following elements:

- Summary of the research
- Theoretical implications
- Practical implications
- Limitations and possibilities for future research
- Final positive statement about what was gained from the research

The discussion typically begins with a summary of the study that provides a clear answer to the research question. In a short report with a single study, this might require no more than a sentence. In a longer report with multiple studies, it might require a paragraph or even two. The summary is often followed by a discussion of the theoretical implications of the research. Do the results provide support for any existing theories? If not, how can they be explained? Although you do not have to provide a definitive explanation

or detailed theory for your results, you at least need to outline one or more possible explanations. In applied research—and often in basic research—there is also some discussion of the practical implications of the research. How can the results be used, and by whom, to accomplish some real-world goal?

The theoretical and practical implications are often followed by a discussion of the study's limitations. Perhaps there are problems with its internal or external validity. Perhaps the manipulation was not very effective or the measures not very reliable. Perhaps there is some evidence that participants did not fully understand their task or that they were suspicious of the intent of the researchers. Now is the time to discuss these issues and how they might have affected the results. But do not overdo it. All studies have limitations, and most readers will understand that a different sample or different measures might have produced different results. Unless there is good reason to think they would have, however, there is no reason to mention these routine issues. Instead, pick two or three limitations that seem like they could have influenced the results, explain how they could have influenced the results, and suggest ways to deal with them.

Many discussions end with some suggestions for future research. If the study conclusions were accompanied by substantial limitations, how might these be addressed in future work? What new research questions has the study raised? This part of the discussion, however, is not just a list of new questions. It is a discussion of two or three of the most important unresolved issues. This means identifying and clarifying each question, suggesting some alternative answers, and even suggesting ways they could be studied.

At the very ending of the Discussion, it is usually effective to make a strong positive statement about the main importance of the findings. Because the preceding paragraphs will list limitations and potential future work, it is good to return to what the strengths of the current report are.

References

The references section begins on a new page with the heading “References” centered at the top of the page. All references cited in the text are then listed in the format presented earlier. They are listed alphabetically by the last name of the first author. If two sources have the same first author, they are listed alphabetically by the last name of the second author. If all the authors are the same, then they are listed chronologically by the year of publication. Everything in the reference list is double-spaced both within and between references.

Online sources can help with preparing references in APA format. The following is an example provided by Google Scholar. The reference format should be followed closely including the authors’ names and initials (but not full names), the publication year, the title, journal and page numbers.

Gino, F., & Wiltermuth, S. S. (2014). Evil genius? How dishonesty can lead to greater creativity. *Psychological science*, 25(4), 973-981.

Appendices, Tables, and Figures

Appendices, tables, and figures come after the references in standard format (but see above for alternate approaches with figures and tables). An appendix is appropriate for supplemental material that would interrupt the flow of the research report if it were presented within any of the major sections. An appendix could be used to present lists of stimulus words, questionnaire items, detailed descriptions of special equipment or unusual statistical analyses, or references to the studies that are included in a meta-analysis. Each appendix begins on a new page. If there is only one, the heading is “Appendix,” centered at the top of the page. If there is more than one, the headings are “Appendix A,” “Appendix B,” and so on, and they appear in the order they were first mentioned in the text of the report.

After any appendices come tables and then figures. Tables and figures are

both used to present results. Figures can also be used to display graphs, illustrate theories (e.g., in the form of a flowchart), display stimuli, outline procedures, and present many other kinds of information. Each table and figure appears on its own page. Tables are numbered in the order that they are first mentioned in the text ("Table 1," "Table 2," and so on). Figures are numbered the same way ("Figure 1," "Figure 2," and so on). A brief explanatory title, with the important words capitalized, appears above each table. Each figure is given a brief explanatory caption, where (aside from proper nouns or names) only the first word of each sentence is capitalized. More details on preparing APA-style tables and figures are presented later in the book.

Peer-review

As a reader of research, you will mainly be a consumer of research publications presented in a final, type-set format that you have seen when reading the literature. The spacing, layout and format of the writing here is aimed at preparation of a manuscript that would be sent to a research journal to be considered for publication.

Manuscripts are generally evaluated through a **peer-review** process where other authors and experts in the related area of science evaluate and provide suggestions for improving the presentation of research. When reading the literature, it is important to be aware of situations where you may be encountering scientific descriptions from sources that are not peer-reviewed. Sometimes news or social media reports are written based on preliminary data, press releases or conference presentations. Because these reports have not yet been subject to peer-review, they should not be relied on. Claims from preliminary data, pre-review, should always be treated with a very healthy skepticism until the formal scientific report has undergone the proper review process preceding publication.

Other Presentation Formats

One of the ways that researchers in psychology share their research with each other is by presenting it at professional conferences. (Although some professional conferences in psychology are devoted mainly to issues of clinical practice, we are concerned here with those that focus on research.) Professional conferences can range from small-scale events involving a dozen researchers who get together for an afternoon to large-scale events involving thousands of researchers who meet for several days. Although researchers attending a professional conference are likely to discuss their work with each other informally, there are two more formal types of presentation: oral presentations ("talks") and posters. Presenting a talk or poster at a conference usually requires submitting an abstract of the research to the conference organizers in advance and having it accepted for presentation—although the peer review process is typically not as rigorous as it is for manuscripts submitted to a professional journal.

In an **oral presentation**, the presenter stands in front of an audience of other researchers and tells them about their research—usually with the help of a slide show. Talks usually last from 10 to 20 minutes, with the last few minutes reserved for questions from the audience. At larger conferences, talks are typically grouped into sessions lasting an hour or two in which all the talks are on the same general topic.

In preparing a talk, presenters should keep several general principles in mind. The first is that the number of slides should be no more than about one per minute of the talk. The second is that talks are generally structured like an APA-style research report. There is a slide with the title and authors, a few slides to help provide the background, a few more to help describe the method, a few for the results, and a few for the conclusions. The third is that the presenter should look at the audience members and speak to them in a conversational tone that is less formal than APA-style writing but more formal than a conversation with a friend. The slides should not be the focus of the presentation; they should act as visual aids. As such, they should present the

main points in bulleted lists or simple tables and figures.

Another way to present research at a conference is in the form of a **poster**. A poster is typically presented during a one- to two-hour poster session that takes place in a large room at the conference site. Presenters set up their posters on bulletin boards arranged around the room and stand near them. Other researchers then circulate through the room, read the posters, and talk to the presenters. In essence, poster sessions are a grown-up version of the school science fair. But there is nothing childish about them. Posters are used by professional researchers in all scientific disciplines and they are becoming increasingly common.

Posters are typically a large size, maybe four feet wide and three feet high. The poster's information is organized into distinct sections, including a title, author names and affiliations, an introduction, a method section, a results section, a discussion or conclusions section, references, and acknowledgments. Although posters can include an abstract, this may not be necessary because the poster itself is already a brief summary of the research.

Given the conditions under which posters are often presented—for example, in crowded ballrooms where people are also eating, drinking, and socializing—they should be constructed so that they present the main ideas behind the research in as simple and clear a way as possible. The font sizes on a poster should be large—perhaps 72 points for the title and authors' names and 28 points for the main text. The information should be organized into sections with clear headings, and text should be blocked into sentences or bulleted points rather than paragraphs. It is also better for it to be organized in columns and flow from top to bottom rather than to be organized in rows that flow across the poster. This makes it easier for multiple people to read at the same time without bumping into each other. Posters often include elements that add visual interest. Figures can be more colorful than those in an APA-style manuscript. Posters can also include copies of visual stimuli, photographs of the apparatus, or a simulation of participants being tested. They can also include purely decorative elements, although it is best not to

overdo these.

Again, a primary reason that posters are becoming such a popular way to present research is that they facilitate interaction among researchers. Many presenters immediately offer to describe their research to visitors and use the poster as a visual aid. At the very least, it is important for presenters to stand by their posters, greet visitors, offer to answer questions, and be prepared for questions and even the occasional critical comment. It is generally a good idea to have a more detailed write-up of the research available for visitors who want more information, to offer to send them a detailed write-up, or to provide contact information so that they can request more information later.

Other Types of Manuscripts

We have focused here primarily on preparing a report of experimental research formatted to the APA scientific writing standard. When preparing a review of the literature to support writing a report, or proposing a research project, there are a few other types of writing that you may encounter.

Some peer-reviewed journals publish Review and/or Theoretical articles that summarize research on a particular topic across a number of published studies, but without presenting new empirical results. Review and theoretical articles are structured much like empirical research reports, with a title page, an abstract, references, appendixes, tables, and figures, and they are written in the same high-level and low-level style. Because they do not report the results of new empirical research, however, there is no method or results section.

Review articles are excellent sources for covering a lot of research quickly and understanding the range of operational definitions and methodologies in common use in a research subdomain. They often lay out the key framework theories that establish the foundation for framing specific research questions. However, they frequently leave out fine-grained details about the implementation of the procedures employed in the research cited. Starting

with a review article to inspire interest and generate research questions is an excellent approach. This should be followed up by seeking out some of the main cited publications that more thoroughly document the research protocol to either compare to results being written or to provide detail for a proposal.

When reviewing literature, it is important to discriminate between peer-reviewed reports and publications that have not been through a review process. Some theoretical or review papers are published in the form of book chapters in collected volumes. You should be aware that those publication outlets are frequently not peer-reviewed. Book chapters can serve the same role of inspiration and general understanding in a research area, but should always be followed by careful review of the cited research, which should have been published following peer-review.

Journalism and social media descriptions of research are also excellent sources of inspiration and interest, but should never be directly relied on or cited. It is important to discriminate between magazines like Psychology Today, which is a journalism based outlet and not peer-reviewed, and Psychological Science, which is a top-tier peer-reviewed research publication. New and exciting research findings are also often reported in various media and social media outlets. These are not cited directly in standard format, but should lead you to the underlying peer-reviewed publication, which you should be able to locate, read and use as background.

It has been suggested that scientists should engage more with these more accessible forms of describing and disseminating research. Making science more available to the broader public is a valuable goal, increasing scientific literacy. The challenge inherent in this is to balance the precision and rigor of the scientific writing style with a more casual and informal language aimed at non-specialists. There are many very high-quality researchers who have developed talents for this kind of communication. However, popular science outlets currently have no built-in protection from low-quality or inaccurate scientific summaries which creates problems from sloppy or even unethical publications in these forms.

Key Takeaways

- APA style is a set of guidelines for writing in psychology. It is the genre of writing that psychologists use to communicate about their research with other researchers and practitioners.
- APA style can be seen as having three levels. There is the organization of a research article, the high-level style that includes writing in a formal and straightforward way, and the low-level style that consists of many specific rules of grammar, spelling, formatting of references, and so on.
- References and reference citations are an important part of APA style. There are specific rules for formatting references and for citing them in the text of an article.
- In APA-style empirical research report consists of several standard sections. The main ones are the abstract, introduction, method, results, discussion, and references.
- The introduction consists of an opening that presents the research question, a literature review that describes previous research on the topic, and a closing that restates the research question to connect to the methodology.
- The method section describes the method in enough detail that another researcher could replicate the study. At a minimum, it consists of a participants subsection and a design and procedure subsection.
- The results section describes the results in an organized fashion. Each primary result is presented in terms of statistical results but also explained in words.
- The discussion typically summarizes the study, discusses theoretical and practical implications and limitations of the study, and offers suggestions for further research.
- Research in psychology can be presented in several different formats. In addition to APA-style empirical research reports, there are theoretical and review articles; final manuscripts, including dissertations, theses, and student papers; and talks and posters at professional conferences.

Exercises

Question 1. Which of the following is the main goal of the methods section of a research report?

- a. Meticulously articulate how you analyzed the data.
- b. Provide enough detail to allow an independent researcher to replicate your study.
- c. Outline the demographic information of your participants so that reviewers can access the generalizability of your research.
- d. Discuss the procedure you used so that readers can decide for themselves if your protocol is biased.

Question 2. Which of the following is usually beyond the scope of the results section of a quantitative research report?

- a. Discussing what statistical techniques were used
- b. Presenting figures and/or tables to portray the data
- c. Providing detailed interpretation of the implications based on the data
- d. Presenting specific statistics that were generated from the data

Question 3. If you state alternative explanations in your discussion, which of the following should you also consider doing?

- a. Tell readers why the alternative explanation falls short of the primary explanation
- b. Conduct statistics tests to test them specifically
- c. Include reviewer opinions of whether they think the alternative explanation is better or worse than the primary explanation

- d. Present a literature review that would allow readers to conduct a follow-up study based on the alternative explanation

Question 4. What is the role played by answering the question "who will you write for" in writing a research report? How should the answer influence the writing process?

Question 5. Suppose you encountered for the following in a manuscript that was intended to be written to APA format:

In their seminal 1972 paper titled "Levels of Processing: A Framework for Memory Research", Fergus Craik and Robert Lockhart claim "Over the past decade, models of human memory have been dominated by the concept of stores and the transfer of information among them..."

Name 3 things wrong stylistically about this writing that make it incorrect for an APA style report.

7 Within-Participants Design

Across the following two lists, say out loud the **color in which the text is presented** (not the word meaning), first on the left, then on the right:

RED
CORAL
TURQUOISE
PURPLE
AMBER

BLUE
CRIMSON
PERIWINKLE
ORANGE
GREEN

Did you notice anything different about the two conditions? Did one of the groups seem to be easier or faster to name the color of the text?

If you thought the words in the right group seemed to take more effort or take longer, then you have experienced a brief replication of the classic Stroop Effect. When trying to name the color of words when the words themselves are color names, there is a tendency to want to say the word out loud instead of naming the color. For example, for the top word on the right list, you might have accidentally said BLUE out loud instead of correctly saying RED (or crimson), which is the color in which it is shown. For the words in the left group, the top word meaning matches the color in which it is shown.

The theory of why the Stroop effect occurs is that we have a tendency to read words due to a lifetime of practice reading, so when we encounter the word BLUE written in red, we automatically want to say "blue" instead of ignoring the word to correctly name the color. The incongruity between the word meaning and presentation color creates interference between the prepotent response and the task, which leads to more effortful and slower responses. No such interference occurs for the words on the left, nor would there be interference from words that were not color words. The use of this paradigm to illustrate this interference has led to many hundreds of studies of this phenomenon to attempt to better understand attention, perception, automaticity and cognitive interference.

This design also serves here as an example of a within-participants design. Just as we have seen in Chapters 3 and 4, there is an independent variable that varies across two levels. In one condition, the color and word meaning match and in the other condition, these are incongruent. However, the design asks the participant to do both conditions of the study rather than assigning half of the participants to the congruent condition and half to the incongruent condition.

In a formal study, we would measure the reaction time to read the words aloud as our dependent variable. We hypothesize that this measure of color-naming speed would be slowed, leading to longer reaction times, for the incongruent condition compared to the congruent condition. The logic of the design is the same as the two independent samples designs previously discussed, but a within-participants design helpfully illustrates several

differences in this methodological approach.

In addition to the design, a study of the Stroop effect requires a protocol for measuring reaction time, the time taken to name the color aloud. Naming a single color word is too rapid to time with a crude method like a stopwatch. In modern practice, we might use a microphone to record participants and precisely measure the time between presenting the color-word on screen and

Using a stopwatch

To use a stopwatch in an experiment, a member of the research team has to be present while the participant performs the task. When starting the reading task, the experimenter starts the stopwatch and then stops it as the participant finishes the list.

The timing is likely imperfect but as long as the extraneous variable of measurement noise is not confounded with the experimental condition, the design should not suffer from a Type 1 error.

However, use of a stopwatch is famously vulnerable to experimenter bias. Knowing the hypothesis and condition, the experimenter might press the stop button a little faster.

How would we design our procedure to avoid this?

the participant's saying the color out loud. The Stroop paradigm, however, dates back to the early 20th century when this technology was not available and instead used the approach of giving participants a printed page with 20-30 words to be read aloud and using a stopwatch to time the whole page.

Even though each color name is only said aloud fractions of a second longer, across many color words, the difference accumulates to be large enough to be measurable with a stopwatch. This provides a low-technology procedure for estimating reaction time that would be suitable for a simple Stroop design.

In this design, we would ask participants to read a page of color-words that were incongruent with the word

meaning and then read another page of color-words that were congruent. Now we should consider what extraneous variables are necessary to control for our study to be rigorous and likely to succeed. We can apply the principle of constancy to the words to be read aloud by having the same colors to be named in both conditions. That should avoid additional variance from aspects like the word *purple* being slower to articulate than *red*. This would be done by using the same colors for the text across lists and only varying the text of the word so that it either matches the color or is incongruent.

One element of the design that cannot be held constant for each participant is the fact that the two conditions are done sequentially and therefore there is the potential for an **order effect**. Order effects are a specific kind of extraneous variable where there is a risk that task performance and the dependent variable change across time. We can imagine the possibility of two common order effects on our design. Participants might get faster at color-word naming with practice across the two pages, producing a **practice effect** that leads to lower scores on the second condition regardless of whether it is congruent or incongruent. Alternately, participants might get tired of doing color-word naming, producing a **fatigue effect**, leading to slower responses on the second page, again regardless of condition.

Order effects are intrinsic to within-participant designs. We can take the approach of counter-balancing the order across participants, by having half the participants first perform the incongruent color-word naming and the other half receive the congruent stimuli first. However, if there is a substantial difference in performance from the first to the second condition, it may be necessary to treat this as another independent variable rather than a simple tool of experimental control. We will discuss designs of this level of complexity in Chapter 10.

Another set of extraneous variables we might consider would be ones related to differences in our participants with respect to the task. Some participants might be more knowledgeable about unusual colors and color-words than others, for example, for the unusual colors included in the example like periwinkle. Others might have differences in basic color perception leading

them to see different shades of color due to perceptual issues related to color-blindness. Some might read faster or slower. However, all of these participants variables are perfectly balanced in our within-participants design because we have exactly the same participants in both groups. The faster readers will identify both congruent and incongruent colors and should still exhibit the hypothesized effect about the difference in performance.

The strength of a within-participants design is in excellent control of the participant variables. The greatest challenge to within-participants design is that they often require administering the conditions in order, leading unavoidable to concerns about order effects.

Design answer

As described in Chapter 4, the way to keep a stopwatch method from being biased is to keep the person with the stopwatch blind to the experimental condition. Since the participant is always reading color-names aloud in both conditions, if the person timing cannot see if the words are congruent with names, then the timing cannot be biased by condition.

THE BERNOULLI-DOPPLER-LEIDENFROST-PELTZMAN-SAPIR-WHORF-DUNNING-KRUGER-STROOP EFFECT STATES THAT IF A SPEEDING FIRE TRUCK LIFTS OFF AND HURLES TOWARD YOU ON A LAYER OF SUPERHEATED GAS, YOU'LL DIVE OUT OF THE WAY FASTER IF THE DRIVER SCREAMS "RED," IN A NON-TONAL LANGUAGE THAT HAS A WORD FOR "FIREFIGHTER" THAN IF THEY SCREAM "GREEN." IN A TONAL LANGUAGE WITH NO WORD FOR "FIREFIGHTER" WHICH YOU THINK YOU'RE FLUENT IN BUT AREN'T.



Learning Objectives

- Understand the strengths and weakness of **within-participants** design
- Identify different kinds of **order**, **carryover**, or **history** effects on measures of the dependent variable
- Identify participant variables that can be effectively controlled with this method
- Apply the dependent samples t-test as the correct statistical approach to data analysis for within-participant designs
- Understand the increased efficient and statistical power of within-participants design for testing experimental hypotheses

Within-Participants Experiments

In a within-participants experiment, each participant is tested under all conditions. Consider an experiment on the effect of a defendant's physical attractiveness on judgments of his guilt. Again, in a between-participants experiment, one group of participants would be shown an attractive defendant and asked to judge his guilt, and another group of participants would be shown an unattractive defendant and asked to judge his guilt. In a within-participants experiment, however, the same group of participants would judge the guilt of both an attractive and an unattractive defendant.

The primary advantage of this approach is that it provides maximum control of extraneous participant variables. Participants in all conditions have the same mean IQ, same socioeconomic status, same number of siblings, and so on—because they are the very same people. Within-participants experiments also make it possible to use statistical procedures that remove the effect of these extraneous participant variables on the dependent variable and therefore make the data less "noisy" and the effect of the independent variable easier to detect. However, not all experiments can use a within-participants design nor would it be desirable to do so.

Carryover Effects and Counterbalancing

The primary disadvantage of within-participants designs is that they can result in **order effects**. An order effect occurs when participants' responses in the various conditions are affected by the order of conditions to which they were exposed. One type of order effect is a **carryover effect**. A carryover effect is an effect of being tested in one condition on participants' behavior in later conditions. One type of carryover effect is a **practice effect**, where participants perform a task better in later conditions because they have had a chance to practice it. Another type is a **fatigue effect**, where participants perform a task worse in later conditions because they become tired or bored. Being tested in one condition can also change how participants perceive stimuli or interpret their task in later conditions. This type of effect is called a **context effect** (or contrast effect). For example, an average-looking defendant might be judged more harshly when participants have just judged an attractive defendant than when they have just judged an unattractive defendant. Within-participants experiments also make it easier for participants to guess the hypothesis. For example, a participant who is asked to judge the guilt of an attractive defendant and then is asked to judge the guilt of an unattractive defendant is likely to guess that the hypothesis is that defendant attractiveness affects judgments of guilt. This knowledge could lead the participant to judge the unattractive defendant more harshly because he thinks this is what he is expected to do. Or it could make participants judge the two defendants similarly in an effort to be "fair."

Carryover effects can be interesting in their own right. (Does the attractiveness of one person depend on the attractiveness of other people that we have seen recently?) But when they are not the focus of the research, carryover effects can be problematic. Imagine, for example, that participants judge the guilt of an attractive defendant and then judge the guilt of an unattractive defendant. If they judge the unattractive defendant more harshly, this might be because of his unattractiveness. But it could be instead that they judge him more harshly because they are becoming bored or tired. In other words, the order of the conditions is a confounding variable. The

attractive condition is always the first condition and the unattractive condition the second. Thus any difference between the conditions in terms of the dependent variable could be caused by the order of the conditions and not the independent variable itself.

There is a solution to the problem of order effects, however, that can be used in many situations. It is **counterbalancing**, which means testing different participants in different orders. The best method of counterbalancing is complete counterbalancing in which an equal number of participants complete each possible order of conditions. With two conditions, this is simple. For example, half of the participants would be tested in the attractive defendant condition followed by the unattractive defendant condition, and others half would be tested in the unattractive condition followed by the attractive condition. With three conditions, the number of orders starts to get larger and there would be six different orders (ABC, ACB, BAC, BCA, CAB, and CBA), so some participants would be tested in each of the six orders. With four conditions, there would be 24 different orders; with five conditions there would be 120 possible orders. With counterbalancing, participants are assigned to orders randomly, just as in between-participants designs. Here, instead of randomly assigning to conditions, they are randomly assigned to different orders of conditions. In fact, it can safely be said that if a study does not involve random assignment in one form or another, it is not an experiment.

More commonly, when the number of potential orders is large, experiments simply use random counterbalancing in which the order of the conditions is randomly determined for each participant. Using this technique every possible order of conditions is determined and then one of these orders is randomly selected for each participant. Use of random counterbalancing will result in more random error, but if order effects are likely to be small and the number of conditions is large, this is an option available to researchers.

There are two ways to think about what counterbalancing accomplishes. One is that it controls the order of conditions so that it is no longer a confounding variable. Instead of the attractive condition always being first and the

unattractive condition always being second, the attractive condition comes first for some participants and second for others. Likewise, the unattractive condition comes first for some participants and second for others. Thus, any overall difference in the dependent variable between the two conditions cannot have been caused by the order of conditions. A second way to think about what counterbalancing accomplishes is that if there are carryover effects, it makes it possible to detect them. One can analyze the data separately for each order to see whether it had an effect.

Simultaneous Within-Participants Designs

So far, we have discussed an approach to within-participants designs in which participants are tested in one condition at a time. There is another approach, however, that is often used when data is collected across multiple trials (events). For example, if participants were asked to judge the guilt of 10 attractive defendants and 10 unattractive defendants, these could be presented in an intermixed order instead of having people make judgments about all 10 defendants of one type followed by all 10 defendants of the other type. The researcher could then compute each participant's mean rating for each type of defendant.

Similarly in the Stroop example earlier in the chapter, if we can measure reaction time of reading aloud for each color-name, we can have the stimuli intermixed between congruent and incongruent. With this kind of mixed stimuli design, order effects are minimized. This approach removes some of the concerns of order and carryover effects across conditions by having them presented roughly at the same time.

Some attention still has to be to the specific order of stimuli in these kinds of approaches. If the stimulus order is constructed using a purely random mechanism, the resulting order can end up not looking very *random* from a human perspective (humans are poor at recognizing truly random sequences) because there can be long subsequences of the same condition across several trials. Typically, the stimulus order in this kind of design is constructed

using a *pseudo-random* sequence that limits the number consecutive stimuli in a row of the same condition. Note that if this is not done, the apparent randomness of the sequence ends up being an extraneous variable that will add noise to the dependent variable, weakening the reliability of the results, but not confounding or weakening the validity of the study.

Whenever this approach can be used, it generally should be as it removes many of the concerns about order effects across experimental conditions. Unfortunately, it does not immediately address concerns about demand characteristics inherent to within-participants designs.

Demand Characteristics on Participants

In Chapter 4, we discussed the general approach of keeping participants unaware of the underlying hypothesis for the experiment to minimize demand characteristics. In a between-participants design, this happens naturally as participants only see one condition and will usually not be able to infer what the independent variable is or how it might affect the dependent variable. In a within-participants design, participants will necessarily see both conditions and may be more aware of the hypothesis.

If there is a concern that the expectations of participants will affect their performance, it may not be possible to use a within-participants design. When participants will have seen both conditions, both levels of the independent variable, it is necessary to consider whether they might come to expect the hypothesis to be true. That expectation might shift their behavior in other ways, for example to try harder in one condition, that would create a significant problem with the internal validity of the study.

In simultaneous or mixed order designs, we can sometimes hide the levels of the independent variable by including additional trials in some additional condition(s) that are not intended for analysis. Such **filler items** can keep participants less aware of the key planned contrasts and reduce expectation effects.

In other cases, such as studies of perception, the processing being measured is sufficiently fast or automatic to be relatively resistant to expectation effects. In some studies, the experimental hypothesis may be subtle or surprising enough that even with full knowledge of the conditions, the participants are unlikely to understand the hypothesis enough to influence their behavior.

Within-participants Statistics

In a within-participants experiment, we will tend to organize the data somewhat differently. In a data spreadsheet, we will continue to generally organize the data with one participant per row in our data matrix. However, now we will have multiple columns of data for scores of the dependent variable across the conditions tested. This allows us to visually inspect the data and often makes condition differences quite easy to see. If in each (or many) of the participants one of the conditions is consistently producing a larger score on the dependent variable, we are likely to be observing a reliable difference. In fact, it is often useful to calculate a subtraction score between the conditions for each participant.

The primary difference in carrying out the basic statistical analysis of a two condition study is to run a **dependent samples t-test** instead of the independent samples t-test done in Chapter 5. Running a dependent (paired) samples t-test using R software requires one extra step for reorganizing the data. The input to analysis in R requires only one data point per row in the data matrix, so the data must be reorganized and recoded so that the two conditions are listed as the same participants on two rows. In addition, a participant number is now required in order to connect the two measures. The specific process for carrying out the data reorganization and the analysis will be reviewed in Chapter 11 as the process is the same as what is required in more complex factorial designs.

An alternative approach to running a paired sample t-test is to do a one-sample t-test on the subtraction scores as these are formally identical

procedures with respect to the math. An advantage of this alternate approach is that the t-statistic is the mean of the subtraction score across participants divided by the standard error of this value. This is simple enough to calculate directly within spreadsheet software if you prefer. If you use this method, you will have to remember that the degrees of freedom in this analysis are the number of participants minus one.

To this point, we have only considered simple designs with a single independent variable that we will either between or within participants. We will shortly extend the general model of experimental design to more complex factorial designs with multiple independent variables (factors). An issue to be aware of with the types of designs discussed here that are intended to be simple within-participants studies is that they can accidentally become more complex designs due to order effects. As noted above, when a design has to be administered in order, e.g., condition A then condition B, we should always counterbalance the order of conditions to ensure that order is not confounded with condition. In this case, we should also always check to see if the first measurement and the second measurement differ from each other regardless of the A/B conditions. This could happen due to a **fatigue effect** (second score is always lower) or a **practice effect** (second score is always higher). If this occurs, the design should no longer be analyzed as a simple paired-sample t-test but requires using a factorial analysis approach that simultaneously considers both condition and order effects (technically this is called a mixed-model factorial ANOVA). Needing to do this does not make the analysis intractable in general, but the risk of discovering the experimental design requires more complex analysis is one to be aware of when planning a within-participants design.

Statistical Power and Planning the Sample Size

Within-participants designs have a big advantage in **statistical power**, effectively meaning that a study is more likely to work with the same number of participants. When planning a study, an important question is to have a

specific sample size to aim for in recruiting. There is a formal process of mathematically estimating the ideal sample size termed a **power analysis**. Calculation of a power analysis depends on estimating the **effect size**, the magnitude of the difference expected on the dependent variable. This is then considered together with an estimation of the variability expected on measures on the dependent variable across participants.

Conceptually, a power analysis is an attempt to try to avoid a Type 2 error. If our experimental hypothesis is correct, we might still not observe a reliable statistical effect if the variability in performance is too large compared to the difference in the performance across conditions or groups. Our work in experimental control is aimed at reducing unrelated sources of variability so that we are more likely to be able to observe effects when they are real.

Within-participant designs are much more statistically powerful than between participants designs for the same number of participants. Part of this large advantage comes from the use of the same participants across conditions that greatly reduces variability arising from participant variables. Another part comes from the fact that each participant is essentially providing data twice, once in each condition. In effect, the same number of participants is producing twice as much data.

Between-Participants or Within-Participants?

For many psychological studies, planning the study procedure involves facing the question of whether the design will be between or within participants. Simply because of the increased power and efficiency of within participant designs, they should generally be preferred if they can be applied to the research question driving the study.

The planned procedure for the two conditions to be contrasted should be evaluated as to whether there is likely to be a carryover or history effect if the conditions are administered in succession. If there is only mild concern about these effects, counterbalancing the order may be sufficient and the more

powerful within design might be preferred. If the history effects are expected to be substantial, then a between participants design will likely be necessary to maintain scientific rigor. Similar attention has to be paid to the demand characteristics of the experiment and the expectations of participants. If even mild deception is needed to keep participants blind to the underlying hypothesis, then it is very likely that a between participants design should be used.

A final consideration in choosing which type of design is to consider the length of the experimental session that a participant is expected to complete during participation. Between-participants experiments have the advantage of being conceptually simpler and requiring less testing time per participant since only one condition is being completed. Shorter tests reduce the risk of fatigue and also interact with other details of the procedure for data collection. For in-person data collection where participants come to a laboratory to complete an experiment under controlled conditions, a significant component of time and effort is scheduling and traveling to the lab. Shortening a protocol from 50 minutes to 25 minutes is of limited value in this case. However, for online data collection, shorter protocols may have better engagement with participants and reduce fatigue and the probability that the participants drop out of the study before completion. Practical questions of recruiting may be important for planning as it may be easier to find volunteers (even paid) for shorter research protocols than ones that run for several hours.

Remember also that using one type of design does not preclude using the other type in a different study. There is no reason that a researcher could not use both a between-participants design and a within-participants design to answer the same research question. In fact, professional researchers often take exactly this type of mixed methods approach.

Pre/post designs

A very specific kind of within participants design is one in which the dependent variable is assessed before and after an experimental intervention. These can be described as pre/post designs because an assessment is run pre-intervention and post-intervention. Intervention research is very common in applied areas within psychology as well as public health, economics, and public policy. We will briefly review some of the common methodologies employed in invention-based research in Chapter 14.

If we think of the two assessments as levels of an independent variable, before and after the intervention, these designs look like the kinds of two condition designs discussed here. However, this simple approach is difficult to make very rigorous because of participant expectations associated with the intervention. If the participants expect to do better after the intervention, that might create a significant validity problem due to demand characteristics. Best practice is to include a control intervention, which now potentially makes this design more complex than a simple two-condition comparison. The natural method to approach data analysis with this design is to consider this a factorial design with one between participants variable and one within participants variable. We will discuss design, analysis and interpretation of these designs in Chapters 10-12.

However, there is a simplifying approach that can in some cases reduce the complexity these designs. In a design with assessments done pre/post, the two scores can be combined as a subtraction score, typically post minus pre. The subtraction score can act as the dependent variable itself and the comparison across two different interventions then follows the guidelines for a simple two group between participants design. We note this approach here as the pre/post subtraction score can look like a within participants design but in practice effectively be a simple between participants design.

Key Takeaways

- Advantage of within participant design: They are highly efficient. Each participant provides data in all conditions so accumulating data collection towards the planned number of participants is more rapid.
- Within participant designs provide perfect control of participant variables. Since all participants provided data on all conditions, the conditions are exactly matched for all extraneous variables related to the participants on the task, e.g., motivation, attention, ability.
- An important disadvantage of within participant designs is the risk of order effects, although known as carryover or history effects. If conditions have to be given in order, there are many ways in which history effects influence the data such that the earlier/later conditions performance on the DV is affected. These can be counterbalanced but never fully controlled.
- Within participant designs may also be affected by participant demand characteristics across the two conditions. By being exposed to all conditions of the experiment, participants will always be aware of all the levels of the independent variable. This increases the chance that they will understand the experimental hypothesis, which may affect their performance, through bias related to expectations.

Exercises

Question 1. Why is it generally impractical to use a within-participants design in studies that have an element of deception (e.g., the implicit bias studies)?

Question 2. For a study assessing time to recognize famous faces upside-down, why would a within-participants design be a good idea? Give two reasons.

Question 3. Why are mood manipulation studies difficult to do as a within-participants design?

Question 4. Why are learning-based studies difficult to do as a within-participants design?

8 Research Ethics

It is somewhat traditional in presenting an introduction to Research Ethics to start with a litany of famous historical ethical lapses to motivate the explanation for the current standard regulatory framework within which science with human participants works. Here we will take a different approach reflecting the hands-on style of learning psychological science. First we will consider the perspective of the research participant given what we have previously discussed about rigorous experimental control. Then we will review how scientists work with and within the regulatory framework that provides oversight and ensures ethical practice.

If you have had the opportunity to participate in a psychological research study, for class credit or pay, you are likely familiar with the sense of trust you need to have towards the experimenter. As we saw in earlier chapters about experimental control, the participant is placing themselves into a situation that has been carefully and meticulously controlled by the experimenter to minimize the impact of extraneous variables. One or a very few small elements of the experience are being explicitly manipulated in order to assess effects on a dependent variable that is being quantitatively assessed from performance. Because of concerns about demand characteristics and expectation effects, the participant may not be fully aware of every aspect of behavior that is being manipulated, controlled or measured. The participant

has to accept being placed in this context and maintain trust that the experimenter will not allow them to be treated unfairly or harmed.

A major element of establishing that trust is that participants should initially be presented with an **Informed Consent** form that will tell them about the research study in which they are asked to participate. That form needs to be read, understood and signed by the participant before the procedure can begin or any data collected. A crucial element of this process is that the participants know that their **participation is entirely voluntary** and no negative consequences of failing to participate can be in place to coerce them into research participation. Because in some cases, some aspects of the experimental hypothesis cannot be made completely clear in advance, this document may not explain every detail fully. Throughout this document will be indications, often marks or stamps, that the exact content of this form has been thoroughly examined by the **Institutional Review Board** and found to appropriately meet the balance of demands between scientific rigor and fair and ethical treatment of the participants.

The involvement of the Institutional Review Board, frequently referred to by its acronym **IRB**, should provide assurance to the participant that the study is operating within the regulatory framework that guarantees compliance with standard ethical practice. This chapter will focus on the elements of ethics related to the **fair and equitable treatment of human participants in research**. Additional topics related to the ethical practice of reporting and disseminating science will be covered when we return to this topic in Chapter 18, Responsible Conduct of Research.

As an experimenter, working to carry out a new research study, you are expected to work within this framework and with institutional oversight over your research process. The experimenter is responsible for creating the informed consent form and sending it together with a detailed summary of the research protocol to the IRB for review and approval. The submitted research protocol has to fully document all experimental procedures and the stimuli to be used in the study. The main review goal of the IRB is to carefully evaluate all the **risks to the participant** that might be incurred

through the procedure. These are then weighed against the **benefits** of carrying out the scientific study. If the benefits outweigh the risks, then the study is judged to be ethically compliant and can be carried out.

As we will see, a large proportion of studies in experimental psychology are easily seen to be **minimal risk**, which is a technical designation reflecting no likely harm could occur to the participants. Nothing bad can happen as a consequence of participating in a study like Experiment 1. However, it is important to note that the decision that a research study is minimal risk is not up to the experimenter leading the study or their research team. The research protocol still needs to be reviewed by the IRB, who provide the formal judgment of that designation.

CITI Training

If you have the opportunity to work within a research lab in the future, you will very likely be asked to complete **CITI training** and obtain certification. The Collaborative Institutional Training Initiative (CITI) is the main source of ethics training for researchers in psychology, health and medicine. Completing the training means signing up with their service, completing training and test modules and connecting that account to your university system. The training modules you will most likely encounter for psychological research are Good Clinical Practice (GCP) and Social, Behavioral, Educational (SBE) research topics. It may also include specialized content in Biomedical Research if biomedical methods like imaging or working with patient populations is part of the research protocol. In addition, all CITI training requires knowledge of Responsible Conduct of Research (Chapter 18). The university IRB then verifies your certification and allowability to be part of a research team associated with the approved research protocol.

Learning Objectives

- Describe a simple framework for thinking about ethical issues in psychological research.
- Give examples of common ethical issues that arise in psychological research that affect research participants.
- Describe the history of ethics codes for scientific research with human participants.
- Describe several strategies for identifying and minimizing risks and deception in psychological research.
- Understand informed consent and debriefing procedures, administering a consent form.
- APA ethics code for reference: <https://www.apa.org/ethics/code>
- The basics of ethical research are all founded on the simple idea of treating research participants with fairness and respect. They should be participating voluntarily in the research protocol with as much information about what is expected of them as can be made available. For most, basic scientific studies in psychological science, these guidelines will maintain compliance with the spirit of best practices in research ethics.

In this chapter we explore the ethics of scientific research in psychology.

We begin with a general framework for thinking about the ethics of scientific research in psychology. Then we look at some specific ethical codes for biomedical and behavioral researchers —focusing on the Ethics Code of the American Psychological Association. Finally, we consider some practical tips for conducting ethical research in psychology.

Ethics is the branch of philosophy that is concerned with morality—what it means to behave morally and how people can achieve that goal. It can also refer to a set of principles and practices that provide moral guidance in a particular field. There is an ethics of business, medicine, teaching, and of

course, scientific research. As the opening example illustrates, many kinds of ethical issues can arise in scientific research, especially when it involves human participants. For this reason, it is useful to begin with a general framework for thinking through these issues.

Moral Principles

Current practice guiding ethical approaches to research are based on a short set of core principles. Five basic guidelines are listed in the APA guidance on ethics:

- 1. Beneficence**
- 2. Fidelity and Responsibility**
- 3. Integrity**
- 4. Justice**
- 5. Respect for People's Rights and Dignity**

These reflect an extension of a core set of three principles from The Belmont Report, a major milestone in the development of a common standard for ethical research (see below). The Belmont Report emphasized three core principles: Beneficence, Justice, and Respect for Persons. The goal of starting with simple, intuitive principles is to emphasize that ethics arises from basic social expectations and customs. Most basic ethical questions are fairly straightforward and standard procedures effectively minimize risk of any costs while obtaining the benefits of new scientific knowledge. Later we will touch on more complex questions in research areas where substantial costs must be considered in the context of potentially important findings.

Respect for people is a core and standard idea. Your participants are participating in your research study, which is of benefit to you. Even if participants are compensated for participation, they have to trust you not to impose extreme or unfair conditions on them that are disproportionate to their expectations. As we will see below when reviewing the historical examples that led to the characterization of the current ethical regulatory framework, most of the egregious violations of ethics arose from a failure of

basic respect for human rights and dignity.

Beneficence reflects the fact that science should be done with the intention of increasing societal benefit. An implication of this idea is that our research studies should be designed to be internally valid and with the intention to publicize findings to the broader scientific community. Poorly designed research is actually unethical in that you are imposing some costs on your participants (even when these are minimal) but if no inferences can be drawn from the results because of design error, the research fails to meet the standard of beneficence.

The idea of justice in ethical research reflects the idea that science should benefit all humanity. The need to include this as a core element of our moral principles for ethical research is sadly due to the history of human experience

Be nice

A lot of the moral framework for research boils down to simply *be nice to your participants*, they are doing you the favor of being in your study. Just treat your participants as humans deserving of respect and dignity.

Of course, if everybody found it really easy to be nice to people all the time, the world would look a lot different than it does. In science, we have our regulatory oversight framework to help.

being dominated by assumptions based on an in-group of people being more important or valuable than out-groups, an “us” versus “them” mentality. Segmenting humanity into groups in which some are more valuable is the core principle of racism (and sexism, classism, and other forms of minority oppression), which unfortunately substantially influenced some historical scientific research. Modern ideas of treating all humans equally have not always been reflected in research, which showed evidence of the cultural expectations of that time, e.g., when racism was more prevalent and/or accepted. To improve on this, we always consider the question of justice and the equity issues implied in research – does the

proposed work aim to benefit all humanity?

The ideas of fidelity and integrity in the APA code reflect the dangers associated with the impact of unethical scientific claims on public policy, understanding and trust. We will return to discuss these ideas in more depth in Chapter 18 in the context of Responsible Conduct of Research as a component of ethical research.

Weighing Risks Against Benefits

We start our understanding of ethical research with these principles and then we need to understand that the risks (costs) of research involve transgressing against these. As noted above, the frequent need for at least some deception is a minor violation of the concept of respect for persons. The goal of the starting principles is to guide the analysis of all of the negative aspects of a research study, i.e., the risks or costs, and then to consider these in balance with the scientific gains that can be obtained by carrying out the research.

Scientific research in psychology can be ethical only if its risks are outweighed by its benefits. Among the risks to research participants are that a treatment might fail to help or even be harmful, a procedure might result in physical or psychological harm, and their right to privacy might be violated. Risk of physical harm in psychology studies is fairly rare as many experimental procedures and manipulations do control physical aspects of behavior. As you might expect, risks of psychological distress need to be assessed and considered. Emotional or psychological distress may occur when participants are exposed to stimuli or tasks that evoke anxiety, fear, or sadness. This is particularly relevant in studies involving sensitive topics or emotional manipulations.

Among the potential benefits are receiving a helpful treatment, learning about psychology, experiencing the satisfaction of contributing to scientific knowledge, and receiving money or course credit for participating. Scientific research can have risks and benefits to the scientific community and to

society too. A risk to science is that if a research question is uninteresting or a study is poorly designed, then the time, money, and effort spent on that research could have been spent on more productive research. A risk to society is that research results could be misunderstood or misapplied with harmful consequences. Of course, the benefits of scientific research to science and society are that it advances scientific knowledge and can contribute to the welfare of society.

The regulatory framework within which the risks and benefits are weighed in order to evaluate the ethics of a psychological science study is the same framework used for all research with human participants, including health and medical research. This may be surprising given that clinical research with patients would seem to have a very different overall profile for both the risk to participants and the reward. The underlying principle of benefits outweighing risks still applies and the evaluation process for a new, proposed research study makes no assumptions about the likely risk level based on the broad domain of the research area. The primary consequence of treating all science the same initially is that researchers considering research hypotheses that are minimal risk must still document how their work does not impose any hidden or unexpected risks to participants.

Historical Overview

The rigor and care with which the scientific ethical framework is applied is motivated by historical examples where scientific work was done that clearly transgressed against these ideas. Identification of events associated with unethical research led to three major efforts in describing sets of guiding principles that would prevent recurrence of these kinds of transgressions.

One of the earliest ethics codes was the **Nuremberg Code**—a set of 10 principles written in 1947 in conjunction with the trials of Nazi physicians accused of shockingly cruel research on concentration camp prisoners during World War II. It provided a standard against which to compare the behavior of the men on trial—many of whom were eventually convicted and either

imprisoned or sentenced to death. The Nuremberg Code was particularly clear about the importance of carefully weighing risks against benefits and the need for informed consent. The **Declaration of Helsinki** is a similar ethics code that was created by the World Medical Council in 1964. Among the standards that it added to the Nuremberg Code was that research with human participants should be based on a written protocol—a detailed description of the research—that is reviewed by an independent committee. The Declaration of Helsinki has been revised several times, most recently in 2004.

In the United States, concerns about the Tuskegee study and others led to the publication in 1978 of a set of federal guidelines called the Belmont Report. **The Belmont Report** explicitly recognized the principle of seeking justice, including the importance of conducting research in a way that distributes risks and benefits fairly across different groups at the societal level. It also recognized the importance of respect for persons, which acknowledges individuals' autonomy and protection for those with diminished autonomy (e.g., prisoners, children), and translates to the need for informed consent. Finally, it recognized the principle of beneficence, which underscores the importance of maximizing the benefits of research while minimizing harms to participants and society.

The Belmont Report became the basis of a set of laws—the **Federal Policy for the Protection of Human Subjects**—that apply to research conducted, supported, or regulated by the federal government. An extremely important part of these regulations is that universities, hospitals, and other institutions that receive support from the federal government must establish an institutional review board (IRB)—a committee that is responsible for reviewing research protocols for potential ethical problems. An IRB must consist of at least five people with varying backgrounds, including members of different professions, scientists and nonscientists, men and women, and at least one person not otherwise affiliated with the institution. The IRB helps to make sure that the risks of the proposed research are minimized, the benefits outweigh the risks, the research is carried out in a fair manner, and the informed consent procedure is adequate.

<https://www.hhs.gov/ohrp/international/ethical-codes-and-research-standards/index.html>

For psychology, additional guidance is provided by the APA's Ethical Principles of Psychologists and Code of Conduct (also known as the APA Ethics Code) was first published in 1953 and has been revised several times since then, most recently in 2016. It includes about 150 specific ethical standards that psychologists and their students are expected to follow. Much of the APA Ethics Code concerns the clinical practice of psychology—advertising one's services, setting and collecting fees, having personal relationships with clients, and so on. For our purposes, the most relevant part is Standard 8: Research and Publication and this serves as a good overall reference description of the key issues. <https://www.apa.org/ethics/code>

Even within these guidelines, it is not necessarily easy to weigh the risks of research against its benefits because the risks and benefits may not be directly comparable. For example, it is common for the risks of a study to be primarily to the research participants but the benefits primarily for science or society. Consider, for example, Stanley Milgram's original study on obedience to authority (Milgram, 1963). The participants were told that they were taking part in a study on the effects of punishment on learning and were instructed to give electric shocks to another participant each time that participant responded incorrectly on a learning task. With each incorrect response, the shock became stronger—eventually causing the other participant (who was in the next room) to protest, complain about his heart, scream in pain, and finally fall silent and stop responding. If the first participant hesitated or expressed concern, the researcher said that he must continue. In reality, the other participant was a confederate of the researcher—a helper who pretended to be a real participant—and the protests, complaints, and screams that the real participant heard were an audio recording that was activated when he flipped the switch to administer the "shocks." The surprising result of this study was that most of the real participants continued to administer the shocks right through the confederate's protests, complaints, and screams. Although this is considered one of the most important results in psychology—

with implications for understanding events like the Holocaust or the mistreatment of prisoners by US soldiers at Abu Ghraib—it came at the cost of producing severe psychological stress in the research participants.

This research was done at a time when the long-term implications of post-traumatic stress disorder (PTSD) were not as well understood as they are now. Although this study was not governed by systematic oversight as modern research is, the idea that the basis of the research could be explained to the participants after the study with an expectation that there would be no long-term effects might have been seen as justifying the scientific benefit. However, the modern understanding of the potential risk of unintentionally creating a long-lasting psychiatric challenge in research participants would clearly indicate that this project cannot be carried out ethically. This highlights one of the very difficult aspects of scientific ethics in that the risk/reward balance depends on current scientific understanding and this can change as science progresses.

The guidelines also reflected an understanding that researchers must conduct their research in a just manner. At a broader societal level, members of some groups have historically faced more than their fair share of the risks of scientific research, including people who are institutionalized, are disabled, or belong to racial or ethnic minorities. A particularly tragic example is the Tuskegee syphilis study conducted by the US Public Health Service from 1932 to 1972 (Reverby, 2009). The participants in this study were poor African American men in the vicinity of Tuskegee, Alabama, who were told that they were being treated for “bad blood.” Although they were given some free medical care, they were not treated for their syphilis. Instead, they were observed to see how the disease developed in untreated patients. Even after the use of penicillin became the standard treatment for syphilis in the 1940s, these men continued to be denied treatment without being given an opportunity to leave the study. The study was eventually discontinued only after details were made known to the general public by journalists and activists. Subsequent investigation identified a number of additional studies in which medical treatments or interventions were not applied in a fair and

equitable manner to all participants. Most of these cases also reflected scientific mistreatment of marginalized communities. It is now widely recognized that researchers need to explicitly consider issues of justice and fairness at the societal level to avoid any possible creation of systematic mistreatment.

Who provides regulatory oversight?

All research universities and major research institutions all maintain an **Office for the Protection of Human Research Participants**. The **Institutional Review Board** is generally housed within this department. Smaller universities or hospitals engaging in research will sometimes contract with external consulting firms to provide regulatory oversight if there is insufficient local need for a whole department.

The office for research oversight consists of specially trained research staff who help manage protocols, submissions for review and any issues that arise related to compliance with ethical practice or adverse events during research procedures. The staff also support researchers who lead research protocols, but they do not participate directly as IRB panel members (large universities will also tend to have several panels with differing specializations). IRB members are mainly faculty from other departments or research domains who can evaluate proposed research protocols to verify compliance with best practices in ethical research. In addition, most IRB's have participation from community members outside the university to provide the perspective of a layperson who has not directly led research projects. These IRB members are the ones charged with applying current research rules. The department staff participate in a broader national (or international) conversation about the current understanding of best practices and how these change with new laws passed, recent court cases or the introduction of novel technologies in research.

Ethical Research in Practice

In a formal research process, all researchers involved in data collection or analysis must meet institutional training requirements for procedures and policies related to human participant research. A very common training tool is provided by CITI, which is an external online training platform that can certify training completion with the researcher's host institution. Everybody in the research process including faculty, laboratory staff, graduate and undergraduate student researchers all must meet the institutional training requirements.

Once training is certified, the researchers prepare a description of the research protocol, recruiting process and informed consent procedures to the IRB for review. The initial assessment of the IRB staff is to determine (a) does the research protocol qualify as *human participants research* and (b) what level of risk is associated with the proposal. All the experimental research described so far in this text qualifies as human participants research (protocols that do not qualify will be discussed briefly in Chapter 19) as they reflect systematic data collection from humans with an intent to disseminate the findings.

The levels of possible risk inherent in the research protocol is designated as one of three levels. **Exempt research** is the lowest level of risk and includes research on the effectiveness of normal educational activities, the use of standard psychological measures and surveys of a nonsensitive nature that are administered in a way that maintains confidentiality, and research using existing data from public sources. It is called exempt because once approved, it is exempt from regular, continuous review. A common heuristic for identifying exempt research is to ask: Would it be reasonably expected to be asked of students in a class? If so, the procedure typically reflects so little risk that the research does not require additional ongoing review from the IRB. It should be noted that only the IRB can designate a protocol as exempt. This decision cannot be made by the lead researcher even if it seems obvious the protocol should be exempt.

Expedited research poses a somewhat higher risk than exempt, but still exposes participants to risks that are no greater than minimal risk (those encountered by healthy people in daily life or during routine physical or psychological examinations). Expedited review is done by one member of the IRB or by a separate committee under the authority of the IRB that can only approve minimal risk research (many departments of psychology have such separate committees). Much psychological research is either exempt or expedited in the risk level. Research that involves deception or some significant privacy risk (e.g., asking sensitive personal questions) is typically expedited and most other protocols are designated exempt.

Finally, research that does not qualify for exempt or expedited review is greater than minimal risk research must be reviewed by the full board of IRB members. This level is common in a wide range of medical research procedures that may embed substantial risk to the participants that needs to be thoroughly evaluated and carefully monitored after approval.

The IRB review designates the risk level and review process for the research protocol and may result in suggestions or requests from the IRB panel to adjust research procedures to improve the risk/reward balance. The review process then considers the process and documentation for obtaining informed consent from research participants.

Before starting a new research project, a detailed description of the research protocol must be made available to the IRB for review in order to evaluate the risks and benefits of the research. In addition, the proposed form to collect informed consent for participants must be provided and all materials used to recruit participants into the study. When possible, all the stimuli, surveys or other research instruments should be provided or described in detail. All of this information is needed to thoroughly evaluate the risks that might be inherent in the research protocol. Once identified, these can be compared to the proposed benefits of carrying out the scientific research to test the experimental hypothesis.

Informed Consent

Informed consent means obtaining and documenting people's agreement to participate in a study, having informed them of everything that might reasonably be expected to affect their decision. This includes details of the procedure, the risks and benefits of the research, the fact that they have the right to decline to participate or to withdraw from the study, the consequences of doing so, and any legal limits to confidentiality.

A critical element of the process of obtaining informed consent is to establish that the participant is aware that **participation in research is completely voluntary**. Nobody should ever be pressured or coerced into accepting the risks of a research protocol. Even the possibility of implicit pressure is considered a potential violation of this principle, making it generally improper for employees of a laboratory to participate in even unrelated research done by that lab for fear they are doing it under the threat of consequences to their employment. Participants should also be aware that they are also free to decline to continue participation at any time during the research protocol if they feel uncomfortable for any reason. Consequences for partial participation should be specified on the consent form together with explicit reassurance that there is no penalty for withdrawing at any time during the protocol.

Although the process of obtaining informed consent often involves having participants read and sign a consent form, it is important to understand that this is not all it is. Although having participants read and sign a consent form might be enough when they are competent adults with the necessary ability and motivation, many participants do not actually read consent forms or read them but do not understand them. For example, participants often mistake consent forms for legal documents and mistakenly believe that by signing them they give up their right to sue the researcher. Even with competent adults, therefore, it is good practice to tell participants about the risks and benefits, demonstrate the procedure, ask them if they have questions, and remind them of their right to withdraw at any time—in addition to having

them read and sign a consent form.

Consent forms have a specific list of required sections that address typical questions about what is asked of participants, what compensation they will receive for participation and any risks they might bear from being in the study. Most IRB's will provide a consent form template that should be followed in preparing a consent form for a research study that will include detailed instructions on form and content. It is generally necessary to follow a provided template to ensure that the proposal's consent form meets standards and criteria for the institution's guidelines.

Recruiting participants

Researchers must also provide information on methods of recruiting participants into the study. These are typically flyers, ads or emails sent

Recruiting Materials

Information provided through ads or other recruiting methods are often the first way that participants learn about the opportunity to participate in a research study. The content and method of recruiting influences whether the sample of participants will fairly represent the population. In addition, the advertising will be aimed at generating enthusiasm for participating, but must also fairly reflect the possible risks to participation.

VOLUNTEERS NEEDED

FOR A SCIENTIFIC STUDY

INVESTIGATING WHETHER
PEOPLE CAN DISTINGUISH
BETWEEN SCIENTIFIC
STUDIES AND KIDNEY-
HARVESTING SCAMS.

(HEALTHY TYPE-O ADULTS ONLY)

TAKE ONE



to potentially interested participants. The review of recruiting materials is primarily aimed to ensure that participants are all volunteers who chose to engage in the research and there is no question of coercion to participate. Examples of unethical coercion include threats of loss of employment, access to medical care or other benefit to participants who choose not to participate. Nobody should ever feel forced or threatened into research participation and detailed procedures are used to ensure that any hint of coercion in recruiting are avoided.

Vulnerable populations

Special guidelines govern research on designated **vulnerable populations** which generally reflect participants who may not easily be able to establish voluntary participation in research. For example, children cannot legally attest to their participation and require a parent or guardian to sign the consent form for them. The same issue occurs in neuropsychological studies with cognitively impaired older adults, requiring a guardian signature or co-signature on the consent form. Research with vulnerable populations very commonly full board review to assess all procedural details and ensure absolute compliance with best practices.

The same idea also applies to research on incarcerated prisoners but for different reasons. Because prisoners have lost their freedom to a controlled environment, it is not expected that they can confidently assert a decision to participate or decline to participate in research. For this reason, a special set of different rules governs research with prisoners with the key requirement being that the research aims to benefit this population in a direct way (not by general benefit to society). Concerns about implicit pressure making it difficult to decline participation in research are also the reason why personnel within the laboratory conducting the research are explicitly proscribed from participating, even in minimal risk studies.

Waiver of consent

There are research situations in which informed consent is either not necessary or cannot be easily obtained within the structure of the research protocol. These include situations in which the research is not expected to cause any harm and the procedure is straightforward or the study is conducted in the context of people's ordinary activities. For example, if you wanted to sit outside a public building and observe whether people hold the door open for people behind them, you would not need to obtain their informed consent. Similarly, if a college instructor wanted to compare two legitimate teaching methods across two sections of his research methods course, he would not need to obtain informed consent from his students.

Research that might qualify for a waiver of consent must explicitly request approval for this alternate process. The request will need to be supported by an explanation of the rationale for not collecting informed consent and processes to manage any negative outcomes during the research process.

Privacy

A very common risk to be aware of in psychological research, is the implied risk to participant's privacy. Research that directly assesses personal information like sexual or other private behavior clearly carries a risk to privacy but the extent of this concern is actually substantially broader. The general guideline on privacy is that all data from research participants is kept private to only research staff and publication reports only average data that obscures any individual performance. Further, the very fact that participants were engaged in a research project is also kept confidential. In medical research, the reason for this is clear as the study may be on treatment of a disease or syndrome that the participant might not wish to have publicly disclosed. In psychological research, an example of participation privacy risk are studies on topics such as implicit racial bias which might document unexpected levels of bias in undergraduate populations. As a participant,

you might prefer not to be known to have been one of the participants in the study. Although these risks are fairly rare, the standard best practice approach is to maintain full privacy and confidentiality for all participants at all times. This poses some technical challenges for data sharing across labs that we will discuss briefly in Chapter 19.

Deception

Deception of participants in psychological research can take a variety of forms: misinforming participants about the purpose of a study, using confederates, using phony equipment like Milgram's shock generator, and presenting participants with false feedback about their performance (e.g., telling them they did poorly on a test when they actually did well). Deception also includes not informing participants of the full design or true purpose of the research even if they are not actively misinformed (Sieber, Iannuzzo, & Rodriguez, 1995). For example, a study on incidental learning—learning without conscious effort—might involve having participants read through a list of words in preparation for a “memory test” later. Although participants are likely to assume that the memory test will require them to recall the words, it might instead require them to recall the contents of the room or the appearance of the research assistant.

Some researchers have argued that deception of research participants is rarely if ever ethically justified. Among their arguments are that it prevents participants from giving truly informed consent, fails to respect their dignity as human beings, has the potential to upset them, makes them distrustful and therefore less honest in their responding, and damages the reputation of researchers in the field (Baumrind, 1985).

Note, however, that the APA Ethics Code takes a more moderate approach—allowing deception when the benefits of the study outweigh the risks, participants cannot reasonably be expected to be harmed, the research question cannot be answered without the use of deception, and participants are informed about the deception as soon as possible. This approach

acknowledges that not all forms of deception are equally bad. Compare, for example, Milgram's study in which he deceived his participants in several significant ways that resulted in their experiencing severe psychological stress with an incidental learning study in which a "memory test" turns out to be slightly different from what participants were expecting. It also acknowledges that some scientifically and socially important research questions can be difficult or impossible to answer without deceiving participants. Knowing that a study concerns the extent to which they obey authority, act aggressively toward a peer, or help a stranger is likely to change the way people behave so that the results no longer generalize to the real world.

Debriefing

Debriefing is the process of informing research participants as soon as possible of the purpose of the study, revealing any deception, and correcting any other misconceptions they might have as a result of participating.

Debriefing also involves minimizing harm that might have occurred. For example, an experiment on the effects of being in a sad mood on memory might involve inducing a sad mood in participants by having them think sad thoughts, watch a sad video, and/or listen to sad music. Debriefing would be the time to return participants' moods to normal by having them think happy thoughts, watch a happy video, or listen to happy music.

The debriefing process plays an important role in studies that use deception as part of the experimental methodology. In cases with significant deception, researchers may use a secondary consent form to be completed after the research protocol. These obviously do not protect participants about their participation beforehand, but allow them to express their concern by refusing to allow their data to be included in research. Times that a participant refuses to sign a consent form at debriefing would be considered an adverse event that is required to be tracked and reported to the IRB.

Nonhuman Animal Subjects

While our course is about experimental psychology with human participants, a small amount of research within psychological departments involves the use of nonhuman animal subjects. In some cases, this work is carried out under the domain of the neighboring discipline of neuroscience, which is a scientific domain often operating within or overlapping with psychological science. There are specific and highly detailed procedures for laboratory work with nonhuman animal subjects that apply to that area of research. The core ethical approach is based on the same underlying idea of balancing the risks with the benefits of the science obtained. Obviously, these procedures are subject to the same ongoing evolution as ethical standards for human research and always aim to maintain the same awareness of current scientific understandings (e.g., of the experience of pain in animals). These guidelines are organized and administered by specialized committees trained in animal welfare and operating as an **Institutional Animal Care and Use Committee (IACUC)**. All researchers working with animals complete training on these ethical issues to ensure that these scientific studies are carried out with attention to appropriate practice.

Common Risks in Psychology Experiments

As noted earlier, the primary risks common to psychology studies are methods that have the potential to create psychological distress. That can occur by exposure to challenging materials, such as stimuli or tasks that evoke anxiety, fear, or sadness. This is particularly relevant in studies involving sensitive topics or emotional manipulation. There can even be simple frustration with difficult tasks that should be monitored as a risk.

In addition, the requirements to respect privacy and maintain confidentiality are more substantial than they may initially appear. Even in a study as low risk as our example Experiment 1, participants were randomly assigned to conditions in which one group was expected to perform significantly worse

on a memory test. If you were in that group and performed worse than your classmates, you might prefer that your personal performance not be public for others to see. Just participating in studies involving sensitive topics may risk stigmatizing participants if the nature of the research becomes known. This is particularly relevant in studies on stereotypes, mental health, addiction, or other stigmatized conditions. Studies conducted in diverse cultural contexts may unintentionally offend or harm participants if researchers are not culturally sensitive. It's crucial to consider cultural norms and values to avoid misunderstandings.

The use of deception has to be managed carefully and is usually monitored very carefully by the IRB. Mild deception by omission related to participants being blind to the underlying hypothesis are considered minimal risk. Needing to overtly mislead participants in order to observe the intended behavior will require a very strong and important scientific question to be answered to justify this imposition.

Know and Accept Your Ethical Responsibilities

As the American Psychological Association (APA) Ethics Code notes in its introduction, "Lack of awareness or misunderstanding of an ethical standard is not itself a defense to a charge of unethical conduct." This is why the very first thing that you must do as a new researcher is to know and accept your ethical responsibilities. Ultimately, you as the researcher must take responsibility for the ethics of the research you conduct.

As you design your study, you must identify and minimize risks to participants. Start by listing all the risks, including risks of physical and psychological harm and violations of confidentiality. Remember that it is easy for researchers to see risks as less serious than participants do or even to overlook them completely. For example, one student researcher wanted to test people's sensitivity to violent images by showing them gruesome photographs of crime and accident scenes. Because she was an emergency medical technician, however, she greatly underestimated how disturbing these

images were to most people. Remember too that some risks might apply only to some participants. For example, while most people would have no problem completing a survey about their fear of various crimes, those who have been a victim of one of those crimes might become upset. It is often important to seek input from a variety of people, including your research collaborators, more experienced researchers, and even from nonresearchers who might be better able to take the perspective of a participant. Once you have identified the risks, you can often reduce or eliminate many of them.

Research protocols should always be designed to take active steps to maintain confidentiality of participants. You should keep signed consent forms separately from any data that you collect and in such a way that no individual's name can be linked to their data. In addition, beyond people's sex and age, you should only collect personal information that you actually need to answer your research question. Be aware also that certain data collection procedures can lead to unintentional violations of confidentiality. When participants respond to an oral survey in a shopping mall or complete a questionnaire in a classroom setting, it is possible that their responses will be overheard or seen by others. If the responses are personal, it is better to administer the survey or questionnaire individually in private or to use other techniques to prevent the unintentional sharing of personal information.

Follow Through

Your concern with ethics should not end when your study receives institutional approval. It now becomes important to stick to the protocol you submitted or to seek additional approval for anything other than a minor change. During the research, you should monitor your participants for unanticipated reactions and seek feedback from them during debriefing. One criticism of Milgram's study is that although he did not know ahead of time that his participants would have such severe negative reactions, he certainly knew after he had tested the first several participants and should have made adjustments at that point.

Severe negative reactions to research protocols are known as adverse events and should be reported to the IRB. It may similarly be necessary to track the number of participants who elect not to participate after reading the informed consent form, or who chose to withdraw after starting to participate. These events may indicate that the research protocol bears more risk than the initial review indicated, and this may in turn require adjustments to the research protocol.

Sometimes mistakes are made in carrying out the procedures associated with the research project. These are technically termed protocol violations and might reflect mistakes or misunderstandings in obtaining informed consent or accurately informing participants about research requirements. Protocol violations should always be reported to the research oversight staff associated with the IRB or office of research protection. Researchers might hesitate to report mistakes for fear of consequences up to and including having the research lab (or even department) shut down entirely, preventing all future research from continuing. However, these events should always be reported promptly and thoroughly as the consequences of minor mistakes are rarely substantial. A typical response process to an error would be to document the error, identify and correct any harm and document improvement in procedures or training to prevent future errors. In general, the IRB strongly prefers hearing about procedural errors from researchers as early as possible. When a problem is raised to them from a participant who has concerns about a negative experience in research (or worse, legal representation), this situation is almost always a much worse problem and requires much more substantial corrective action.

Evolution of ethical practice

The idea that the standards for ethical practice change over time is sometimes viewed with surprise by researchers engaging systematically with research ethics training for the first time. On reflection, this should not be surprising. Not only do societal and cultural expectations about ethics and morality

change over time, but technology changes and with it, new issues with respect to concerns such as privacy become relevant. Widespread access to communication over the internet has opened up new possibilities for research, but also raised additional questions about privacy and standard assurances related to informed consent, e.g., how confidently can we establish the identity of participants who are solely interacted with via the internet.

A minor example of this evolution over time is the preference for the word *participant* for humans who engage in research. Older publications and writing will generally refer to participants as *subjects* but in modern parlance, subjects are used exclusively for nonhuman animals. This was done acknowledging that the idea of being “subject” to experimenter control might be considered objectionable and to focus more directly on the need for humans in research to be participating voluntarily.

A more significant example has to do with an understanding of how to handle difficult questions of participant privacy. In general, we attempt to fully respect everything about participation in a research study and to maintain privacy of all information about participants. Research might be based on asking participants about personal, high-risk or even illegal behavior. This kind of research is handled specifically with minimizing the risk to the participants in mind. However, more recently the question of how to handle accidental awareness of potential issues separate from the research protocol, specifically, concern over mistreatment of minors participating in research. Most IRB's have decided that the need to report these concerns outweighs the maintenance of privacy and much developmental psychology research uses a *mandatory reporting* guideline where any such concerns are reported to oversight agencies.

Tension between Ethics and Science

Science is intended to reflect a gain in knowledge for the benefit of humanity. As described in the APA Ethics Code: “Psychologists are committed to increasing scientific and professional knowledge of behavior and people’s

understanding of themselves and others and to the use of such knowledge to improve the condition of individuals, organizations, and society." In experimental research, we observe or measure people's behavior in controlled or manipulated conditions to carry out this gain of knowledge. However, subjecting people to controlled conditions and measuring their responses is fundamentally not really a nice thing to do to them. When participants are unaware of experimental conditions, there is at least a minimal aspect of deception in the research protocol, a "lie" of omission, and some protocols use considerably more deception. When we observe or measure participants' behavior, we are at some level invading their privacy. The manipulation of experimental conditions may very well put participants in a challenging situation not directly of their own choice. We consider these impositions both costs of doing research and also acknowledge the risk of negative consequences created by the research protocol.

The research process therefore cannot be carried out without some cost imposed on our research participants. It is therefore necessary to always consider the costs and benefits of each specific research study and ensure that the value of carrying out the research exceeds the costs imposed on the participants. The regulatory framework for this process is centered around the Institutional Review Board (IRB) that reviews and provides oversight of research processes. For research with human participants, there is almost always a process of obtaining informed consent from participants before they engage with the research protocol.

The regulatory framework in place to ensure that scientific work maintains appropriate ethical balance between risks and rewards applies to all kinds of research done at a university or research institution. The same review and evaluation process applies to all medical research, human research, animal research, community research, epidemiological, economic, and public health studies. The majority of psychological science research actually bears fairly little risk or cost to participants, meeting a technical consideration termed minimal risk. However, the review and evaluation processes follow a common procedure across all types of research, which is often a surprise to

beginning researchers in psychology. Practically it means that the oversight paperwork process involves a lot of questions that do not appear to directly apply to simple psychological experiments. The rationale for this approach is based on the historical occurrences of unethical research that highlight the potential problem that scientists in the past have made substantial mistakes in understanding and applying a proper understanding of the risk and costs being imposed on participants. As a consequence, the IRB, as an oversight committee, is required to evaluate the potential risks of all research studies from basic principles and not simply trust that the lead researcher on a project will carry out the research ethically.

Key Takeaways

- A wide variety of ethical issues arise in psychological research. Thinking them through requires considering how each of four moral principles (weighing risks against benefits, acting responsibly and with integrity, seeking justice, and respecting people's rights and dignity) applies to each of three groups of people (research participants, science, and society).
- Ethical conflict in psychological research is unavoidable. Researchers must think through the ethical issues raised by their research, minimize the risks, weigh the risks against the benefits, be able to explain their ethical decisions, seek feedback about these decisions from others, and ultimately take responsibility for them.
- There are several written ethics codes for research with human participants that provide specific guidance on the ethical issues that arise most frequently. These codes include the Nuremberg Code, the Declaration of Helsinki, the Belmont Report, and the Federal Policy for the Protection of Human Subjects.
- The APA Ethics Code is the most important ethics code for researchers in psychology. It includes many standards that are relevant mainly to clinical practice, but Standard 8 concerns informed consent, deception, debriefing, the use of nonhuman animal subjects, and scholarly integrity in research.
- Research conducted at universities, hospitals, and other institutions that receive support from the federal government must be reviewed by an institutional review board (IRB)—a committee at the institution that reviews research protocols to make sure they conform to ethical standards.
- Informed consent is the process of obtaining and documenting people's agreement to participate in a study, having informed them of everything that might reasonably be expected to affect their decision. Although it often involves having them read and sign a consent form, it is not equivalent to reading and signing a consent form.
- It is your responsibility as a researcher to know and accept your ethical responsibilities.
- You can take several concrete steps to minimize risks and deception in your research. These include making changes to your research design, prescreening

to identify and eliminate high-risk participants, and providing participants with as much information as possible during informed consent and debriefing.

- Your ethical responsibilities continue beyond IRB approval. You need to monitor participants' reactions, be alert for potential violations of confidentiality, and maintain scholarly integrity through the publication process.

Exercises

Questions

1. What is the main goal and purpose of the Institutional Review Board?
2. Before participating in a research experiment, all participants should generally read and sign an informed consent form. What are 3 main goals of this process intended to maintain ethical standards for the scientific work?
3. In a study of attitudes about extramarital affairs, a researcher finds that an acquaintance has participated in the study and reports having cheated on their spouse. They then find themselves torn about whether to report this information to the spouse. What kind of research ethics problem has occurred here? What research procedures should be used to keep this from occurring?

Additional Questions

- Practice: Imagine a study testing the effectiveness of a new drug for treating obsessive-compulsive disorder. Give a hypothetical example of an ethical issue from that could arise in this research.
- Discussion: It has been argued that researchers are not ethically responsible for the misinterpretation or misuse of their research by others. Do you agree? Why or why not?
- Practice: Read the Nuremberg Code, the Belmont Report, and Standard 8 of the APA Ethics Code. List five specific similarities and five specific differences among them.

- Discussion: In a study on the effects of disgust on moral judgment, participants were asked to judge the morality of disgusting acts, including people eating a dead pet and passionate kissing between a brother and sister (Haidt, Koller, & Dias, 1993). If you were on the IRB that reviewed this protocol, what concerns would you have with it? Refer to the appropriate sections of the APA Ethics Code.
- Discussion: How could you conduct a study on the extent to which people obey authority in a way that minimizes risks and deception as much as possible? (Note: Such a study would not have to look at all like Milgram's.)
- Practice: Find a study in a professional journal and create a consent form for that study. Use a standard informed consent template as a guide.

9 Factorial Design

So far, we have focused our consideration of experimental design on the simplest possible designs with a single independent variable with just two conditions administered either between or within groups of participants. Most studies in psychology are more complex than this and, in this chapter, we start to discuss slightly more complex designs. As the complexity of the experimental design increases, the relationship of the data to the experimental hypotheses also increases. These designs allow us to test more interesting and complicated ideas about how psychological constructs interact with each other. However, these designs make the process of drawing inferences from experimental data more challenging. In this chapter, we will be concerned with how to design an experiment with multiple factors (independent variables). Then in Chapter 10, we will review the process of evaluating various patterns of data that can arise from these designs and how we draw conclusions from these.

Cleanliness and moral judgments

Simone Schnall and her colleagues carried out a series of simple studies examining an interesting effect where, in which they found that priming the idea of cleanliness (Experiment 1) or washing one's hands (Experiment 2) led

people to view moral transgressions as less wrong (Schnall, Benton, Harvey, 2008). In a separate study, Schnall and her colleagues investigated whether feeling physically disgusted causes people to make harsher moral judgments (Schnall, Haidt, Clore, Jordan, 2008). In this experiment, they extended this idea to both include disgust created by the testing environment but also accounting for differences in the participants' sensitivity to their own bodily sensations. Participants' feelings of disgust were manipulated by testing them in either a clean room or a messy room that contained dirty dishes, an overflowing wastebasket, and a chewed-up pen. In addition, a self-report questionnaire to measure the amount of attention that people pay to their bodily sensation, described as *private body consciousness*. The primary dependent variable remained the same as in the previous simpler 2-group designs. They measured the harshness of people's moral judgments by describing different behaviors (e.g., eating one's dead dog, failing to return a found wallet) and having participants rate the moral acceptability of each one on a scale of 1 to 7. The primary results of this study were that participants in the messy room were, in fact, more disgusted and made harsher moral judgments than participants in the clean room—but only if they scored relatively high in private body consciousness.

A diagram of the design helps with understanding how the two independent variables or factors here are affecting participants' behavior.

		Tested in Messy Room	
		No	Yes
Private Body Consciousness	Low	Typical moral judgments	Typical moral judgments
	High	Typical moral judgments	Very harsh moral judgments

There are effectively four conditions in this study. There are participants high in private body consciousness (PBC) tested in a messy room, high PBC tested in a clean room, low PBC tested in a messy room and low PBC tested in a clean room. In this chapter we will use diagrams like the one provided here to see how studies like this are organized and carried out. In the next chapter, we will use these diagrams to present data from each of the conditions and review how to interpret the results of studies with this kind of design.

The conclusion drawn in this study depends on describing an interaction between two different variables that were hypothesized to affect the dependent variable (moral acceptability rating). To be able to say that environmentally elicited disgust affected moral judgments but only for people with high sensitivity to their bodily sensations requires a factorial design to incorporate both variables simultaneously. With information about both the environmental variable and the participant variable, the researchers could consider three hypotheses simultaneously. First, did the messiness of the room by itself affect the moral judgments? Second, did participants who scored higher in private body consciousness rate moral judgments differently than those who scored lower? And third, did the messy room affect the higher scoring participants more than the lower scoring participants? This third hypothesis is based on an interaction among the variables. The main conclusion of the study is actually focused on this interaction and looking for these interactions is the primary reason to employ factorial designs in experimental research.

Factorial designs depend on all the same basic experimental design elements discussed in previous chapters, including operational definitions of psychological constructs and control of extraneous variables. The key difference is the use of multiple independent variables that are manipulated (or measured) simultaneously. Studies with multiple dependent variables are also possible, but *multivariate* research design is beyond the scope of this introductory research methods text.

Learning Objectives

1. Understanding one factor designs with more than two levels of the independent variable
2. Explain why researchers often include multiple independent variables in their studies.
3. Define factorial design and use a factorial design table to represent and interpret simple factorial designs.
4. Understand the core hypotheses embedded in a factorial design: main effects and interactions among effects

One Factor Design

The simplest extension from the designs we have discussed so far with two groups or conditions is to consider an experimental design with three different options for the independent variable. For more complicated designs, the term **factor** is often used instead of or synonymously with the term **independent variable**. The conditions that are implemented are described as **levels of the factor**.

As an example, consider a hypothetical design where participants listed to one of three kinds of auditory input while performing a spatial cognition task. The type of music is the experimental factor and the three levels are classical music, electronic dance music, and soothing ocean sounds. Participants completed as many problems as they could in 10 minutes on a “paper cutting and folding” test. Just as in prior designs, the hypothesis is that the type of sounds listened to would influence the score on the test. However, it should be clear that there are already more outcomes to consider than we would have with a 2-group design. With a 2-group design, either the independent variable affects the dependent variable (test scores) or it does not. With three groups, the null hypothesis is that the sounds have no effect and that all three conditions are essentially identical. But we can reject that

null hypothesis if any of the 3 groups shows different performance on the test from the others. The statistical tool to carry out this type of inferential statistic is the Analysis of Variance (ANOVA) which will be discussed in Chapter 10. This analysis provides a p-value that indicates the probability of the data occurring under the null hypothesis and if less than .05, we can conclude the different levels of the factor affected the dependent variable score.

In general, that is only the first step in analyzing data with more than 2 levels. We typically want to know not just that performance differs, but which of the conditions differ from each other. The statistical tool used here is the post-hoc t-test to do all the possible pairwise comparisons and find the differences. Conceptually, what we want to know is (1) "did classical music lead to different performance than ocean sounds?" (2) "did classical music lead to different performance than electronic dance music?" and (3) "did electronic dance music lead to different performance than ocean sounds?" Each of those potential conclusions may have very different meanings for a theory of how auditory input affects spatial cognitive performance. The first thing to note is that this simple extension to just 3 conditions instead of 2 requires us to bring in a new statistical tool, ANOVA, and do a total of 4 analyses to try to understand our data.

In practice, the challenge of drawing inferences from the data in these designs can be even harder when the data are messy. For example, we might observe that performance during classical music is reliably better than ocean sounds, but neither of the other two comparisons is statistically reliable, e.g. performance during the dance music is in-between ocean sounds and classical. These data would leave us in a difficult position for summarizing the findings of our study because different kinds of music both are and are not affecting the dependent variable.

The potential for problem in getting a strong conclusion from this kind of design makes factors with many levels less common in psychological research than just using two levels. Very complex designs with many levels on the factors do get used but often in specific cases with very strong theoretical

foundations and in conjunction with more complex analytical tools. All the statistical tools that will be described in this class are simplified cases derived from a more general approach based on general linear models. Extrapolating to these more complex types of analysis is beyond the scope of this text.

Factorial design

A very common approach in psychological science is to design studies with more than one factor. Researchers' inclusion of multiple independent variables in one experiment is further illustrated by the following actual titles from various professional journals:

- The Effect of Age and Divided Attention on Spontaneous Recognition
- The Effects of Temporal Delay and Orientation on Haptic Object Recognition
- Opening Closed Minds: The Combined Effects of Intergroup Contact and Need for Closure on Prejudice
- Effects of Expectancies and Coping on Pain-Induced Intentions to Smoke
- The Effects of Reduced Food Size and Package Size on the Consumption Behavior of Restrained and Unrestrained Eaters

In each of these cases, we see research that is assessing the effect of at least two factors (independent variables) on some behavior of interest. In each of these studies, the researchers are looking simultaneously at two different IV's that may affect the DV measure. This approach goes importantly beyond examining the effect of each factor by also allowing the researchers to identify interactions between these variables that could not be assessed by doing two successive studies looking at each factor in isolation. Taking the first headline above as an example, we might find that divided attention leads to worse performance on spontaneous recognition but also that this effect is much larger for older adults than younger adults. This would be an example of an interaction among the experimental factors. These are generally the most

interesting effects to study in psychological research but also ones that can pose more difficulties in drawing accurate inferences from.

By far the most common approach to including multiple factors (independent variables) in an experiment is the factorial design, which assesses both the effects of these factors and their interactions. In a factorial design, each level of one independent variable is combined with each level of the others to produce all possible combinations. Each combination, then, becomes a condition in the experiment. Imagine, for example, an experiment on the effect of cell phone use (yes vs. no) and time of day (day vs. night) on driving ability. This is shown in the factorial design table below. The columns of the table represent cell phone use, and the rows represent time of day. The four cells of the table represent the four possible combinations or conditions: using a cell phone during the day, not using a cell phone during the day, using a cell phone at night, and not using a cell phone at night. This particular design is referred to as a 2×2 (read “two-by-two”) factorial design because it combines two variables, each of which has two levels.

		Cell Phone	
		No	Yes
Time of Day	Daytime		
	Nighttime		

If one of the independent variables had a third level (e.g., using a handheld cell phone, using a hands-free cell phone, and not using a cell phone), then it would be a 3×2 factorial design, and there would be six distinct conditions. Notice that the number of possible conditions is the product of the numbers of levels. A 2×2 factorial design has four conditions, a 3×2 factorial design has six conditions, a 4×5 factorial design would have 20 conditions, and so on. Also notice that each number in the notation represents one factor, one independent variable. So by looking at how many numbers are in the notation, you can determine how many independent variables there are in the experiment. 2×2 , 3×3 , and 2×3 designs all have two numbers in the notation and therefore all have two independent variables. The numerical value of each of the numbers represents the number of levels of each independent variable. A 2 means that the independent variable has two levels, a 3 means that the independent variable has three levels, a 4 means it has four levels, etc. To illustrate a 3×3 design has two independent variables, each with three levels (9 conditions), while a $2 \times 2 \times 2$ design has three independent variables, each with two levels (8 conditions). As noted in the discussion of one-factor designs, having 3 levels adds surprising amounts of complexity to interpretation. As a result, it is more common to extend designs to additional factors such as a $2 \times 2 \times 2$ design.

In principle, factorial designs can include any number of independent variables with any number of levels. For example, an experiment could include the type of psychotherapy (cognitive vs. behavioral), the length of the psychotherapy (2 weeks vs. 2 months), and the sex of the psychotherapist (female vs. male). This would be a $2 \times 2 \times 2$ factorial design and would have eight conditions. The table below shows one way to diagram this design. In practice, it is unusual for there to be more than three independent variables with more than two or three levels each. This is for at least two reasons: For one, the number of conditions can quickly become unmanageable. For example, adding a fourth independent variable with three levels (e.g., therapist experience: low vs. medium vs. high) to the current example would make it a $2 \times 2 \times 2 \times 3$ factorial design with 24 distinct conditions. Second, the number of participants required to populate all of these conditions (while

maintaining a reasonable ability to detect a real underlying effect) can render the design unfeasible. For a 2×2 design, we might determine that we need 20 participants in each of the four conditions to have adequate statistical power (sensitivity to detect an effect, discussed in more detail in Chapter 12) for a total of 80. In this hypothetical $2 \times 2 \times 2$ design diagrammed below, we would need to double that number to 160. In general, increasing complexity in factorial designs increases the number of participants required exponentially, making some complex designs essentially infeasible. As a result, we will primarily focus on designs with two independent variables. The general principles discussed here extend in a straightforward way to more complex factorial designs.

		Psychotherapy Type			
		Cognitive		Behavioral	
Length	Two weeks	Female Therapist	Male Therapist	Female Therapist	Male Therapist
	Two months	Female Therapist	Male Therapist	Female Therapist	Male Therapist

Assigning Participants to Conditions

The diagrams in the preceding section are useful in experimental design for planning how to assign participants to conditions and planning the total number of participants to be enrolled in the study. Recall that in a simple between-participants design, each participant is tested in only one condition. In a simple within-participants design, each participant is tested in all conditions. In a factorial experiment, the decision to take the between-participants or within-participants approach must be made separately for each independent variable. In a between-participants factorial design, all of the independent variables are manipulated between participants. For example, each participant would be tested either while using a cell phone or while not using a cell phone and either during the day or during the night. This would mean that each participant would be tested in one and only one of the four possible conditions. This type of design avoids any possible problems with order effects but does generally require a lot of participants to be recruited and enrolled in the study. In modern psychological studies, we prefer having 20-30 participants in each of the conditions meaning a 2×2 design might require 80-120 participants.

It's perfectly acceptable to organize the design as an entirely within-participants factorial design with all of the independent variables are manipulated within participants. In this case, all participants are tested in all four of the conditions, that is, each participant is tested both while using a cell phone and while not using a cell phone and both during the day and during the night. The advantages and disadvantages of these two approaches are the same as those discussed in Chapter 7. The between-participants design is conceptually simpler, avoids order/carryover effects, and minimizes the time and effort of each participant. The within-participants design is more efficient for the researcher and controls all extraneous participant variables.

Since factorial designs have more than one independent variable, it is also possible to manipulate one independent variable between participants and another within participants. This is called a mixed factorial design. For

example, a researcher might choose to treat cell phone use as a within-participants factor by testing the same participants both while using a cell phone and while not using a cell phone (while counterbalancing the order of these two conditions). But they might choose to treat time of day as a between-participants factor by testing each participant either during the day or during the night (perhaps because this only requires them to come in for testing once). Thus, each participant in this mixed design would be tested in two of the four conditions.

An important difference to keep in mind across these design choices is that there are slightly different statistical tools for analyzing data when there is at least one within-participants factor. For statistical analysis, the within-participants factor is typically referred to as having repeated measures in the design. This changes some details of how the analytical tools are run and how the data are formatted for analysis. This will be reviewed in Chapter 11.

Regardless of whether the design is between participants, within participants, or mixed, the actual assignment of participants to conditions or orders of conditions is typically done randomly. A diagram of the design can be used to both plan the total sample size and also track the accumulation of data so that the number of participants in each condition stays relatively balanced. For statistical analysis, it is best if the number of participants in each of the design cells (conditions) is the same or similar when data collection is completed.

Non-Manipulated Independent Variables

In many factorial designs, one of the independent variables can also be a non-manipulated independent variable. In this case, the researcher measures but does not manipulate the factor and is often a characteristic that varies across participants. The study by Schnall et al. (2008) is an example of this that incorporated the participants rating of their “private body consciousness” in the design. Scores on a measure of this characteristic were used to assign participants to either high or low “condition” on this measure. In design of

this kind of factor, it is necessary to have a plan for the distinction of the rating scale into the high/low categories. This can be done by using prior research with the scale provide definitions of the categories. It can also be done by using a median split of participants. Since the median value in a group is defined as the number that splits the groups into two equal halves, this technique is guaranteed to give equal sized samples across the two levels of this factor.

Studies with this generally approach to design are extremely common and can provide important insight into how the manipulated independent variable might have different effects on different people. In the Schnall et al. (2008), the manipulated variable was the environment, specifically how messy the room was in which participants made moral judgments. In the process of science, it is not uncommon to have developed a hypothesis that the messy room might cause people to make harsher moral judgments, implying a typical two-condition research study that is also consistent with prior research published by the same group. However, in data collection, it might become clear that the effect of the room is not statistically reliable in the simple design, leading researchers to examine why this effect might be influencing some participants but not others. That might provide the insight that the participants varied in their sensitivity to the room, leading to the incorporation of the second factor in which private body consciousness was measured and leading to the study's conclusions. The end result is a richer theoretical understanding of the idea that disgust can cause harsher moral judgments but that this effect will likely vary across people at least by differences in what causes them to experience disgust.

In considering this type of design, it is important to remember that when non-manipulated independent variables are participant variables, they are by definition between-participants factors. These variables are generally assumed to be static, which is why they are measured instead of manipulated (unlike mood, for example). As long as one independent variable is manipulated, the design is still considered an experimental design overall, no matter how many other non-manipulated factors are included. However,

conclusions about the non-manipulated variables need to incorporate the fact that these were not manipulated. We would want to avoid statements such as "high private body consciousness caused harsh moral judgment in a messy room" because it implies a causal effect on a variable that was not controlled. We would prefer to state the conclusion as a "messy room caused harder moral judgments in participants with high private body consciousness." As we will review in Chapter 16, non-experimental relationships among variables are more difficult to interpret due to needed to consider and attempt to rule out alternate explanations.

Hypochondria and Memory for Health-related Words

Another example of a design with one manipulated factor and one non-manipulated participant variable is a study in which participants were exposed to several words that they were later asked to recall (Brown, Kosslyn, Delamater, Fama, Barsky, 1999). The manipulated independent variable was the type of word. Some were negative health-related words (e.g., tumor, coronary), and others were not health related (e.g., election, geometry). The non-manipulated independent variable was whether participants were high or low in hypochondriasis (excessive concern with ordinary bodily symptoms). The result of this study was that the participants high in hypochondriasis were better than those low in hypochondriasis at recalling the health-related words, but they were no better at recalling the non-health-related words.

Non-Experimental Studies With Factorial Designs

Thus far we have seen that factorial experiments can include manipulated independent variables or a combination of manipulated and non-manipulated independent variables. But factorial designs can also include only non-manipulated independent variables, in which case they are no longer experiments but are instead non-experimental in nature.

Consider a hypothetical study in which a researcher simply measures both the moods and the self-esteem of several participants—categorizing them as having either a positive or negative mood and as being either high or low in self-esteem—along with their willingness to have unprotected sexual intercourse. This can be conceptualized as a 2×2 factorial design with mood (positive vs. negative) and self-esteem (high vs. low) as non-manipulated between-participants factors. Willingness to have unprotected sex is the dependent variable. But because neither independent variable in this example was manipulated, it is a non-experimental study rather than an experiment. This is important because, as always, one must be cautious about inferring causality from non-experimental studies because of the directionality and third-variable problems. Directionality is a challenge in some non-experimental designs where you are not sure which factor happened first. For example, willingness to have unprotected sex could be affecting mood or participant's rating of self-esteem so we do not necessarily know the direction of causality. There may also be additional extraneous variables that are causing all of these measures to increase together, creating the apparent correlation.

Hypotheses in Factorial Designs

The primary goal of using a factorial design is to look for interactions among the design factors. An interaction is defined as one of the design factors modifying the effect of another design factor. For example, in the very first experiment diagrammed above, the effect of a cell phone on driving quality

might be moderate during the daytime, but much larger at night. We would then say that the time of day influences the effect of the cell phone. Factorial designs are always designed to explore the interaction of factors. If we simply wanted to look at the effect of cell phones on driving and time of day on driving independently, we would run two parallel studies that each had a simpler 2-condition design.

At the same time, in the evaluation of the results of a factorial design, we have to systematically consider all the embedded hypotheses. In a 2×2 design, there are three hypotheses that are automatically being tested. We describe these as two main effects and the interaction term. In this example, one main effect is the overall effect of cell phone use on driving, but note that this is evaluated while ignoring any effect of time of day. Main effects measure the overall impact of that factor's levels on the DV independently of everything else (the other factor or any interactions). A second main effect in this design is the effect of time of day on driving quality not including any effect of cell phone use. The technique for visualizing these main effects is to calculate marginal means from the results, which will be discussed in Chapter 10. Although the goal of the experiment may be to examine the interaction between factors, the results should always be presented comprehensively and include the main effects and interactions of interest.

The number of embedded hypotheses goes up quickly as design complexity is increased. For a $2 \times 2 \times 2$ design, we now have 3 main effects and 4 interaction terms to consider. The main effects are one for each of the three factors. However, we now have potential interactions between the first and second factor, the first and third factor and the second and third factor. Then there is a potential three-way interaction among all the factors. In Chapter 10, we will review how to interpret results from factorial designs, identify the most common kinds of interactions and how to connect these to the experimental hypotheses.

Key Takeaways

- Researchers often include multiple independent variables in their experiments. The most common approach is the factorial design, in which each level of one independent variable is combined with each level of the others to create all possible conditions.
- Each independent variable can be manipulated between-participants or within-participants.
- Non-manipulated independent variables (gender) can be included in factorial designs, however, they limit the causal conclusions that can be made about the effects of the non-manipulated variable on the dependent variable.
- In a factorial design, the main effect of an independent variable is its overall effect averaged across all other independent variables. There is one main effect for each independent variable.
- There is an interaction between two independent variables when the effect of one depends on the level of the other. Some of the most interesting research questions and results in psychology are specifically about interactions.
- A simple effects analysis provides a means for researchers to break down interactions by examining the effect of each independent variable at each level of the other independent variable.

Exercises

Question 1

- After watching a group of nursery school children, we get the idea that some toys are more popular with children than others are. We would like to test the difference in time spent playing with toys that are used for building (e.g. blocks) and toys that are not (e.g., stuffed animals). Since there are many differences between boys and girls, we would also like to look at gender as an independent variable.
- Outline a factorial design for this study and describe the operational definitions of the factors (independent variables) and dependent variable. Speculate about hypotheses for what you might see if you ran this study (effect of toy, gender and any interaction).

Additional Questions

- Practice: Return to the five article titles presented at the beginning of this section. For each one, identify the independent variables and the dependent variable.
- Practice: Create a factorial design table for an experiment on the effects of room temperature and noise level on performance on the MCAT standardized test. Be sure to indicate whether each independent variable will be manipulated between-participants or within-participants and explain why.

10 Interpretation of Factorial Data

Even though factorial designs are a relatively small conceptual increase in the complexity of an experimental protocol, they can pose surprisingly difficult challenges when trying to make sense of the data. Even experienced researchers can make mistakes in characterizing the effects observed and drawing inferences from factorial results. In this chapter, we will describe a systematic process of working through factorial data examining the main effects and then any interactions among the factors. Data visualizations are very helpful in providing an overview and with some practice, common outcome patterns can be recognized as visual patterns in data graphs.

In addition, we will introduce the statistical concept of **effect size** to help describe interaction effects. As we will see later, modern psychological science is working to incorporate improved models of statistical inference and shift away from a reliance on a simple rejection of the null hypothesis by the familiar standard $p < .05$. Here we will examine the unstandardized effect sizes, which are the simple difference in the mean performance across conditions. These will help to identify types of interaction effects: super-additive, 3:1, and crossover. The statistical tool of a post-hoc t-test is used to assess specific contrasts between conditions within a factorial design.

Learning Objectives

1. Distinguish between **main effects** and **interactions** in the results from a factorial design and recognize each.
2. Understand factorial data tables by looking at both individual conditions as **cells**, and **marginal means**.
3. Interpret and understand bar graphs and line graphs showing the results of studies with factorial designs.
4. Understand and know how to describe basic types of interactions: **super-additive, 3:1, crossover**.
5. Know the role of **post hoc t-tests** to further characterize data and test targeted hypotheses.

Main Effects

In factorial designs, the main hypotheses are tested as main effects and interactions. A main effect is the effect of one independent variable on the dependent variable—averaging across the levels of the other independent variable. Common patterns of data are illustrated here with both means tables and figures to illustrate the results.

In a means table, the average performance of participants in each of the experimental conditions is shown separately, typically with means and the standard deviations shown below in parentheses. Note that in APA format, tables are accompanied by a table note indicating what the parenthetical numbers are.

In the tables and graphs below, we have an abstract design based on two factors, Factor 1 and Factor 2. Factor 1 has two conditions, A and B. Factor 2 has two conditions, X and Y. Each participant in the study is in one of the four possibilities: AY, BY, AX, BX. For illustration, we will assume a dependent variable that is scored on a 1 to 10 scale and that 20 participants were run in each of the 4 conditions in a fully between-participants design. In Chapter 9,

we reviewed a variety of examples of manipulations and participant variables that can be used to test hypotheses in 2×2 designs. Here, we will keep the design specifics abstract to focus on how to interpret quantitative data.

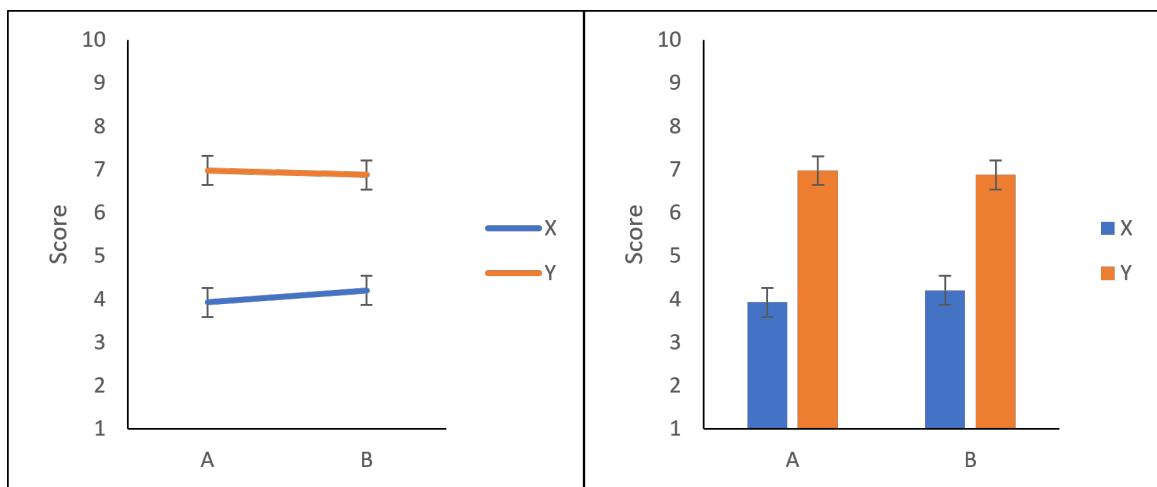
Six simulated data outcomes will be shown, each with both a **means table** and two figures illustrating the results. A means table includes the mean performance from all the participants in that specific condition of the design, typically with a measure of variance underneath in parentheses. The bottom row shows the **marginal means** across Factor 1 conditions, combining across the two levels of Factor 2, essentially setting Factor 2 aside. The very rightmost column shows the marginal means of Factor 2, combining across Factor 1 levels.

Each dataset is shown graphed as both a line graph and a bar graph. In general, experimental reports will never include both of these as they are effectively completely redundant. However, line graphs and bar plots emphasize different elements of the differences across factors slightly differently. When preparing a research report that includes a figure to illustrate the results from a factorial design, the decision of which kind of plot to use should reflect a choice of the graph that most effectively communicates the results. For these simulated data, both figures types are shown so that these differences are concretely visual in order to help with that choice.

Simulated Data 1

In the first simulated dataset, one main effect (Factor 2) is reliable, the other main effect is not affecting the DV and no interaction between the factors occurred.

Means Table		Factor 1		Mean
		A	B	
Factor 2	X	3.27 (1.51)	7.09 (1.35)	5.18 (2.39)
	Y	3.65 (1.08)	7.44 (1.44)	5.55 (2.28)
	Mean	3.46 (1.33)	7.26 (1.41)	



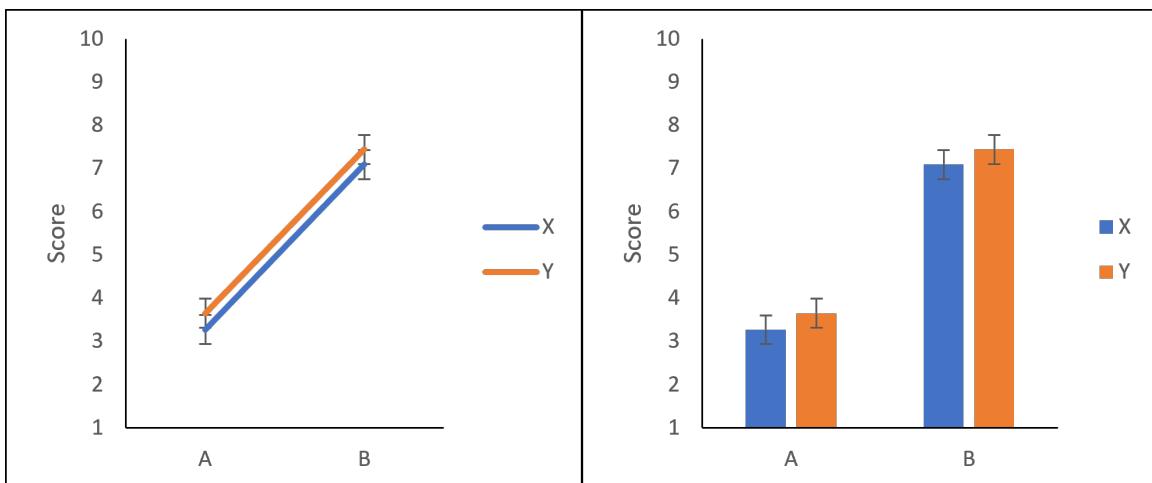
In the line graph, this pattern creates two roughly parallel, nearly horizontal lines. The distance between the lines is the effect of Factor 2. The fact that they are flat (horizontal) is reflecting the absence of an effect of Factor 1. The magnitude of the Factor 2 difference is also seen in the marginal means for Factor 2, the rightmost column of the data table ($M = 4.07$ for X, $M = 6.93$ for Y). The bar plot also clearly shows the Y condition (orange) producing higher values than the X condition (blue).

With the outcome being an effect on just one factor, we can also see the magnitude of this effect in the marginal means. The score is about 3 points higher for the Y condition than the X condition. The uncorrected effect size for Factor 2 in these data is an increase in the DV of 2.86. Later in Chapters 12 and 16 we will discuss how to use effect sizes for more sophisticated assessment of experimental effects and how to incorporate these into predicting the power and sensitivity of our designs to our hypothesized findings.

Simulated Data 2

In the second simulated dataset, one main effect (Factor 1) is reliable, the other main effect is not affecting the DV and no interaction between the factors occurred. In the line graph, this pattern creates two nearly overlapping lines that slope up the same way. The increase across the graph (from A to B) reflects the effect of Factor 1. The lack of vertical separation is due to the non-effect from Factor 2. The magnitude of the Factor 1 difference is also seen in the marginal means for Factor 1, the bottom row of the data table ($M = 3.46$ for A, $M = 7.26$ for B). The bar plot also clearly shows the B condition (right 2 bars) producing higher values than the A condition (left 2 bars).

Means Table		Factor 1		Mean
		A	B	
Factor 2	X	3.93 (1.43)	4.21 (1.22)	4.07 (1.33)
	Y	6.98 (1.28)	6.88 (1.37)	6.93 (1.33)
Mean		5.45 (2.04)	5.55 (1.86)	



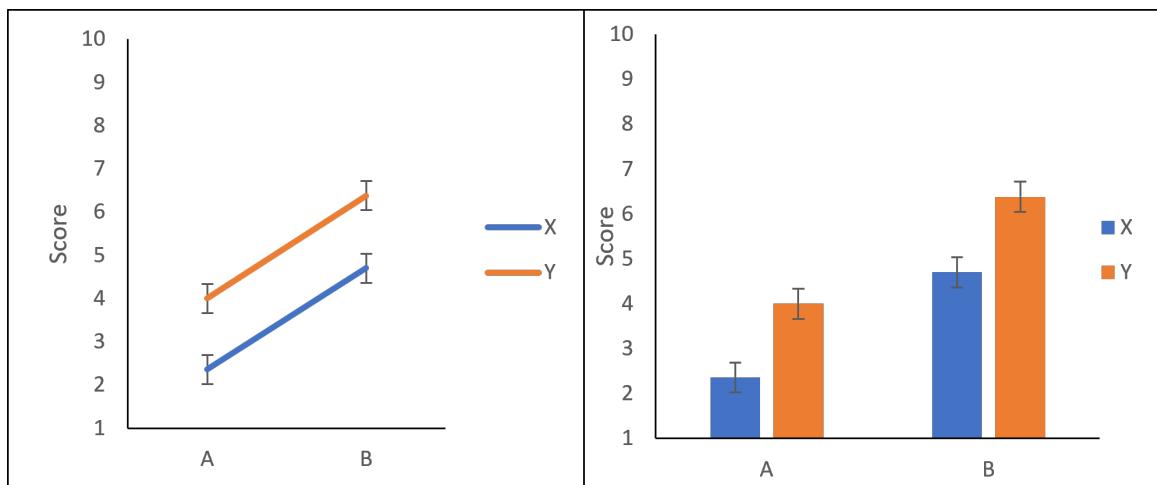
The results of this simulation are essentially the same as the first data set. We could easily transform the data by simply relabeling the factors, there are no specific rules telling you which of your independent variables in your design are Factor 1 or Factor 2. The point of providing both graphs is to indicate that when there is one reliable main effect on the dependent variable in a factorial design, you can see that in one of two ways on a line graph. It can be visually presented as two separated, roughly flat parallel lines as in the first simulation, or two nearly overlapping, steeply sloped lines as in this second graph. For bar plots, there will always be two higher bars and two lower bars reflecting the mean differences affected by that factor.

We can also still characterize the effect of Factor 1 by its uncorrected effect size seen in the marginal means. Condition B is scoring an average of 3.8 points higher than condition A.

Simulated Data 3

In this third simulated data set, both main effects are reliable but there is no interaction between them. This pattern is often mistaken for suggesting an interaction between the factors but there is none. The highest performing condition (BY) is showing the effects of both Factor 1 and Factor 2 additively.

Means Table		Factor 1		Mean
		A	B	
Factor 2	X	2.36 (1.54)	4.70 (1.54)	3.53 (1.93)
	Y	4.00 (1.08)	6.38 (1.44)	5.19 (1.84)
	Mean	3.18 (1.63)	5.54 (1.74)	



In the line graph, this pattern creates two parallel, separated lines with the same slope. As we will see below, differing slopes on a line graph is a useful visual signal of the occurrence of an interaction between factors. The increase across the graph (from A to B) reflects the effect of Factor 1. The vertical separation is due to the effect of Factor 2 (from X to Y). The marginal means

for both factors show the magnitude of the two effects independent of each other (note that these would not be independent if there was an interaction). The bar plot also shows both effects but does not imply the parallel slopes quite as easily as the line graph.

Once again looking at the marginal means for uncorrected effect sizes we see the Factor 1 effect is 2.36 points higher from A to B. The effect for Factor 2 is 1.66 points higher from X to Y. The highest scoring condition, BY, is larger than the lowest scoring condition, AX, by roughly the sum of these two effects. This is synonymous with saying there is no interaction, i.e., that the two main effects simply add together when both factors are present.

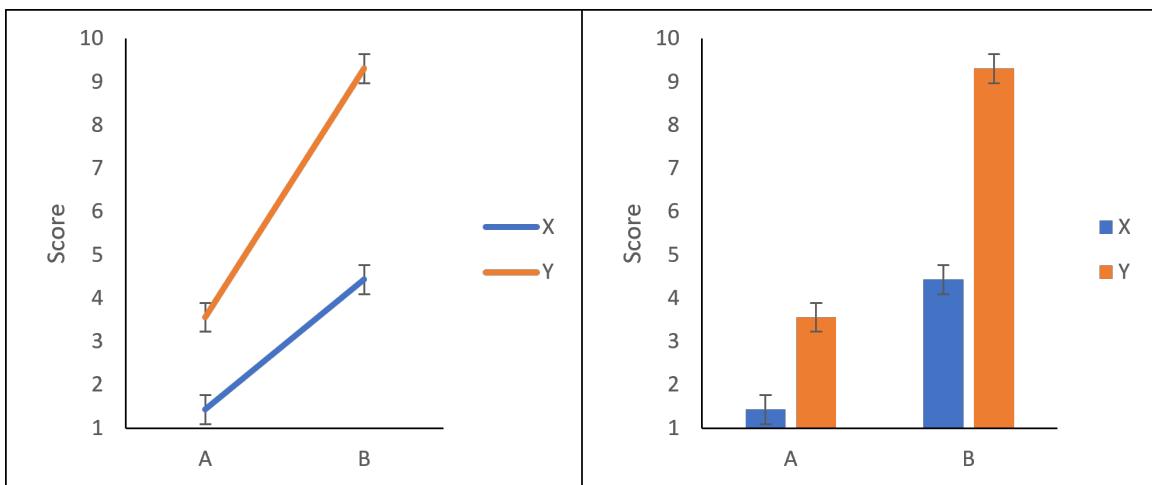
The idea that each factor can affect the DV is the basic meaning of what a main effect is and is the same way we discussed effects of our IV on the DV in the simpler 2-group designs in Chapters 2-5. With two factors, we now have two things that can affect the DV. Technically, the question of whether these factors interact with each other is mathematically the same as whether you can simply add them up to describe the data, that is, are they additive? You can see that they are additive here in the means table. The line graph also provides an important visual element to help see that. The fact that the lines are parallel is an easy way to see that the factors are combining additively. In the next three examples, we will see that the line slopes differ whenever there is an interaction. This is a visual indication that one of the factors is affecting how the other factor influences the DV, which is the definition of observing an interaction among the factors in the design.

Simulated Data 4

In these data, we have two main effects and a **super-additive interaction**.

The effect of both Factor 1 and Factor 2 are to increase scores on the DV. In addition, performance in the combined condition (BY) is higher than would be predicted if the two factor effects summed together.

Means Table		Factor 1		Mean
		A	B	
Factor 2	X	1.43 (1.32)	4.43 (2.00)	2.93 (2.26)
	Y	3.56 (1.39)	9.31 (1.38)	6.43 (3.19)
	Mean	2.50 (1.72)	6.87 (2.98)	



We can describe this effect as saying the effect of Factor 1 was particularly strong in the Y condition of Factor 2. It would also be correct to describe this effect as the effect of Factor 2 was particularly strong in the B condition of Factor 1. This is another example of the symmetry of factorial designs. The design does not inherently prioritize one factor over another and there are

usually at least two equivalent ways to describe the results.

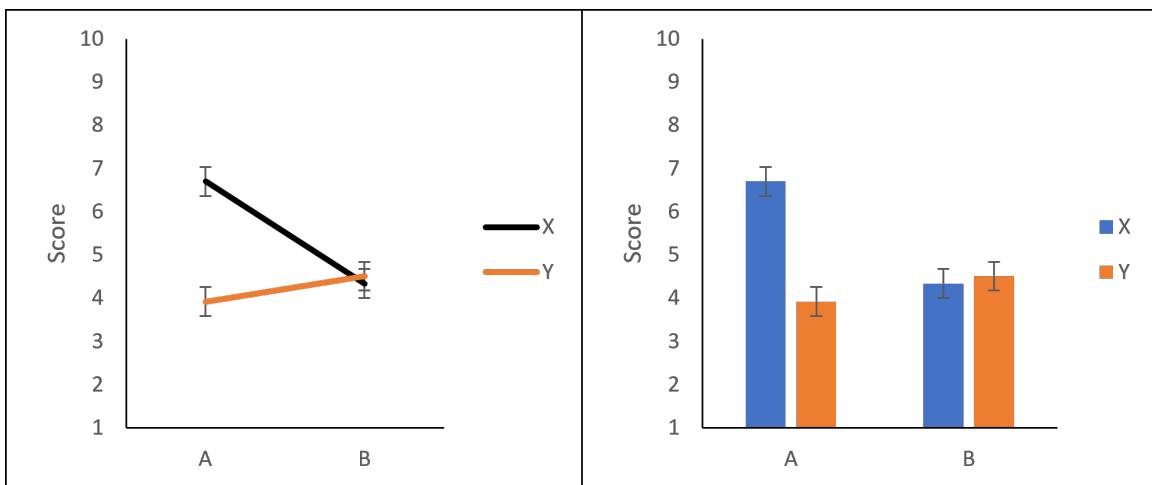
In the line graph, this pattern creates two separated lines that do not have the same slope. The different slopes are what visually signal the presence of an interaction. The main effects are visible in the same way as previous graphs. The left to right increase reflects Factor 1; the separation in lines reflects Factor 2. The marginal means for both factors show an estimate of the effects somewhat independently of each other but note how the BY condition stands out from the marginal means to show the super-additivity. The bar plot also shows both effects but the difference in slopes is again not quite as visible as in the line graph.

Compare the different here in the BY condition to the AX condition from Example 3 above. Here the BY condition is scoring even higher than the two main effects would predict independently. This kind of interaction is described as super-additive because scores in the BY condition are greater than the sum of the main effects A/B and X/Y. This interaction reflects something additional pushing up the DV score when both factors are present simultaneously that is different than either factor in isolation. The *something additional* is usually the goal in a 2x2 design and aimed to learn something new about the component variables that requires measurements across manipulations of both.

Simulated Data 5

A common data pattern in 2x2 designs is a **3:1 interaction**. In this case, one of the conditions is producing a different score from the others, which are all roughly similar (e.g., 1 high score, 3 low scores). We could describe this result using language similar to the super-additive interaction by saying the effect of Factor 2 had a large effect in the A condition of Factor 1 but little or no effect in the B condition.

Means Table		Factor 1		Mean
		A	B	
Factor 2	X	6.70 (1.76)	4.34 (1.58)	5.52 (2.04)
	Y	3.92 (1.42)	4.51 (1.29)	4.21 (1.39)
	Mean	5.31 (2.11)	4.43 (1.45)	



Technically this is just the opposite of the super-additive case (sub-additive) where the effect of Factor 2 is smaller than predicted by the main effects in the B condition. However, it is quite common to describe this result as

saying Factor 2 only matters in the A condition and does not apply to the B condition. When Factor 1 is a participants variable (e.g., an personality variable or measure like “math identification”) we are observing a case where an experimental manipulation only appears to affect one subgroup of the population. This type of design and outcome are fairly common in psychological science.

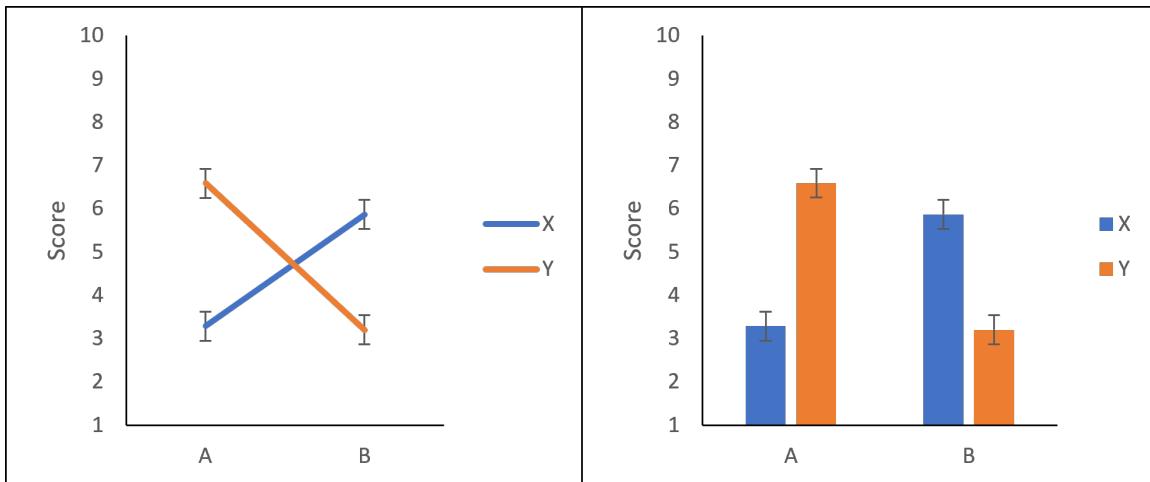
In this type of interaction we see that the bar plot graph illustrates the form of the data more effectively than the line graph. The AX condition stands out from the other three conditions which produce similar levels of performance. The one standing out from the other three is the basis for describing this as a 3:1 interaction type. We can still see that the lines are not parallel in the line graph, which reflects the fact that there is an interaction between factors.

This is also a special case in that the marginal means on the table are not very helpful in understanding the data. Both factors are showing a hint of a main effect on the marginal means, but describing the data in terms of the main effects does not necessarily contribute effectively to communicating the 3:1 interaction in the data. We could characterize this kind of effect as saying the influence of Factor 2 goes away in condition B (of Factor 1). Synonymously, we could say that the influence of Factor 2 only occurs in condition A (of Factor 1). The preferred language depends on the domain being studied and the experimental hypothesis.

Simulated Data 6

Here we see the case where the effects of the factors are essentially inverted across conditions of the other factor, which produces a **crossover interaction**. This are somewhat complex to describe fully. We can say that the effect of Factor 2 in the A condition of Factor 1 is to increase performance while the effect of Factor 2 decreases performance in the B condition. Written descriptions tend to be wordy and these interactions are examples of where data visualizations are very valuable.

Means Table		Factor 1		Mean
		A	B	
Factor 2	X	3.29 (1.07)	5.87 (1.54)	4.58 (1.85)
	Y	6.59 (1.63)	3.20 (1.31)	4.90 (2.25)
	Mean	4.94 (2.15)	4.53 (1.96)	



In this type of interaction, we see that the line graph can make a clear visual signal of the crossover, essentially drawing an X on the graph. It does not always create this shape as this does not occur if there is also a large main effect of Factor 2 that separates the lines. As in all interactions, the slopes of the two lines are markedly different. The structure of the data is also quite visible on the bar plot graph.

In these particular simulated data, there are no main effects of either factor as can be seen in the similar numbers across all the marginal means. This kind of data is often of theoretical importance because it suggests that neither factor has simple effects on the dependent variable and that prior work with simpler 2-group designs might have produced inconsistent or null results. As a reminder, studies that produce a null result do not establish that the IV had no effect on the DV, but only tell us that the study did not work – the IV did not affect the DV under the procedure and sampling conditions of that study. For these types of findings, we would not focus on the effect sizes characterized by the main effects. Instead, we might characterize the differences in specific conditions contrasted by use of a post-hoc t-test.

Post-hoc t-tests

The analytical tool used to identify reliable main effects and interactions is the ANOVA (analysis of variance). As we shall see, the output of the ANOVA will give us a statistical measure (the F ratio) and a p value (the probability of the data under the null hypothesis, as always) for each of the two main effects and interaction. However, the ANOVA itself does not indicate anything about the direction of the effects or the form of the interaction. For the main effects, we can use the descriptive statistics like the marginal means to support the inferential statement about differences in conditions. For the interaction term, it is often necessary to actually graph the data and visualize the form of the interaction in order to describe it accurately in the text. Nothing in the output of the ANOVA procedure indicates the form of the interaction.

In addition, for some interaction types, particularly 3:1 and cross-over interactions, it may be theoretically important to further characterize the data by targeted analysis of a subset of the data. For example, in the 3:1 interaction above, we may want to know whether the X condition is producing reliably higher numbers than the Y condition when only considering participants from the A condition of Factor 1. The tool for this is a **post-hoc t-test** on just the participants from condition A comparing scores in the X and Y conditions. This type of test is described as post-hoc because it is done after obtaining the ANOVA results and may also be described as a protected t-test. The standard analytical tool for carrying out this analysis is the **Tukey's HSD test**, which provides some additional statistical rigor for cases where multiple post-hoc comparisons may be evaluated.

For a crossover design, we can use post-hoc t-tests to evaluate both conditions separately to see if the effect of Factor 2 is reliable in just the A condition (Factor 1) and also is it reliable in the B condition. It should be noted that the fact that the interaction is statistically reliable does not automatically require that these post-hoc t-tests are individually reliable. A reliable interaction means the effects of one factor (e.g., Factor 2) are reliably

different across conditions of the other factor (e.g., Factor 1). That does not mean the effects are reliably within each condition considered separately. A reliable post-hoc t-test allows for a slightly stronger description of the results, which is often theoretically relevant.

We should also be aware when using these follow-up t-tests that there is some risk of weakening our conclusions through running too many parallel comparisons. Care needs to be taken in larger factorial designs where there can be a lot of individual conditions that we might be interested in specific comparisons between. For complex designs, we might need to consider using a Bonferroni correction for multiple comparisons or restricting our analysis to pre-specified hypotheses of specific importance.

Key Takeaways

- Main effects describe the effect of a single factor, separate from the other factors or any interaction
- Each factor in a design has the potential for a main effect. In a 2x2 design, there are two main effects.
- Main effects are visible in the marginal means in a data table.
- An interaction among factors reflects one factor modifying the effect of another factor.
- When one factor increases the impact of the other factor, it creates a super-additive interaction.
- When one factor reverses the effect of the other factor, it creates a cross-over interaction.
- When the effect of a factor is only visible in conjunction with the other factor, it creates a 3:1 interaction.
- The description and interpretation of interaction terms requires explaining the relationships among all related cells of the design, all four conditions for a 2x2

Exercises

TBA

11 Participant Sampling

In every psychological research study, we have a group of participants who are recruited to voluntarily participate in the research project. These participants are our sample, and they are drawn from a larger population. This is part of our statistical model as we take our descriptive statistics about the mean performance and variance from our sample and try to evaluate how accurate these numbers are with respect to the population. For statistics and our inferences about difference, the number of participants is a critical aspect of this calculation with more participants, or larger n , always being better.

In drawing inferences from our data, it is also necessary to consider where this sample came from, how it was recruited and how it might relate to or differ from the broader population. When we draw a conclusion from our data, who do we expect (hope) the conclusions will apply to? The answer to that question will vary across subdisciplines within psychology. Some areas are ambitious, such as cognitive psychology, which hopes to draw inferences that apply to all humans about memory, perception, or other basic human cognitive processes. Some areas are much more specific, such as developmental psychology, which may aim to draw inferences about

***Psychological science is a science of people,
by people and for people.***

behavior in a very specific age range (e.g., 6-month old infants, or 12-16yo adolescents). Clinical psychology and neuropsychology often aim to draw inferences about people with specific psychological challenges, sometimes aiming to both understand these challenges and what they might apply to the broader population (healthy controls). Some areas are more complex, such as social psychology which looks for behavior that may be general to all humans, but may also be strongly culturally or socially influenced (e.g., stereotype bias). Biologically-oriented psychology disciplines such as health and neuroscience look for generalities but acknowledge that these can be influenced by biological differences such as genetics.

In some of these domains, the research hypothesis provides a very specific approach to identifying participants who can be invited to participate in the research study. However, for a wide range of research topics, participants are recruited broadly from the available population and in these cases, some attention needs to be paid to the sampling procedure to identify if that affects the conclusions of the study. Thus far, most of our research design has been aimed at good operational definitions and control of extraneous variables to avoid confounds to maximize our internal validity. The separate question of whether the conclusions from an internally valid study apply broadly to the whole human population is a question of external validity.

External validity describes the degree to which the conclusions of the research study can be applied to the rest of the population outside the specific sample who participated. Good external validity means the results can be **generalized** to the whole population. External validity can be limited if some aspect of the procedure to recruit participants into the sample accidentally introduced some bias such that the sample is no longer representative of the population. Examples of this are studies run exclusively on undergraduate populations, which typically have a very restricted age range (among other characteristics). Inferences drawn about the operation of memory or perception from this range may not apply to all other ages. Some of these limitations are obvious and implicit. Nobody expects studies on memory for reading prose passages to apply to pre-verbal infants. Others are subtler,

such as the finding that some visual illusions are not perceived universally across cultures but may reflect the experience of the participants with stimuli related to the illusions.

In some cases, the recruiting (sampling) method used to carry out the research can introduce bias into the results. For example, research on high-risk behavior has to carefully consider how to find participants, e.g., an advertisement for a study on sexual attitudes or behavior may not recruit a sample of participants that is representative of the population. This can be a difficult issue to resolve since we ethically require participants to voluntarily participate in a research study. People who are reluctant to talk about this topic even when their privacy is guaranteed may be persistently underrepresented in those studies and it can be difficult even to assess the size of the problem (which in this case, likely also varies substantially across different social and cultural groups).

The question of sampling methodology is most often considered in the context of non-experimental research, especially survey research and the related area of polling. Here we will introduce the main underlying ideas and methodologies as they can be applied to sampling and generalizability in experimental research. In Chapter 17 we will return to this topic in the context of non-experimental research. Many of the sophisticated sampling methods described here, such as stratified random sampling are critical for survey/polling methodology but less commonly used explicitly in experimental research.

Learning Objectives

1. Understand the differences in different kinds of **sampling** and their implications for drawing conclusions from the research.
2. Define **sampling bias** in general and **non-response bias** in particular.
3. Understand how to explain the limitations of the recruiting process using in a research study and how this might affect the **external validity** of the conclusions.

Sampling and Measurement

In addition to identifying which variables to manipulate and measure, and operationally defining those variables, researchers need to identify the **population** of interest. Researchers in psychology are usually interested in drawing conclusions about some very large group of people. This is called the population. It could be all American teenagers, children with autism, professional athletes, or even just human beings—depending on the interests and goals of the researcher. But they usually study only a small subset or sample of the population. For example, a researcher might measure the talkativeness of a few hundred university students with the intention of drawing conclusions about the talkativeness of men and women in general. It is important, therefore, for researchers to use a representative sample—one that is similar to the population in important respects.

One method of obtaining a sample is **simple random sampling**, in which every member of the population has an equal chance of being selected for the sample. For example, a pollster could start with a list of all the registered voters in a city (the population), randomly select 100 of them from the list (the sample) and ask those 100 whom they intend to vote for. Unfortunately, random sampling is difficult or impossible in most psychological research because the populations are less clearly defined than the registered voters in

a city. How could a researcher give all American teenagers or all children with autism an equal chance of being selected for a sample? The most common alternative to random sampling is convenience sampling, in which the sample consists of individuals who happen to be nearby and willing to participate (such as introductory psychology students). Of course, the obvious problem with convenience sampling is that the sample might not be representative of the population and therefore it may be less appropriate to generalize the results from the sample to that population.

Essentially all psychological research involves sampling—selecting a sample to study from the population of interest. Sampling falls into two broad categories. The first category, **probability sampling**, occurs when the researcher can specify the probability that each member of the population will be selected for the sample. The second is **non-probability sampling**, which occurs when the researcher cannot specify these probabilities. Most psychological research involves non-probability sampling. For example, **convenience sampling**—studying individuals who happen to be nearby and willing to participate—is a very common form of non-probability sampling used in psychological research. Other forms of non-probability sampling include **snowball sampling** in which existing research participants help recruit additional participants for the study, **quota sampling** in which subgroups in the sample are recruited to be proportional to those subgroups in the population, and **self-selection sampling** in which individuals choose to take part in the research on their own accord, without being approached by the researcher directly.

Compared with non-probability sampling, probability sampling requires a very clear specification of the population, which of course depends on the research questions to be answered. The population might be all registered voters in Washington State, all American consumers who have purchased a car in the past year, women in the Seattle over 40 years old who have received a mammogram in the past decade, or all the alumni of a particular university. Once the population has been specified, probability sampling requires a sampling frame. This sampling frame is essentially a list of all the members

of the population from which to select the respondents. Sampling frames can come from a variety of sources, including telephone directories, lists of registered voters, and hospital or insurance records. In some cases, a map can serve as a sampling frame, allowing for the selection of cities, streets, or households.

There are a variety of different probability sampling methods. Simple random sampling is done in such a way that each individual in the population has an equal probability of being selected for the sample. This type of sampling could involve putting the names of all individuals in the sampling frame into a hat, mixing them up, and then drawing out the number needed for the sample. Given that most sampling frames take the form of computer files, random sampling is more likely to involve computerized sorting or selection of respondents. A common approach in telephone surveys is random-digit dialing, in which a computer randomly generates phone numbers from among the possible phone numbers within a given geographic area.

A common alternative to simple random sampling is stratified random sampling, in which the population is divided into different subgroups or *strata* (usually based on demographic characteristics) and then a random sample is taken from each *stratum*. **Proportionate stratified random sampling** can be used to select a sample in which the proportion of respondents in each of various subgroups matches the proportion in the population. For example, because about 12.6% of the American population is African American, stratified random sampling can be used to ensure that a survey of 1,000 American adults includes about 126 African-American respondents. Disproportionate stratified random sampling can also be used to sample extra respondents from particularly small subgroups—allowing valid conclusions to be drawn about those subgroups. For example, because Asian Americans make up a relatively small percentage of the American population (about 5.6%), a simple random sample of 1,000 American adults might include too few Asian Americans to draw any conclusions about them as distinct from any other subgroup. If representation is important to the research question, however, then disproportionate stratified random sampling could be used to

ensure that enough Asian-American respondents are included in the sample to draw valid conclusions about Asian Americans a whole.

Yet another type of probability sampling is **cluster sampling**, in which larger clusters of individuals are randomly sampled and then individuals within each cluster are randomly sampled. This is the only probability sampling method that does not require a sampling frame. For example, to select a sample of small-town residents in Washington, a researcher might randomly select several small towns and then randomly select several individuals within each town. Cluster sampling is especially useful for surveys that involve face-to-face interviewing because it minimizes the amount of traveling that the interviewers must do. For example, instead of traveling to 200 small towns to interview 200 residents, a research team could travel to 10 small towns and interview 20 residents of each. The National Comorbidity Survey was done using a form of cluster sampling.

Sampling Bias

Probability sampling was developed in large part to address the issue of sampling bias. Sampling bias occurs when a sample is selected in such a way that it is not representative of the entire population and therefore produces inaccurate results. This bias was the reason that the Literary Digest straw poll was so far off in its prediction of the 1936 presidential election. The mailing lists used came largely from telephone directories and lists of registered automobile owners, which over-represented wealthier people, who were more likely to vote for Landon. Gallup was successful because he knew about this bias and found ways to sample less wealthy people as well.

There is one form of sampling bias that even careful random sampling is subject to. It is almost never the case that everyone selected for the sample actually responds to the survey. Some may have died or moved away, and others may decline to participate because they are too busy, are not interested in the survey topic, or do not participate in surveys on principle. If these survey non-responders differ from survey responders in systematic

ways, then this difference can produce **non-response bias**. For example, in a mail survey on alcohol consumption, researcher Vivienne Lahaut and colleagues found that only about half the sample responded after the initial contact and two follow-up reminders (Lahaut, Jansen, van de Mheen, Garretsen, 2002). The danger here is that the half who responded might have different patterns of alcohol consumption than the half who did not, which could lead to inaccurate conclusions on the part of the researchers. So to test for non-response bias, the researchers later made unannounced visits to the homes of a subset of the non-responders—coming back up to five times if they did not find them at home. They found that the original non-responders included an especially high proportion of abstainers (nondrinkers), which meant that their estimates of alcohol consumption based only on the original responders were too high.

Although there are methods for statistically correcting for non-response bias, they are based on assumptions about the non-responders—for example, that they are more similar to late responders than to early responders—which may not be correct. For this reason, the best approach to minimizing non-response bias is to minimize the number of non-responders—that is, to maximize the response rate. There is a large research literature on the factors that affect survey response rates (Groves et al., 2004). In general, in-person interviews have the highest response rates, followed by telephone surveys, and then mail and Internet surveys. Among the other factors that increase response rates are sending potential respondents a short pre-notification message informing them that they will be asked to participate in a survey in the near future and sending simple follow-up reminders to non-responders after a few weeks. The perceived length and complexity of the survey can also make a difference, which is why it is important to keep survey questionnaires as short, simple, and on topic as possible. Finally, offering an incentive—especially cash—is a reliable way to increase response rates. However, ethically, there are limits to offering incentives that may be so large as to be considered coercive.

Online Data Collection

An increasingly popular methodology for psychological research is based on using web sites that provide access to research participants as a service. One popular option has been Amazon's Mechanical Turk service, often referred to by the shortened **mTurk**. The mTurk service was not originally designed for human participants for research but has been applied to this purpose by many researchers who were able to carry out data collection with online methodologies.

Initially, several concerns were raised about online data collection related to data collection methodologies that did not rely on ever directly interacting with experiment participants. For example, if participants were effectively anonymous, how could we ensure that research was carried out ethically with respect to vulnerable populations such as children. Researchers need to trust and rely on the company running the online marketplace to be rigorous with respect to knowing their customers so that they can certify participants meet standard ethical practice. In the years that psychological science has been carried out with mTurk, no evidence of ethical compliance problems have yet been identified so this type of concern has abated somewhat.

There were also several preconceptions about samples recruited through the internet potentially not being representative of the broader population. Concern was initially raised that internet-based samples might not be demographically diverse. However, studies of online participant demographics have shown that these samples tend to be more diverse than traditional samples that have depended mainly on undergraduate students. There was an early stereotype that heavy internet users might be more likely to be maladjusted, socially isolated, or depressed, which has not been found to be true in practice. In general, there was also concern that internet-based findings might differ from those obtained with other methods but this has also not appeared to be the case whenever methodologies could be compared directly.

Some of the concerns initially raised reflected the fact that the first available

History of mTurk

The name Mechanical Turk refers to a 19th century fraud where a machine was purported to be a chess-playing automaton, an early robot. The device was shown for a fee and wealthy patrons could pay to play against it. In reality, the machine was simply a device that allowed a very short human hiding within the machine to manipulate pieces and play chess. The name reference was likely chosen by Amazon to reflect the fact that the original goal of the mTurk service was to hire humans inexpensively to do cognitive tasks that could not be completed accurately by artificial intelligence programs some years ago. For example, evaluating the accuracy of key words or search terms being related to online postings or determining if photos provided on a site selling cars were actually usable pictures of automobiles. Their model was to create a marketplace where requesters could post *human intelligence tasks* that could be completed by workers for pay. Since psychological research can easily be thought of as a human intelligence task, this online marketplace presented an interesting opportunity for psychological science for paradigms that could be carried out entirely online.

marketplace for recruiting participants online, mTurk, was not designed for the purpose of systematic, scientific data collection. One consequence of this is that it is possible to collect data entirely anonymously and not even be aware of the demographics of the participants sampled. While this seems to protect participant privacy very effectively, there is no way to know if there was some unexpected sampling bias in data collection that might be important for understanding the validity of conclusions.

Another newer marketplace company, **Prolific**, has recently become available for recruiting human participants for data collection. Their system was built explicitly around the idea of access to research participants. As a result, they can provide averaged demographic information for participant samples without the researcher needing to directly collect identifiable information together with research data.

Data collection online still faces the general issue that the context in which the research protocol is being carried out is under far less control than is possible in laboratory conditions. Participants might be engaged with the research task on mobile devices, in loud or distracting environments or not able to provide their full attention to the research task. A tendency to not comply with more complex tasks online might inadvertently embed a new kind of non-response bias. To date, the fact that online methodologies appear to result in similar patterns of data as in-person protocols suggests this is not a major problem for the kinds of tasks currently run online.

Deploying a protocol within an online environment can also place an implicit burden on the researchers to be familiar with tools for online protocol development. For many paradigms, the wide range of online survey tools makes this process very straightforward. For some forms of online studies where precise control is desired, researchers need to become familiar with online programming tools.

Overall, the ability to access a very large population quickly and easily online appears to provide an opportunity to accelerate research processes in psychological science. Participants can be recruited and complete protocols online at much more rapid rates than can be accomplished with in-person procedures. Further technological advances will likely continue to influence how these processes work in practice and how these affect the process of recruiting human participants into psychological studies. Attention will have to be paid to challenges that arise, such as the potential for more effective AI to simulate human behavior, while capitalizing on the potential to carry out effective science at a much greater pace.

WEIRD samples

As we noted in Chapter 8, psychological research is increasingly developing a sensitivity to the fact that a great many well-known findings about psychology depend largely or entirely on undergraduate participants who are at universities with substantial research programs. University students are already a very restricted demographic based on age and education. In addition, research universities tend to be ones that are more competitive with respect to admissions and therefore reflect populations that have succeeded in that competition. That may bias samples with respect to both individual difference variables and also factors like socio-economic status, which have substantial impacts on student preparation, success, and ability to attend competitive universities.

The acronym WEIRD, from Western Educated Industrialized Rich Democratic, has been used to describe the potential sampling issue involved in depending on undergraduate participants. Note that the Western and democratic elements of the acronym reflect the fact that to date, the overwhelming bulk of published psychological research that has been done with populations drawn from the United States, Canada and Western Europe. Acknowledgment of this issue has mainly been used to be more explicit about the demographic characteristics of the participants in research in publication. Some effort has also been made to increase outreach to broader and more diverse communities.

This is not a simple problem to solve as research at universities where research tends to depend on convenience samples of undergraduates. Because this kind of recruiting is far easier than investing time and energy into community outreach, the scientific research is therefore less costly to carry out. Making research more difficult or expensive will lead to less science being accomplished, which is not necessarily the goal of broadening our sampling procedures.

Online/internet based recruiting holds some promise for improving this, but will still tend to over-represent aspects of the WEIRD demographics.

Within research on internet use, there was documented a **digital divide** that reflected less access to the internet among poorer communities. This effect may be attenuating with greater accessibility through increasingly sophisticated mobile devices. It is still likely the case that online-based recruiting procedures are not reaching a fully diverse and representative population, although they are likely similar or better than standard WEIRD dependent approaches.

Considerations of the broader population the research sample is drawn from correctly brings attention to the question of: who is the conclusion drawn from research aimed at? This is essentially a question of **external validity**. Reliable data is known to reflect an effect seen in the participants sampled. The question is then how broad is the population this sample was drawn from to which the conclusion can be generalized?

Limitations in External Validity

Identifying that a research study is based on a WEIRD sample does not necessarily imply that the results are importantly limited by this fact. To identify a limitation, we need to be able to communicate an alternate hypothesis. Specifically, we should be able to identify a different sample that might plausibly not be expected to show the same behavior as the participants in a research study being reported. For example, in our Experiment 1 study examining the effect of encoding “depth” on word memory, the data were collected from undergraduates in a research methods class. We might note this and worry that it limits our conclusions. However, we would need to identify what different samples might not show the depth effect. There is no existing work that suggests that non-college-attending participants, or older/younger participants do not show the effect of memory enhancement following a study process that connects items to be remembered to existing semantic knowledge. Obviously, participant who cannot read (young children, illiterate) would probably not show the depth encoding effect for word list stimuli. We might also see this as a limitation of

the stimuli that could potentially be addressed in future studies.

In many cases, the limitation arising from sampling is completely clear. Research on stereotype bias based on race that are run in the USA are likely to show different patterns of behavior than bias studies carried out in other countries. There may be important commonalities that provide insight into human behavior for all humans on the planet, but the sample context is an important part of understanding the result of a research study run in one location. Psychological research based on attitudes, identity, or moral values are all examples of research that is very likely to be related to the population from which the participants were sampled.

Identifying external validity limitations is typically done by brainstorming as many conditions as you can think of where the effects of the study might not apply, then decide if any of these are important limitations to include in the discussion of your results. Hypotheses about effectiveness of limitations in generalizability usually must be done based on general knowledge of people's behavior. Our intuitions are often useful here, but expertise within the specific subdiscipline of the research is also very helpful. Obviously, the more experience you have in psychological research, the better your intuition about what sampling issues may be relevant.

External validity judgments can virtually never be made perfectly or with absolute confidence. They may look accurate, but then a new idea about differences across people advances our scientific understanding and modifies previous broad statements. Interval validity of studies, when established, rarely changes when new evidence about the phenomenon at study becomes available. However, the external validity of findings may change as science progresses and new factors and context elements are discovered in subsequent research. Often these advances further refine our understanding of the groups of people to whom the results apply, demonstrating the need to be complete and accurate about the samples participating in each research study.

Key Takeaways

- The method by which participants are recruited into the study to be part of the sample can affect the **external validity** and **generalizeability** of the scientific findings to the broader population.
- Participant samples drawn from undergraduate classes may over-represent WEIRD populations and reflect sampling bias.
- Methods to sample participants from broad populations may use either probability or non-probability sampling approaches.
- Population-based surveys and polling techniques can use complex, balanced **stratified sampling** approaches to avoid bias.
- **Non-response bias** reflects the possibility of a shift in the overall patterns of responses based on participants who elect not to participate due to the content of the research.

Exercises

TBA

12 Statistics 2: ANOVA

As we did in Chapter 5, here we will document practical steps required to carry out ANOVA, Analysis of Variance, analysis within R/Rstudio. We will review hands-on examples of three different analysis from three hypothetical experiments.

The first will demonstrate analysis across a single factor with three levels, a one-way ANOVA. This demonstrates the simple extension of the two independent samples t-test to experimental designs with three conditions instead of two.

The second will demonstrate analysis of a 2x2 factorial design with both factors having two levels between participants. This is the simplest factorial design. From the output of the ANOVA analysis, we will extract the key statistical parameters including the F-ratio, the degrees of freedom and the p-value. As with earlier t-test analysis, a simple reporting frame will be provided for reporting the results. However, it should be noted that the simple report of statistics from the output of an ANOVA is particularly uninformative without supporting statements about the descriptive statistics, statements of the direction of the results and ideally, a good data visualization.

In a third example, a mixed-model ANOVA will be demonstrated in which

there is one factor between participants and one factor within-participants. This changes the output information from the analysis as well as requires some reformatting of the input data files. Once the correct information is identified in the table, reporting and visualizing the results is a similar process to other ANOVA analysis.

In our return to hands-on statistical analysis, we will also review how reports of observed **effect sizes** are increasingly a part of modern statistical reporting in psychological science. Several different measures of corrected effect sizes are used to attempt to provide context for conditions where the independent variable has a small, medium or large effect on the dependent variable. These can be used to support the $p < .05$ formalism, but different effect size measures require familiarity with their underlying ranges.

At the end of this chapter, we will touch very briefly on the idea of **Bayesian analysis** as an alternate model for statistical inference. The Bayesian approach has aspects that are very intuitive and reflect a natural way to think about accumulating evidence for a hypothesis. However, the mathematics of employing a Bayesian approach require making assumptions about the experimental hypothesis that have proven difficult to accept broadly.

Learning objectives

- Carrying out an ANOVA in R/RStudio
- Reporting the ANOVA results in APA format, extracting key numbers from the output table
- Understanding how to read and how to make figures for factorial designs to illustrate main effects and interactions.
- Modern reproducibility theory: effect sizes
- Power analysis and sensitivity to observing reliable effects when planning research
- Bayesian analysis as an potential alternate approach to drawing inferences

In this chapter, we will present a series of analysis examples using R/RStudio and the function `ezANOVA` to carry out an ANOVA on simulated factorial data. The data files for these analyses should be available so that you can run these analyses in parallel to become familiar with the general process. The goal of these examples is to review how to extract the information to report from the output of the ANOVA calculation and how to format it for reporting in an APA scientific report.

This is the process we will use to analyze the data from the in-class Experiment 2. The results of this experiment will be reported in the second major writing assignment as an extension of the ideas from Experiment 1. You will also need to be able to carry out your ANOVA analysis for the in-class research projects, which are reported in the final term paper.

Example 1: One-way ANOVA in R

To test a Mozart Effect hypothesis, participants were assigned to listen to one of three kinds of audio while performing a spatial cognition test with 21 challenging problems. The audio sounds were either soothing Ocean noise, Folk dance music or Classical music. The number of problems solved was the dependent variable.

Simulated data are shown as the mean number of problems solved while the different sounds are playing. The standard deviations are shown under the means for each condition.

Music type	Problems solved
Ocean sounds	11.6 (2.72)
Folk music	13.1 (2.38)
Classical music	15.4 (2.27)

The analysis output is shown to the right as it would be printed in RStudio after running the `ezANOVA` command. The command parameters are included here for your reference. The key part of the output occurs after the `print(anova_result)` command, which reports the statistical output from the analysis. As written, the tabular format of the output is not completely clear.

As a table we can improve the formatting:

	Effect	DFn	DFd	F	p	p<.05	ges
2	Music	2	27	6.04215	.006782334	*	0.3091855

Now we can see the connection from the statistical information to the numbers. For the factor *Music*, the degrees of freedom in the numerator are 2 (DFn) and 27 in the denominator (DFd). The F-ratio value is 6.04. This

R Output

```
> anova_result = ezANOVA(
+   music
+   , dv = .(Problems.Solved)
+   , wid = .(N)
+   , within = NULL # NULL if no within factors
+   , between = .(Music) # NULL if no between factors
+   , observed = NULL
+   , diff = NULL
+   , reverse_diff = FALSE
+   , type = 3
+   , white.adjust = FALSE
+   , detailed = FALSE
+   , return_aov = FALSE # TRUE for showing details
+ )
Warning: Converting "N" to factor for ANOVA.
Warning: Converting "Music" to factor for ANOVA.
Coefficient covariances computed by hccm()
> print(anova_result)
$ANOVA
  Effect DFn DFd      F      p p<.05      ges
2  Music    2  27  6.04215 0.006782334     * 0.3091855

$`Levene's Test for Homogeneity of Variance`
  DFn DFd SSn SSd      F      p p<.05
1    2   27  0.2   65  0.04153846 0.9593736
```

would be written as $F(2,27) = 6.04$. The p-value is just as in our previous analysis and would be written rounded as, $p < .001$ or $p = .0068$ (one or the other, not both).

This is a reliable result where the different audio input types affected the score on the problem solving test. In the R output, the reliability of the results can be accidentally mis-read because of the two rightmost columns. The very rightmost column that is labeled **ges** is reporting a generalized eta-squared effect size to help characterize not just how reliable the effect is but how large it is. We will discuss measures of effect sizes at the end of this

chapter. It is slightly unfortunate that the ges measure is in the range from 0.0 to 1.0, so when there is very little effect of the IV, it can sometimes look initially like a p-value and mislead the reader into thinking an non-reliable effect is reliable. The second column from the right is only an asterisk when the p-value is less than .05 and is designed to help find reliable effects in much larger, more complex analysis with more factors and interactions. It will not usually be very helpful in our simpler designs.

Example 2: 2x2, Anagrams and Ink Color

In the example below, we have simulated data from a hypothetical experiment on stress and eating preferences. In this experiment, participants were given anagrams to solve which were either hard or easy. This difficulty factor was intended to create more stress for the harder problems. The problems were presented in either red or black ink under a theory that red ink presentation implicitly stresses participants more than traditional black ink. After several minutes of solving anagram puzzles, participants were offered candy and the number of pieces of chocolate taken was scored as the dependent variable. As an exercise, you might consider all the potentially questionable operational definitions in this study, but for our simulation we are concerned with interpreting the analysis.

R Output

```
> print(anova_result)
$ANOVA
      Effect DFn DFD          F          p p<.05      ges
2        Color    1   76 31.9657273 2.618672e-07     * 0.29607291
3  Difficulty    1   76  9.0441736 3.571545e-03     * 0.10634678
4 Color:Difficulty    1   76  0.3617669 5.493165e-01      0.00473754

$`Levene's Test for Homogeneity of Variance`
  DFn DFD      SSn      SSd          F          p p<.05
1     3   76 0.1375 17.25 0.2019324 0.8947498
```

The output of analysis using R/Rstudio is shown in the table above which just shows the **ANOVA table** output from the ezANOVA function (not the function call itself or the descriptive statistics). For this analysis, which is a 2x2 design, we have three main possible effects reported. These are the two main effects, of ink color and difficulty, and the interaction between these effects. The interaction term is listed in the row with the Effect, Color:Difficulty.

The first effect reported is the main effect of Color (line following "2"). The F column contains the F-ratio and the two columns to the left indicate the degrees of freedom in the numerator and denominator. This would be written as $F(1,76) = 32.0$. The p-values are all in scientific notation but we should be able to see that for the main effect of Color, this would be .00000026, which we can simply write as $p < .001$. The rightmost two columns are just the asterisk for a reliable result and the ges effect size report.

Similarly, the main effect of Difficulty was found to be reliable as well. Reading on line 3, we can find that $F(1,76) = 9.04$ and $p < .01$ (or $p = .0036$) for this effect.

However the interaction between the two factors is not reliable here. On line 4, the Color:Difficulty interaction produced an $F(1,76) = 0.36$ and translating the scientific notation for the p-value, we see that it is $p = 0.55$ which is greater than .05. Between the scientific notation and the very low ges score, it is possible to mis-read the output for a non-reliable effect like this, so care must be taken when understanding the analysis output.

We might also note at this point that we have no idea what the reliable effects in this study actually are. We have confirmation that the ink color affected the amount of chocolate eaten but ANOVA output itself provides no information about the direction of the effect. Obviously, we need to describe the direction of the effects in order to effectively communicate the results of this kind of analysis to a reader. To do this, we will need to look at the descriptive statistics

	Red Ink	Black Ink
Easy problems	3.45 (1.0)	4.75 (0.91)
Hard problems	2.95 (0.89)	4.00 (0.92)

Above is the means table for the average number of chocolate pieces taken after completing the stressful problem solving exercise. The numbers below the means in parentheses are the standard deviations. Remember to always check the descriptive statistics in both R and in Excel to be sure they are the same values. The output format from R will be somewhat harder to read quickly but may serve as an example of why the above format for means tables is preferred in order to quickly see the data pattern.

Note that in our simulated data, the group who completed the puzzles written in black ink are taking more chocolate pieces than the red ink condition. This was counter to our initial hypothesis. Nothing in the ANOVA report itself would have alerted us to this surprising finding. Careful review of the descriptive statistics is always necessary to accurately explain and interpret our experiment results.

General 2x2 ANOVA Heuristics

For a 2x2 design, the degrees of freedom in the numerator, the first number in parentheses after the F, will be 1 for all three contrasts, both main effects and the interaction term. For each factor, this value is the number of levels minus one, which is 1. For a design with both factors being between-participants, the degrees of freedom in the denominator is the total number of participants minus 4. You can think of this as starting with the sample size and reducing this by 1 for each of the three contrasts plus one more.

In the ANOVA report, there will be 3 lines reporting the reliability of effect results. The first two lines report the main effects, that is, the difference

Reporting the F-ratio

Different statistical programs may format the information describing the evaluation of the statistical analysis in different ways. They should all provide the same core information somewhere in the output. The main statistical parameter resulting from an ANOVA analysis is an F-ratio, typically written as F. The F statistic is reported with two degrees of freedom, for the numerator and the denominator, which are included in parentheses. First is the numerator df (DFn), which is related to the number of levels within the condition being reported on. The second df is the denominator (DFd), which is related to the number of participants in the study across all conditions. There will also be a report of the p value, which is the probability of the data occurring by chance under the null hypothesis.

In a written description of the results, the format follows the frame below for each of the main effects and interactions and all should be reported:

$$F(df_n, df_d) = X. xx, p < 0. yy$$

between levels of that factor ignoring the other factor. The third line is the interaction term, typically listed as something like *Factor1:Factor2* and will tell you whether there is a reliable influence across factors.

As with all other inferential statistics, we also obtain a p-value which means the probability of having observed the difference occurring in the data under the null hypothesis. We use the same standard criterion for this, $p < .05$.

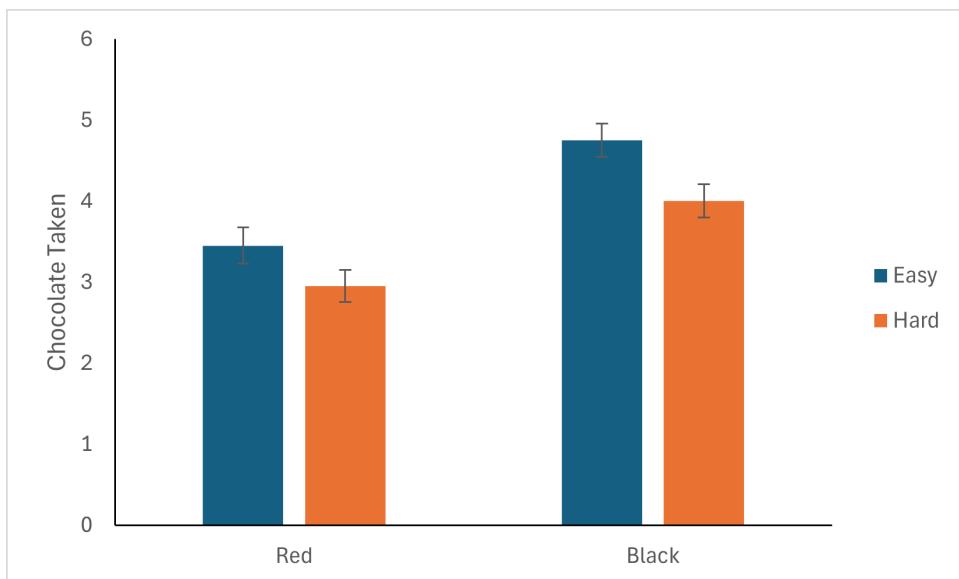
Importantly, the statistical reports of the main effect and the interaction term do not tell you anything about the direction of differences or what kind of interaction might have occurred. A reliable interaction could super-additive,

a 3:1 interaction (or sub-additive) or a cross-over interaction. Just looking at the ANOVA table cannot tell you which. It is necessary to review the descriptive statistics to understand the interaction so that you can report it accurately in the Results.

Making a 2x2 bar plot

To make a figure with 2x2 data in MS Excel, start by creating a labeled means table in a spreadsheet that contains the mean performance in each condition. It will look like the table above, but without the SD information. Start by selecting these 9 cells and Insert a 2-D Column chart.

With a little formatting it should look like this



The formatting applied here was to remove the Chart Title, add a vertical axis label, replace the Legend to the right side, change both axes to be black instead of gray and remove the y-axis gridlines. To add the standard error bars, it is necessary to prepare a separate 2x2 table of just the SE values for each of the cells in the design. Then select the Custom Error Bars option and select values across the row for each of the two series. That will get

accurate error bars for each of the four cells in the design, which each have a slightly different standard error. As a reminder, Google Sheets and the online versions of Excel do not currently have a method for individualized error bars across the conditions within a series of data. As a result, you should not use these programs because your error bars are inaccurate and it is very important not to present your data in a misleading or inaccurate way.

Example 3: 2x2 Mixed-model ANOVA

Consider the adage *the grass is always greener on the other side*. If we were to design an experiment to test whether this adage is true, we would need to come up with operational definitions of the metaphor that is based on viewing somebody else's situation more positively than one's own. For the purposes of this example, we might add an additional element that we hypothesize that this effect interacts with the personality variable optimism/pessimism such that the effect is much larger for optimists than pessimists.

For our hypothetical design, we will suppose that participants are given a description of a moderately lucky event, like winning money in a charity raffle, and asked to rate how happy they would be on a 1 to 7 scale. Participants will also be asked to rate how happy somebody else would be after the same event (order balanced, of course). This is a within-participants factor in this design since every participant answers twice, from the metaphor, once about the *other side* and once about their own side. In addition, we would use a personality scale to measure optimism and split our participants into two groups of 15, optimists and pessimists. As a participant variable, this is necessarily a between-participants factor.

For this 2x2 design, we have one between participants factor and a within-participants factor, which is referred to as a **mixed-model** ANOVA. We will see that with R/RStudio, the ANOVA results for this approach are presented in a very similar way with the only difference being slightly different df in the denominator.

However, this design approach requires some additional work with the spreadsheet tabulations of the data. In a typical data table, data are organized with one row per participant and all data collected from that participant listed across columns. For the within-participant variable, we would simply list the data as two columns. This is a useful format for reviewing data because it is easy to quickly compare scores across conditions within each participant. It is also relatively easy to calculate the descriptive statistics across conditions from this format.

However, the ANOVA analysis within R requires the data input to have a single variable per row and multiple rows for within-participant data. As a result, the sample data provided in the examples Excel (.xlsx) file has the same information organized differently than the file to be used as input for R (.csv). The need to re-organize the data is one of the many reasons why it is always

R Output

```
anova_result = ezANOVA(
+   greener
+   , dv = .(Green)
+   , wid = .(N)
+   , within = .(Side) # NULL if no within factors
+   , between = .(Personality) # NULL if no between factors
+   , observed = NULL
+   , diff = NULL
+   , reverse_diff = FALSE
+   , type = 3
+   , white.adjust = FALSE
+   , detailed = FALSE
+   , return_aov = FALSE # TRUE for showing details
+ )
Warning: Converting "N" to factor for ANOVA.
Warning: Converting "Side" to factor for ANOVA.
Warning: Converting "Personality" to factor for ANOVA.
> print(anova_result)
$ANOVA
  Effect DFn DFd          F          p p<.05      ges
2  Personality    1  28  5.316854 2.873922e-02      * 0.13530825
3            Side    1  28 53.200000 6.101031e-08      * 0.25052047
4 Personality:Side    1  28  7.221053 1.198682e-02      * 0.04340124
```

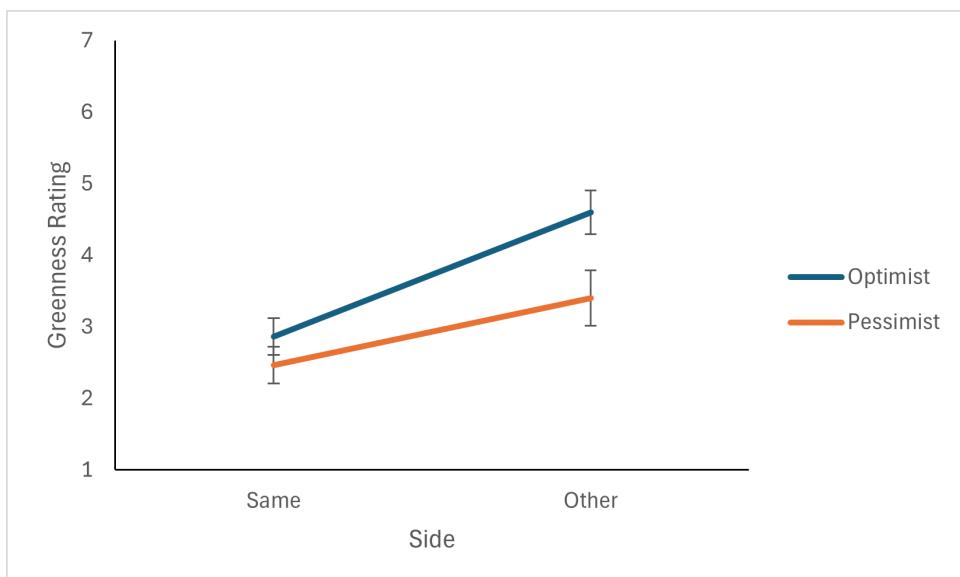
best to redundantly calculate sample descriptive statistics and verify the values across formats.

The output of the ANOVA looks very similar to the prior example. You might note that the df in the denominator (DFd) is the number of participants in the study, 30, minus 2 instead of 4 due to the within-participants factor of same/other side. The rest of the table is read the same way. There are reliable effects of Personality, Side and an interaction between these factors for the simulated data. To see what these effects are, we inspect the means table and a data visualization.

Means table:

	Same Side	Other Side
Optimists	2.87 (0.99)	4.6 (1.18)
Pessimists	2.47 (0.99)	3.4 (1.5)

Here we will use a line graph to visualize the effects the ANOVA indicated are reliable. As in the prior example, the graph is constructed from a means table in Excel with the four cell means used. In this case, we select a 2d Line graph as the starting template to make the following figure.



The additional formatting applied here to the standard Excel template was: remove the Chart Title, add vertical and horizontal axis labels, move the Legend to the right side, change both axes to be black instead of gray and remove the y-axis gridlines. Standard error bars are added to a line graph the same way they are added to a bar graph, using the Custom Error Bars option and selecting the SE values from a table of these in order to show the SE values correctly for each of the cells in the design.

The line graph allows us to see what the reliable effects reported in the ANOVA are. The Optimists are rating higher on the metaphorical greenness measure than the Pessimists, reflecting the main effect of the personality variable. Both groups are rating the Other Side higher than the Same Side, reflecting the main effect of the within-participants variable. These data would be consistent with the hypothesis implied by the adage that a good event happening to somebody else (the other side) would be perceived as having a bigger effect on their happiness. In addition, we have a reliable interaction between optimism and the side variable which the graph shows us is the hypothesized super-additive effect. Optimists see an even stronger effect of good things happening to somebody else than pessimists (in this made up data).

In the results section, these effects would be reported indicating the direction of the effect and supported by the statistical reports:

Optimists rated the benefit of a lucky event as more impactful than the pessimists, $F(1,28)=5.31, p<.05$. All participants rated the impact of the lucky event as being larger for somebody else compared with themselves, $F(1,28)=53.2, p<.0001$. There was a reliable interaction between the personality variable and the side factor reflecting the fact that the optimistic participants felt the lucky event would have an even larger effect on other people than themselves, $F(1,28)=7.22, p<.05$.

Modern Reproducibility Theory

You may have heard that psychology, as well as a variety of other scientific domains, is currently experiencing a *replicability crisis*. This has been inspired by a series of attempts to replicate well-known findings that have not produced reliable differences among conditions that were originally observed as reliable. There are substantial issues with the replication methodology that has been used that likely indicate that the term “crisis” is more extreme than warranted. However, the concern has usefully drawn attention to some aspects of how we carry out statistical inference in psychological science that we can use to improve our overall scientific progress.

The statistical model we have used so far reflects the approach used in the bulk of psychological research aimed at rejecting the null with a criterion of $p < .05$. As a reminder, this mathematically means that there is less than a 5% chance of the data appearing as observed if the null hypothesis were true and randomness produced the apparent effect. This leads to reporting results with a binary outcome: either the effect was reliable or not. There are several difficulties created by trying to make the outcome as simple as yes/no.

Marginal effects. It is not uncommon for research to be carefully carried out, analyzed properly and find that the probability of rejecting the null does not meet the .05 threshold but is instead in the range of .051 to 0.10. This poses some challenges for drawing interpretations of the results. We cannot claim that the results are reliable because they are not. However, the null hypothesis has actually been found to be somewhat improbable so simply saying that the effects are not reliable seems to miss an important aspect of the data. The simple binary model does not provide guidance for how to deal with these kinds of results.

Minuscule effects. It is also possible to have a statistically reliable effect that is actually extremely small. For example, if we found that an alternate studying method led to an reliable increase in memory performance of 1% accuracy, we would have a significant but somewhat uninteresting effect. This problem is fairly uncommon in experimental work as even small effects

can have theoretical implications, but comes up in more applied research or in some large-scale non-experimental studies. Here the simple binary model does not help us explain a reliable but not particularly useful effect.

Null findings. Sometimes our experimental hypothesis depends on providing evidence for a null effect. For example, we might want to show that sugar does not lead to hyperactivity in children. The simple binary model does not provide a method for evaluating this hypothesis since a *non-significant* finding could reflect a marginal effect or a true absence of an effect.

Effect sizes

Increasingly, the way researchers have sought to improve communication of results is to focus more on measures of the effect size. This changes our inference from "did the IV affect the DV?" into, "how much does the IV affect the DV?" In this approach, note that the null hypothesis is now the same as saying the effect size equals zero. Whenever we carry out an analysis, we are estimating the effect size based on our sample, which is a subgroup from a broader population. Unless we measure the entire population, we can never assert that the effect size is exactly one specific value. This is the difficulty of arguing for the null hypothesis. Our estimates can provide evidence that the effect size is not very different from zero, but not that it is exactly zero. When we fail to reject the null, we can only say that we are not sure that our current effect size estimate is different from zero.

As we reviewed earlier, an unstandardized effect size is simply the difference in the DV between conditions of interest across the average (mean) scores. In some cases, this can help communicate the results of an experiment, but it has the weakness of not incorporating any information about the variability of performance that was observed. Standardized effect sizes all incorporate a measure of variance to rescale the difference in means with the intention of providing a common scale for denoting effects across a scale something like 'small,' 'medium,' and 'large.' Unfortunately, the field of psychological science has not yet converged on a standard methodology analogous to the reporting

of p-values. Instead, there are several different forms of standardized effect sizes that are used depending on methodology and analysis type. Here we will briefly review two of these.

One common standard effect size measure is **Cohen's d**, which is often reported with t-tests to help communicate the findings. It is calculated as a ratio of the mean difference to variance and produces a number than can be used to scale the effect size into categories: small, medium, large, very large. Large effect sizes intuitively reflect factors that are particularly important to understand in their relationship to the dependent variable measure.

Another common effect size measure that is reported in the ANOVA results above is **generalized eta-squared** or η^2 in the column titled **ges** in the ezANOVA output in R. This can be treated as an effect magnitude estimate like Cohen's d, but the scale is different. In the table below, values for both of these effect size measures are shown for the common effect size descriptive categories.

Table of effect size ranges

Effect Magnitude	Cohen's d	Generalized eta-squared
Small	0.2	0.02
Medium	0.5	0.13
Large	0.8	0.26
Very Large	1.2	0.40

While this effect size approach improves on the simple binary categorization based on whether p is less than .05 or not, the effect size statistics require becoming familiar with their relative scale values. It is also obviously very important to know what effect size measure is being provided by the analysis function. The ges values in the analyses reported above are generally very robust, many being medium or large, but if one accidentally compared them to Cohen's d effect sizes, they could be mistaken for small effects.

One of the advantages of the effect size approach is to identify reliable but small effects. Small effect sizes can be reliable but reflect factors that do not have a large effect on the dependent variable. In the third example above, the interaction between optimism and side is a small effect. This could help us accurately communicate the results that the more substantial effects were due to the main effects and while there is a reliable interaction, it has less impact on the scores.

Effect sizes are also very helpful for planning research and understanding conditions where effects are found to not be reliable. In both effect size types, the null hypothesis that the IV does not influence the DV is the same as indicating that the effect size is zero. While our statistical models do not provide a method for establishing confidence in a null finding, consistently observing effect sizes around zero would be a method for eventually supporting a conclusion that an IV has reliably no effect on the DV.

For planning research, if we have an estimate of the effect size we can use that to help plan the sample size for our study. If we think the effect size may be small, we know that we will likely need a large sample and very rigorous procedure to avoid a Type 2 error. When the effect size is expected to be large, smaller sample sizes are likely to be enough to observe a reliable effect. The process of planning the sample size from the effect size is carrying out a **power analysis**.

Power Analysis and Sensitivity in design

When planning a research study, particularly a rigorous Randomized Clinical Trial (RCT), it is important to be able to specify in advance exactly how many participants are expected to be in the research study. This is done by carrying out a power analysis, which is based on an a priori estimate of the effect size to be observed in the study. A power analysis takes a standardized effect size and with a specific number of participants expected to be recruited, provides a probability estimate of the chance of obtaining a reliable statistical difference between conditions. The math of carrying out this analysis is

beyond our scope here, but the underlying idea is that even where there is a real, true difference between conditions, data can still be variable enough that our statistics do not work (we fail to reject the null, a Type 2 error). In many formal research proposals, studies are designed around a power analysis based estimate of 80% or 90% likelihood of success.

In many experimental research studies, the researchers do not start with a strongly held numerical estimate of the expected effect size. In this case, it is impossible to carry out a formal power analysis before starting research. However, if the data indicate no reliable statistical differences, it may lead the researchers to consider that their design lacks sensitivity to the observed effect size. That is, the effect size is smaller than could be detected with the sample size available. This is often the case in results termed “marginal” above. The best practice in this case is to estimate the effect size from the “failed” study and use this to design a better follow-up study with larger n and/or a more powerful manipulation.

A consideration of power and sensitivity points out the difficulty of interpreting findings that “fail to replicate” prior studies that have been commonly reported as inspiring a “replicability crisis.” We should actually expect studies to fail to replicate some of the time, even with real effect sizes when the effect is subtle, as many interesting effects are. Power analysis with effect sizes in the ‘small’ range can indicate that it may take several hundred participants to have a high probability of obtaining a reliable effect. There are certainly publications that have found reliable effects with smaller sample sizes, suggesting the researchers may have been lucky. We will consider the implications of this later in Chapter 19 (Responsible Conduct of Research).

Bayesian analysis

An entirely alternate approach to statistical inference exists based on Bayesian analysis. This approach focuses on the probability that the experimental hypothesis is correct and how this is influenced and updated as data becomes available. The probability of truth of the hypothesis acts like a quantitative effect size measure and follows a very robust mathematical tradition. The core element of this approach is to start with an estimate of the probability that your hypothesis is correct before you begin your research study. This number is referred to shorthand as the experimental **prior**, or **prior odds**.

After a study has been completed, if the data are consistent with the hypothesis, we can say that the probability that the hypothesis is true has increased. The data from the experiment has made us more confident in our hypothesis. The probability that the hypothesis is true including the experimental data is the **posterior probability** or **posterior odds**.

The Bayesian model is very intuitive because it reflects the way scientists think as research is carried out. We generally start planning an experiment with some confidence than the hypothesis is true and then over a series of studies, this confidence increases with additional consistent data. It also provides a mathematical approach to gaining confidence in the null hypothesis when it may be true.

Unfortunately, the mathematical basis of this approach has a major limitation in that calculating the posterior probability depends very heavily on the specific prior probability. That means if two researchers have different priors, for example one of them does not believe the hypothesis, they evaluate the statistics of the study completely differently. Since researchers often do not agree, it is very difficult to objectively quantify the effect of data on everybody's beliefs. As a result, this approach has not replaced our more traditional statistical models in spite of its benefits.

Key Takeaways

- Learn how to carry out an ANOVA analysis of factorial data in R
- Understand how to read the statistical output of this analysis and translate the result into the format you would use in the Results section of a scientific report
- Learn how to make a figure to visualize the results of a 2x2 ANOVA, both as a bar plot and a line graph so that you can choose the most effective presentation for your data
- Understand **effect sizes** in results reporting and how to use these to interpret large or small overall effects
- Understand the meaning of **marginal effects**, which do not meet the reporting requirements for reliability but do not indicate that the IV has no effect on the DV.
- Understand how a **power analysis** is derived from estimates of effect size to help plan sample sizes for proposed research.

Exercises

Analyze the data from Experiment 2. Start by review the data in Excel using the provided .xlsx file.

Calculate descriptive statistics as we did with our earlier analysis of Experiment 1. Note that there are now four conditions to calculate condition means, SD and SE. You should also examine marginal means, where two conditions are combined, so that you can observe the magnitude of the main effects.

To run the analysis, use the RStudio program to start an analysis session. Open the provided Exp2 ANOVA.R file. Set your *working directory* to the location of the source file (if that is where your .csv data file is located). Step through the commands that load the ezANOVA and related packages.

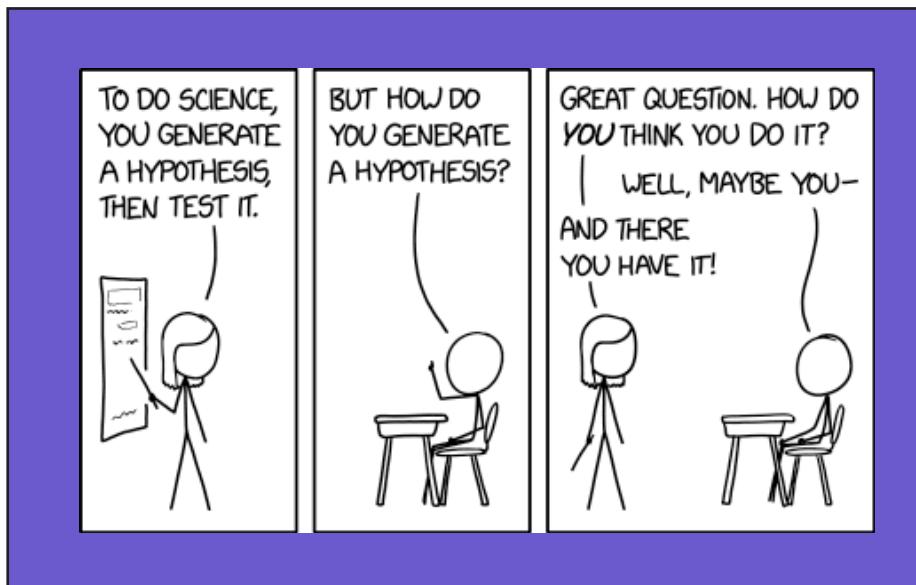
Load the datafile, run the describeBy function and verify that the output matches the descriptive statistics you calculated in Excel.

Run the ezANOVA command. Review the output, locate the key statistical parameters, the F-ratio's, the degrees of freedom and p-values. Prepare your report of the analysis, remembering to include directional statements about the main effects and a thorough description of any observed interaction.

Save your work for when you prepare the next experimental report write-up.

13 Proposing Research

For our hands-on approach to learning experimental research methods, an important part of the learning experience is to develop, propose, carry out and write up a small project using experimental methods. Given the constraints of class, this necessarily is a very constrained process with limited scope. These projects can be carried out with generally good adherence to experimental design principles but cannot be considered formal, publishable research. Should a class project identify something that looks like a novel and interesting research finding, it is recommended that interested students seek out a laboratory group working in that area.



Learning Objectives

- Understand how to prepare and write a research proposal
- Explaining the experimental hypothesis related to prior research
- Describing the novel procedure including details of the independent and dependent variables
- Communicate what is novel about the new study, how it adds to or goes beyond the findings of previous research
- Prepare a protocol description at the level of detail that would be required for IRB review

Developing a research hypothesis: intuition

It's very common to start the research process with a statement that begins "I wonder what would happen if..." These ideas are usually more exploratory in nature and not initially expressed in the form of a testable hypothesis.

You can turn these ideas into a testable hypothesis by not stopping with the exploratory statement. Make a prediction about what would happen and then argue for your prediction. Now you have a testable hypothesis.

Keep in mind that your hypothesis may very well be wrong at this point. One reason why it is often more comfortable at this stage of idea development is that you might not be very confident that your hypothesis is correct. A highly effective hypothesis for driving a research idea needs to be specific and testable, but it does not have to be correct. Often the most interesting research ideas are where the outcome is difficult to predict, or if there are two different theoretical approaches that make different predictions. A very useful heuristic for research proposals: it is far better to be wrong than to be vague. Make a specific, testable prediction and motivate your rationale. Even if you

are wrong about the outcome, the study will likely be well-designed.

Preliminary background research

The original idea can come from background research, or it might come from a textbook or even more purely from intuition. Developing the idea requires examining the relevant published research around the question in the subdomain of interest. Google Scholar is the currently best recommended search engine for scientific work. It works through a familiar search term interface and covers all available published research across psychology and broader scientific fields. The amount of research included is vast enough that it is important to identify the correct key terminology, the right search terms, to find the related published work.

Many subfields of psychology have specific technical terms or jargon that are precisely defined and identifying these is very important to being able to do thorough background research. Many areas of science have jargon as

Literature Searching with Google Scholar

<https://scholar.google.com/>

This will provide lists of articles matching search terms. Clicking the article title will take you to the journal where the article was published. However, you may only have access to the title and abstract for that publication without paying a large fee.

Your university likely has a license to access the full text of most published work in reputable peer-reviewed journals. If you connect through your university, either on the campus internet or via a VPN connection, you should see another option to access the publication. At Northwestern, the link is marked “Find@NU” on the right side of the page. This link will take you past the paywall to give you direct access to the full content of the vast majority of published scientific work.

shorthand, but this is a particular feature of psychological science because we are often studying concepts that have familiar names. For example, a concept like “depression” is one that most people have an idea of but not everybody’s idea may actually be the same. Research in this area will tend to cluster around the term “mood disorder” which is a less common but more precisely defined construct. Similarly, the idea of “altruism” is often done as part of research on “pro-social behavior.” Getting started on background research is often a matter of first learning the best keywords for searching.

Once you have found the terms that lead you to the prior work in the area, you can start reviewing the methodologies that are commonly used. As an aside, if you cannot find prior published work, it is much more likely that you have not found the right search terms than nobody has ever thought to do research in this area before. Most research areas will have reported findings using a variety of methodological approaches. Many of these approaches may not be suitable for your proposal. They might require expertise in a complex or demanding technique requiring specific training. They might require

Open Access Science

There is a movement in the scientific community, often tied in with efforts to improve the reliability and reproducibilty of research, to make more scientific reports more generally accessible. When publishing a research paper, the authors often have an option to designate their report as *Open Access*, which means the publication will never be placed behind a paywall and difficult to reach for non-university researchers. However, many journals currently charge large fees at publication to authors who wish to have this designation and not all research is carried out with funds set aside to cover these costs. As a result, you may only occasionally find recent interesting research to be Open Access and easier to reach. You might note that when you do, the authors have made an explicit effort to make their work available to you.

access to a special population, or a very large number of participants to be adequately powered.

When a tractable methodological approach has been identified, the next step is to incorporate the original ideas behind an experimental hypothesis into that methodology. That requires constructing the operational definitions of the key constructs, identifying what can be manipulated and how and what the key dependent measure will be.

Operational Definitions

As noted previously in our approach to basic experimental design, we first need to come up with operational definitions of the constructs embedded in our hypothesis. Existing published research is the best place to start with ways to implement complex psychological ideas with tractable methodology. Examine the methodologies used in published work and consider how effective they are with respect to face validity, that is, how obviously the capture the idea. Where they seem imperfect, there may be necessary compromises made to make experimentation possible. Or it can be the case that the idea is so complex that there are many different ways to reconceptualize the idea for a research paradigm. Making adjustments to the methodology can improve the design, especially if the published work might have been constrained by technology of the time in which the research was accomplished.

In general, the process of setting up the operational definitions is the same as described earlier. Identify the key independent variable(s) and the level across each that can be controlled (for experimental designs) or measured (for non-experimental factors such as participant variables). The dependent variable needs to be a measured operational definition that can be collected and exhibits a roughly normal (Gaussian) distribution. The main issues to assess for the dependent variable are that participant scores will not tend to cluster at floor, the lowest possible value, or ceiling (perfect performance).

Operational definitions will often require defining the stimuli that will be used in the research. Any surveys to be used for measurement or words, images, pictures to be shown to the participants should be characterized. The published work may provide exact examples of the stimuli and instruments used or may give a broad overall description. Many published studies are accompanied on the journal's website by Supplementary Materials that may contain the exact stimuli or questionnaires used in the research. In other cases, it is necessary to go and obtain exact stimuli to be used in the research protocol. This should be done early in the research development process to be sure that the stimuli are obtainable and any surveys that are planned to be used are available. Some research depends on research instruments that are held under copyright and may not be openly available to other researchers. In some cases, authors indicate that the stimuli used are "available upon request" but are not as responsive as would be desired. Before committing to the research plan, the availability of the key research elements needs to be assured. In addition, evaluation of the specific operational definitions used helps guide the analysis of possible extraneous variables to consider.

Extraneous Variables

For planning the experimental procedure, it is necessary to identify as many possible extraneous variables as possible. There is no guaranteed approach to figure out all of these in advance, unfortunately. Looking at the detailed procedure from prior published work will provide a lot of insight into known factors that influence the dependent variable. General knowledge of the research area is the other main source of ideas. Increasing your background knowledge of the theoretical ideas in the area through additional research and reading the published literature is a great benefit.

Once the known extraneous variables are identified, the tools to manage these are the same as we have seen before: constancy, counterbalancing, and random assignment to conditions. Across the manipulated levels of your design, keep as much constant as possible. Anything you cannot

keep constant, counterbalance across groups to keep this variable from confounding your research. This process will give you a detailed structure for your research protocol.

For data collection to be carried out in-person with the participants, it is often a good idea to fully script out the research procedure. This helps maintain constancy across multiple participants and especially when research is done by a collaborative team of experimenters who should all try to administer the task exactly the same way.

Data collection that will be carried out online is often very efficient. It relieves the burden of scheduling meetings with individual participants and reduces the need to carefully script participant interactions. However, it does require attention to the method by which the procedure is implemented online. Survey systems such as Qualtrics are very popular for online studies. Some time and effort will need to be invested in learning how to use the system and how to configure the presentation of stimuli or survey questions.

Recruiting Plan

Once the procedure is known, the next step is to develop a recruiting plan. The two key questions to answer are (1) how many participants will be included in the study? and (2) how will these participants be found? Since all research participation is voluntary, the plan involves outreach to the population of interest with the opportunity to participate. If a specific subpopulation is the focus of research, a plan for finding and recruiting participants is necessary. The number of participants can be technically accurately estimated via the use of a power analysis (from Chapter 12). In many research projects based on convenience sampling, the main constraint is how many people can be recruited making the answer to this question "as many as you can." A good rule of thumb is 15-20 participants per manipulated condition, i.e., 40 for a two-group design and 80 for a fully between-participants 2x2 factorial design.

In formal research, an important aspect of the recruiting plan is developing a fair compensation plan for participants who volunteer. In some cases, this is based on class credit and therefore the experimental protocol is generally highly constrained in length (e.g., 30 m or 1 hr). The length of time needed to carry out the experimental protocol is important for this step as both financial compensation and credit are generally scaled on an hourly basis.

Analysis Plan

Best practice for experimental design is to have a formally written analysis plan for the DV as a function of the IV's before starting data collection. This can be as simple as noting that the analysis will depend on independent samples t-tests or a factorial ANOVA. It can also require more complex analysis approaches planned in advance. However, in a lot of research cases where a novel set of ideas are being tested against each other, unexpected findings will inspire additional analytic ideas in the course of the research process.

As a rule of thumb, if the analysis plan is significantly different than originally planned, the research should most likely be further explored with additional studies. Those studies can be planned with a more accurate understanding of the analytic needs. Using very creative and flexible analytic strategies runs the risk of research being bias by *p-hacking* as will be discussed in Chapter 19 (Responsible Conduct of Research).

IRB approval

Once the entire research plan is complete, the protocol is submitted for review to the Institutional Review Board for approval and/or revision. No systematic data collection from human participants intended for broad distribution should ever be carried out without review. Classroom research by not being intended for broad distribution is typically seen as not under the

purview of the IRB. However, it is still important that class projects be carried out under the general principles of ethical research: Respect for Persons, Beneficence and Justice.

Participants should be informed that they are participating in a research study and indicate that they agree to this of their own choice. This can be done by including that information on paper for in-person data collection. For online data collection, the first element should be a notification that they are participating in a research study, what is expected of them and that they can decline to participate. Continuing with the protocol from that point is consenting to participate.

Practical Guidelines for Class Research

The most important first step for planning a psychological science project that can be completed in a classroom is to find a published report in a peer-reviewed journal to work from. You may start from intuition, interesting results you have seen in other classes or elsewhere, but it is extremely valuable to have a closely related publication for reference. The reason for this is that the operational definition process in psychological science is often a lengthy one with false starts, mistakes and gradual improvements. Most published research implicitly relies on a series of pilot studies that guided the design through a variety of pitfalls. In a new subdomain, the first paper could easily reflect several years of preliminary research developing the methodological tools to test the hypothesis. Those often do not get included in the final publication – making science often look a lot easier than it is – but for classwork there is not time to do this methodological exploration. A published report will contain information on a set of definitions that worked, which is a good place to start.

As noted above, Google Scholar is the tool to use to find this first background publication. Be aware that it may take some exploration to identify the key technical terms used in your area of interest to find the published work. Also

be aware that Google Scholar indexes outside of psychology. Pay attention to the journal the work is published in to identify if your search has drifted into related areas that are more physiological in nature (e.g., neuroscience, health) that may be impractical for class. Try to verify that the journal is peer-reviewed if the name is unfamiliar by checking if the publication is cited in recognizable outlets (use the Cited By link) and avoid publications with "Proceedings" in the name as these are conference proceedings which may not robust findings.

Once you have the first paper, you should look for something new to add to their approach. Even for class projects, we should approach research with the idea of extending findings to something novel and not just simply replicating a famous finding. The new idea to add can come from intuition, from the authors discussion of future research in their Discussion section, or from another related publication in the field. Blending two papers together often works well to create a 2x2 design from two publications that each had contrasts between two groups. Note that even if the two published papers used more complex designs, you may be able to take their main effect findings as evidence that a two-group study would work and use this as a factor in your design. Check the interaction terms in their work, of course, to ensure that these are not indicating critical extraneous variables that you need to plan for.

For classroom work, you will prepare a 2x2 design with at least one manipulated variable. If you are combining published papers, you may come up with a design plan that is more complex. If you find that the design that best captures the previous work is a 3x2 or a 2x2x2 design, you will want to simplify down to 2x2 even if it weakens the scientific impact of your potential findings. Anything more complex than a 2x2 adds too much difficulty to be plausibly carried out in a classroom context. They require too much data, extending the time needed to recruit and test participants. The analysis is also necessarily much more complex and will significantly slow both the analysis and interpretation of the data when writing up the results later.

As discussed previously, there are a variety of ways to design a factorial study

with 2x2 complexity. In general, for the manipulated variable, it is best to try to follow a published successful study as much as possible. The second factor can be a participant variable that is measured or recruited for instead of manipulated. However, avoid the temptation to lazily use men versus women as the second variable. This is an area where intuitions are often not at all grounded in a theory that can be articulated to motivate the study. To make the case that this is an important question to ask in your study, you must find research that shows your manipulated variable is explicitly affected by gender. Even so, be aware that modern understanding of gender does not reduce this variable to a simple choice of two options which will make this factor not suitable for a 2x2 unless you restrict recruiting.

With good sources, most of your work establishing the operational definitions can be taken from those publications. Use existing surveys, stimuli, or other materials from those papers as much as possible. If you need to create something new, keep it as simple as possible and maximize face validity, e.g., 1-10 scales asking participants to subjectively rate their current state.

Once you have the basic design and materials, you need a plan for carrying out the procedure. It is very popular to collect data using online tools such as Qualtrics. Many aspects of experimental control can be implemented within these robust systems. Simpler systems such as Google Forms may also work. Be careful of fees associated with systems not affiliated with the university. Systems with university site licenses often provide access to a great deal of technical documentation to help set up the design and will have local experts to answer questions (e.g., Northwestern University has a site license with Qualtrics and it is very effective for this purpose).

If you are not doing data collection online, write out a script for how participants will carry out the design procedures. The script helps maintain consistency in interactions with participants through the 40-80 repetitions of the process needed to accumulate the data. It also helps maintain consistency across a collaborative group where 4 people might each be responsible for portions of the data collection.

The recruiting plan should also be specified in advance as part of the research proposal. It may be as simple as social media posts or emails to a locally available convenience sample. If your research plans to recruit from specific populations such as athletes or engineers, be sure to plan how that group will be reached.

Once all the pieces are in place, the entire research protocol is written and provided to class instructors for review. This must include all stimuli that will be used in the planned research. That is, you should not at this point say, "we will collect images of famous celebrities from the internet." You should collect the images you will use and include those in your protocol submission.

Given time constraints, there is generally not time for a formal IRB review of these research plans and the instructor and teaching assistant will act as an informal IRB. As a result, all research should be absolutely minimal risk. All aspects of deception or any issues with privacy should be minimized or eliminated as much as possible. This may render some very interesting and motivated scientific research unable to be carried out in the classroom environment, but this should not be surprising given how important adherence to ethical research is in science.

Data collection can not start until the entire protocol is reviewed and explicitly approved by classroom instructors. This is necessary for ethical research but has the risk of delaying projects and placing classroom researchers under severe time constraints. Prepare your proposal early and expect feedback about adjustments and revisions to your plan. Make those and resubmit the proposal as quickly as possible. Data collection can take significant time and there is a lot of work still to do after collecting data. The results need to be organized, analyzed and then the writeup of the results needs to be prepared. It is very ambitious to try to carry out an independent project in the scope of a month. It is possible but requires good time management throughout the process.

Grant proposals to funding agencies

The process of preparing a research proposal bears some resemblance to the process of writing grant applications that is an important part of the operation of major research laboratories. This process is somewhat more focused on obtaining funds to support these research projects. Many of the staff in most large research labs are not supported by the institution or university housing the lab but are entirely paid through outside funding to the lab. Research funds also support more expensive methodologies and participant compensation to carry out a series of studies organized around a core theoretical framework.

These proposals often look like research papers to some degree, although written in future tense rather than past tense. They will typically include a fair amount of "preliminary data" that has already been collected but not yet published that indicates that the research plan is feasible. The research plan will detail a series of experiments over a time frame that can vary from less than a year up to five years. These proposals have three major components: the collaborative team (led by a Principal Investigator, PI), a budget (cost/year) and a specific scientific research plan. The format of these sections varies very widely across funding agencies. Research staff supporting grant applications spend a lot of time reading detailed formatting requirements and necessary levels of detailed information. The research plan is generally reviewed by a committee of scientific peers in a competitive fashion. Grants are reviewed on an annual cycle and depending on availability of funds to the funding agency only the top 5%-15% of proposals may be awarded funding.

Ideas for research proposals to granting agencies virtually never start with intuitive ideas and background research. Agencies tend to award grants to established experts in a field, so most grants build on the prior work of the collaborative team and PI. This does have some known issues in potentially creating a barrier to entry for researchers to become established or to move into a new area. At the same time, much of the money available for research funding comes from governmental sources which have a requirement to

obtain some value from those funds. It is very hard to tell in proposal review which projects are going to have the largest scientific impact. Practically speaking, experts with robust track records in an area are most likely to produce scientific advances.

Within the USA, two major institutions that fund psychological science are the National Science Foundation (NSF) and the National Institute of Health (NIH). Within NSF, most psychological research is in the broad category of Social, Behavioral, Economic Sciences (SBE) which is then further subdivided into Behavioral and Cognitive Sciences and Social and Economic Sciences. The NIH is much larger in size and budget than NSF and houses 21 divisions across a very wide range of health-related research areas. Examples of programs that fund psychological science research include National Institute of Mental Health (NIMH), National Eye Institute (NEI), National Institute of Child Health and Human Development (NICHD), National Institute of Aging (NIA), National Institute of Deafness and Other Communication Disorders (NICHD), National Institute of Neurological Disorders and Stroke (NINDS).

There are also research projects funded through scientific divisions within the Department of Defense (DoD). These include a collection of laboratories such as the Air Force Research Laboratory (AFRL) and Army Research Laboratories (ARL). The Office of Naval Research (ONR) acts as a funding agency similar to NSF but with research aimed at application at military personnel. Most DoD research is aimed at more immediate application of findings rather than long-term scientific understanding. However, it should be noted that these projects can be aimed at psychological questions across the large range of both active and retired (veteran) military personnel, making this sample fairly similar to the overall population. There are also specialty agencies within the DoD such as the Defense Advanced Research Projects Administration (DARPA) which fund very basic science aimed at extremely novel ideas (which has, unfortunately, led historically to support of ideas with little credible scientific support).

There are also private foundations that support psychological science that often have specific areas of interest. Many of these foundations approach

scientific support with the same goal of highly rigorous, robust and internally valid research. However, there are some foundations that look for work that advances an agenda regardless of the robustness of science. Most universities or large research institutions have a Development office that provides guidance on private funding sources that support high quality psychological and other science.

Most of this information is not immediately relevant to undergraduate researchers but if you have the opportunity to work in a university laboratory, you may encounter some work aimed at seeking external funding. Some universities have some internal funds set aside to support undergraduate research and if you have the opportunity to apply for these, you will find yourself working through the same process as the lab PI. For example, Northwestern University has undergraduate research funding available for projects done over the summer as well as during the academic year. These can be a great opportunity to do formal, high quality research within a professional laboratory context.

Key Takeaways

- Preparing a research proposal is similar to writing a research report, only in the future tense.
- Providing a planned research protocol in enough detail for IRB evaluation includes at least as much detail as the Methods section of a report, usually also including all the stimuli to be used in the study.
- Recruiting and sample size planning are done with both experimental rigor and ethical considerations in mind.
- New research builds on prior research for robustness and guidance in design and tools for experimental control.

Exercises

Prepare a research proposal outline for a project to be carried out as a short class project for a final paper.

The outline should contain all of the following information:

- Name of the researchers carrying out the project, including all group members
- Tentative project title
- Identify a first main background source and provide the APA-style reference to this peer-reviewed, published research. This source experiment will provide some theoretical background and starting ideas for the operational definitions.
- Describe the design of the main inspirational experiment in this paper including the IV(s), the DV, the number of participants and the outcome.
- For the proposed research, clearly indicate what new element you are planning to add to this design to expand on this published work. Describe your experimental hypotheses driving your proposed study.
- Diagram your 2x2 design, describing both factors and both levels of each factor. Identify how many participants you think you will need to test your hypothesis.

14 Non-experimental Design

Researchers who are simply interested in describing characteristics of people, describing relationships between variables, and using those relationships to make predictions can use non-experimental research. Using the non-experimental approach, the researcher simply measures variables as they naturally occur, but they do not manipulate them. For instance, research analyzing the number of traffic fatalities in America last year that involved the use of a cell phone would be of substantial importance. It would not be done by actually manipulating cell phone, but by collection of available data on accidents and would be categorized as non-experimental research. Alternatively, if a researcher stood at a busy intersection and recorded drivers' genders and whether or not they were using a cell phone when they passed through the intersection to see whether men or women are more likely to use a cell phone when driving, then this would also be non-experimental research. It is important to point out that non-experimental does not mean nonscientific. Non-experimental research is scientific in nature. It can be used to fulfill two of the three goals of science (to describe and to predict). However, unlike with experimental research, we cannot easily make causal conclusions using this method; we generally cannot confidently say that one variable causes another variable using this method.

In spite of the difficulty with causal inference, there are a wide range of

important questions in psychological science and related fields that depend on data about people that can only be addressed with non-experimental methods. One such area is questions about behavior that affect public health, such as during the COVID pandemic. Wallace, Goldsmith-Pinkham and Schwartz (2023) addressed the question of whether political party affiliation was a risk factor associated with COVID mortality during the pandemic. Using public health data from Florida and Ohio, they found that after vaccines became available, excess mortality was significantly higher for Republican voters than Democratic voters in these states. Although it is tempting to infer a causal relationship about political or personal beliefs that influence party membership, vaccine use and mortality outcome, even with a great deal of careful quantitative effort to measure and control for other extraneous variables, the authors final conclusions are stated cautiously:

"In this cross-sectional study, an association was observed between political party affiliation and excess deaths in Ohio and Florida after COVID-19 vaccines were available to all adults. These findings suggest that differences in vaccination attitudes and reported uptake between Republican and Democratic voters may have been factors in the severity and trajectory of the pandemic in the US."

We should note the careful use of reporting an association and "may have been factors" reflecting the challenges inherent to this approach. The research question is one of substantial practical importance and understanding the relationship requires consideration of the role of behavior, attitudes and beliefs on public health, which are questions related to psychological constructs.

Findings from large scale surveys are frequently encountered in public scientific reporting and these highlight a set of scientific disciplines related to psychological science. Much economic reporting data, such as the National Household Income and Expenditure Survey, is carried out with the same methods and analytical tools. We might more easily identify research like the National Survey of Mental Health and Well Being as being a psychological study but many fields depend on data collected from people, human research

participants, to assess scientific questions. Research in public health, sociology and educational research will have similar kinds of approaches.

In this chapter, we will generally focus on how to identify non-experimental methods and the inference challenges rather than the quantitative tools for tackling these questions. The analytical tools for large scale, multivariate data collection are beyond the scope of this text and class.

What do the following classic studies have in common?

- Stanley Milgram found that about two thirds of his research participants were willing to administer dangerous shocks to another person just because they were told to by an authority figure (Milgram, 1963).
- Elizabeth Loftus and Jacqueline Pickrell showed that it is relatively easy to “implant” false memories in people by repeatedly asking them about childhood events that did not actually happen to them (Loftus Pickrell, 1995).
- John Cacioppo and Richard Petty evaluated the validity of their Need for Cognition Scale—a measure of the extent to which people like and value thinking—by comparing the scores of university professors with those of factory workers (Cacioppo Petty, 1982).
- David Rosenhan found that confederates who went to psychiatric hospitals claiming to have heard voices saying things like “empty” and “thud” were labeled as schizophrenic by the hospital staff and kept there even though they behaved normally in all other ways (Rosenhan, 1973).

The answer for purposes of this chapter is that they are not experiments.

In this chapter, we look more closely at non-experimental research. We begin with a general definition of non-experimental research, along with a discussion of when and why non-experimental research is more appropriate than experimental research. We then look separately at two important types of non-experimental research: correlational research and observational research.

Learning Objectives

1. Define non-experimental research, distinguish it clearly from experimental research, and give several examples.
2. Explain when a researcher might choose to conduct non-experimental research as opposed to experimental research.
3. Define correlational research and give several examples.
4. Explain why a researcher might choose to conduct correlational research rather than experimental research or another type of non-experimental research.
5. Explain why correlation does not imply causation

What Is Non-Experimental Research?

Non-experimental research is research that lacks the manipulation of an independent variable. Rather than manipulating an independent variable, researchers conducting non-experimental research simply measure variables as they naturally occur (in the lab or real world). In many cases this kind of research depends on large scale data collection and quantitative techniques for assessing relationships among measurements.

Studies of differences in human personality, as noted in Chapter 2 on Measurement, also reflect the approach of measuring constructs that characterize consistent characteristics of human traits. This approach is often used to address questions about nature versus nurture, which also depends on variables that cannot be directly manipulated.

Non-experimental research techniques also include a range of observational and qualitative techniques that will be discussed in greater detail in Chapter 17. These approaches often involve simply observing behavior in context to identify features of behavior. Research in which aspects of the environment are manipulated outside the lab and behavior is measured can

look like observational research, but if the experimenter is manipulating an independent variable, this approach is a kind of Field Research, which will be discussed in Chapter 19.

Most researchers in psychology consider the distinction between experimental and non-experimental research to be an extremely important one. This is because although experimental research can provide strong evidence that changes in an independent variable cause differences in a dependent variable, non-experimental research generally cannot. As we will see, however, this inability to make causal conclusions does not mean that non-experimental research is less important than experimental research. It is simply used in cases where experimental research is not able to be carried out.

When to Use Non-Experimental Research

Experimental research is appropriate when the researcher has a specific research question or hypothesis about a causal relationship between two variables—and it is possible, feasible, and ethical to manipulate the independent variable. It stands to reason, therefore, that non-experimental research is appropriate—even necessary—when these conditions are not met. There are many times in which non-experimental research is preferred, including when:

- The research question or hypothesis relates to a single variable rather than a statistical relationship between two variables, e.g., how accurate are people's first impressions?.
- The research question pertains to a non-causal statistical relationship between variables, e.g., is there a correlation between verbal intelligence and mathematical intelligence?.
- The research question is about a causal relationship, but the independent variable cannot be manipulated or participants cannot be randomly assigned to conditions or orders of conditions for practical or ethical reasons, e.g., does damage to a person's hippocampus impair the formation of long-term memory traces?.

- The research question is broad and exploratory, or is about what it is like to have a particular experience, e.g., what is it like to be a working mother diagnosed with depression?.

Again, the choice between the experimental and non-experimental approaches is generally dictated by the nature of the research question. Recall the three goals of science are to describe, to predict, and to explain. If the goal is to explain and the research question pertains to causal relationships, then the experimental approach is typically preferred. If the goal is to describe or to predict, a non-experimental approach is appropriate. But the two approaches can also be used to address the same research question in complementary ways. For example, in Milgram's original (non-experimental) obedience study, he was primarily interested in one variable—the extent to which participants obeyed the researcher when he told them to shock the confederate—and he observed all participants performing the same task under the same conditions. However, Milgram subsequently conducted experiments to explore the factors that affect obedience. He manipulated several independent variables, such as the distance between the experimenter and the participant, the participant and the confederate, and the location of the study (Milgram, 1974).

Types of Non-Experimental Research

Non-experimental research falls into two broad categories: correlational research and observational research. In this chapter, we will focus primarily on correlational research. Methods related to observational research will be discussed further in Chapter 19, Qualitative Research.

The most common type of non-experimental research conducted in psychology is correlational research. Correlational research is considered non-experimental because it focuses on the statistical relationship between two variables but does not include the manipulation of an independent variable. More specifically, in correlational research, the researcher measures two variables with little or no attempt to control extraneous variables and

then assesses the relationship between them. As an example, a researcher interested in the relationship between self-esteem and school achievement could collect data on students' self-esteem and their GPAs to see if the two variables are statistically related.

Observational research is non-experimental because it focuses on making observations of behavior in a natural or laboratory setting without manipulating anything. Milgram's original obedience study was non-experimental in this way. He was primarily interested in the extent to which participants obeyed the researcher when he told them to shock the confederate and he observed all participants performing the same task under the same conditions. The study by Loftus and Pickrell (1995) is also a good example of observational research. The dependent variable was whether participants "remembered" having experienced mildly traumatic childhood events (e.g., getting lost in a shopping mall) that they had not actually experienced but that the researchers asked them about repeatedly. In this particular study, nearly a third of the participants "remembered" at least one event. As with Milgram's original study, this study inspired several later studies using experimental research methods on the factors that affect these phenomena.

Internal Validity in Non-Experimental Design

Internal validity is the extent to which the design of a study supports the conclusion that changes in the independent variable caused any observed differences in the dependent variable. Experimental research is highest in internal validity, when properly controlled, because the use of manipulation (of the independent variable) and control (of extraneous variables) help to rule out alternative explanations for the observed relationships. If the average score on the dependent variable in an experiment differs across conditions, it is quite likely that the independent variable is responsible for that difference. Non-experimental (correlational) research is lowest in internal validity because these designs fail to use manipulation or control. There are

also research approaches that blend both experimental and non-experimental approaches described as quasi-experimental research (which will be described in more detail in a subsequent chapter). This type of research contains some, but not all, of the features of a true experiment. For instance, it may fail to use random assignment to assign participants to groups or fail to use counterbalancing to control for potential order effects. Imagine, for example, that a researcher finds two similar schools, starts an anti-bullying program in one, and then finds fewer bullying incidents in that “treatment school” than in the “control school.” While a comparison is being made with a control condition, the inability to randomly assign children to schools could still mean that students in the treatment school differed from students in the control school in some other way that could explain the difference in bullying (e.g., there may be a selection effect).

What Is Correlational Research?

Correlational research is a type of non-experimental research in which the researcher measures two variables (binary or continuous) and assesses the statistical relationship (i.e., the correlation) between them with little or no effort to control extraneous variables. There are many reasons that researchers interested in statistical relationships between variables would choose to conduct a correlational study rather than an experiment. The first is that they do not believe that the statistical relationship is a causal one or are not interested in causal relationships. Recall two goals of science are to describe and to predict and the correlational research strategy allows researchers to achieve both of these goals. Specifically, this strategy can be used to describe the strength and direction of the relationship between two variables and if there is a relationship between the variables then the researchers can use scores on one variable to predict scores on the other (the specific analytic approach will be discussed in Chapter 16).

Another reason that researchers would choose to use a correlational study rather than an experiment is that the statistical relationship of interest is

thought to be causal, but the researcher cannot manipulate the independent variable because it is impossible, impractical, or unethical. For example, while a researcher might be interested in the relationship between the frequency people use cannabis and their memory abilities, they cannot ethically manipulate the frequency that people chronically use cannabis. As such, they must rely on the correlational research strategy; they must simply measure the frequency that people use cannabis and measure their memory abilities using a standardized test of memory and then determine whether the frequency people use cannabis is statistically related to memory test performance.

Another strength of correlational research is that it is often higher in external validity than experimental research. Recall there is typically a trade-off between internal validity and external validity. As greater controls are added to experiments, internal validity is increased but often at the expense of external validity as artificial conditions are introduced that do not exist in reality. In contrast, correlational studies typically have low internal validity because nothing is manipulated or controlled but they often have high external validity. Since nothing is manipulated or controlled by the experimenter the results are more likely to reflect relationships that exist in the real world.

Finally, extending upon this trade-off between internal and external validity, correlational research can help to provide converging evidence for a theory. If a theory is supported by a true experiment that is high in internal validity as well as by a correlational study that is high in external validity, then the researchers can have more confidence in the validity of their theory. As a concrete example, correlational studies establishing that there is a relationship between watching violent television and aggressive behavior have been complemented by experimental studies confirming that the relationship is a causal one (Bushman & Huesmann, 2001).

Does Correlational Research Always Involve Quantitative Variables?

A common misconception among beginning researchers is that correlational research must involve two quantitative variables, such as scores on two extraversion tests or the number of daily hassles and number of symptoms people have experienced. However, the defining feature of correlational research is that the two variables are measured—neither one is manipulated—and this is true regardless of whether the variables are quantitative or categorical. Imagine, for example, that a researcher administers the Rosenberg Self-Esteem Scale to 50 American college students and 50 Japanese college students. Although this “feels” like a between-subjects experiment, it is a correlational study because the researcher did not manipulate the students’ nationalities. The same is true of the study by Cacioppo and Petty (1982) comparing college faculty and factory workers in terms of their need for cognition. It is a correlational study because the researchers did not manipulate the participants’ occupations.

The statistical tool based on calculating a **correlation coefficient** (Chapter 16) does require having two variables that are quantitative and not categorical. While it is commonly the case that correlational research depends on calculating correlation coefficients, be aware that the terms are not actually synonymous. While it is technically possible to have an experimental design that is evaluated statistically with a correlation analysis (calculating a correlation coefficient), in practice this is extremely rare. It would require having a large number of possibilities across the manipulated independent variable, which is a very rare case. In all other cases, the term ‘correlation’ should be avoided in describing experimental research as the neither the statistical tool nor the general method applies (if necessary, the related word ‘associated’ should be preferred over ‘correlated’).

Data Collection in Correlational Research

Again, the defining feature of correlational research is that neither variable is manipulated. It does not matter how or where the variables are measured. A researcher could have participants come to a laboratory to complete a computerized backward digit span task and a computerized risky decision-making task and then assess the relationship between participants' scores on the two tasks. Or a researcher could go to a shopping mall to ask people about their attitudes toward the environment and their shopping habits and then assess the relationship between these two variables. Both of these studies would be correlational because no independent variable is manipulated.

Correlation Does Not Imply Causation

You have probably heard repeatedly that "Correlation does not imply causation." An amusing example of this comes from a 2012 study that showed a positive correlation (Pearson's $r = 0.79$) between the per capita chocolate consumption of a nation and the number of Nobel prizes awarded to citizens of that nation. It seems clear, however, that this does not mean that eating chocolate causes people to win Nobel prizes, and it would not make sense to try to increase the number of Nobel prizes won by recommending that parents feed their children more chocolate.

There are two reasons that correlation does not imply causation. The first is called the directionality problem. Two variables, X and Y, can be statistically related because X causes Y or because Y causes X. Consider, for example, a study showing that whether or not people exercise is statistically related to how happy they are—such that people who exercise are happier on average than people who do not. This statistical relationship is consistent with the idea that exercising causes happiness, but it is also consistent with the idea that happiness causes exercise. Perhaps being happy gives people more energy or leads them to seek opportunities to socialize with others by going to the

gym. The second reason that correlation does not imply causation is called the third-variable problem. Two variables, X and Y, can be statistically related not because X causes Y, or because Y causes X, but because some third variable, Z, causes both X and Y. For example, the fact that nations that have won more Nobel prizes tend to have higher chocolate consumption probably reflects geography in that European countries tend to have higher rates of per capita chocolate consumption and invest more in education and technology (once again, per capita) than many other countries in the world. Similarly, the statistical relationship between exercise and happiness could mean that some third variable, such as physical health, causes both of the others. Being physically healthy could cause people to exercise and cause them to be happier. Correlations that are a result of a third-variable are often referred to as spurious correlations.

User: give me an example of a research study where correlation does not imply causation

ChatGPT: Certainly! Here's an example of a study where correlation does not imply causation:

Example: Ice Cream Sales and Drowning Incidents

Scenario: Researchers observe a strong positive correlation between ice cream sales and drowning incidents. They notice that as ice cream sales increase during the summer months, so do the number of drowning incidents.

Correlation: There is a positive correlation between ice cream sales and drowning incidents. As one variable (ice cream sales) increases, the other variable (drowning incidents) also tends to increase.

Interpretation: One might hastily conclude that eating ice cream leads to an increase in drowning incidents or vice versa.

"Lots of Candy Could Lead to Violence"

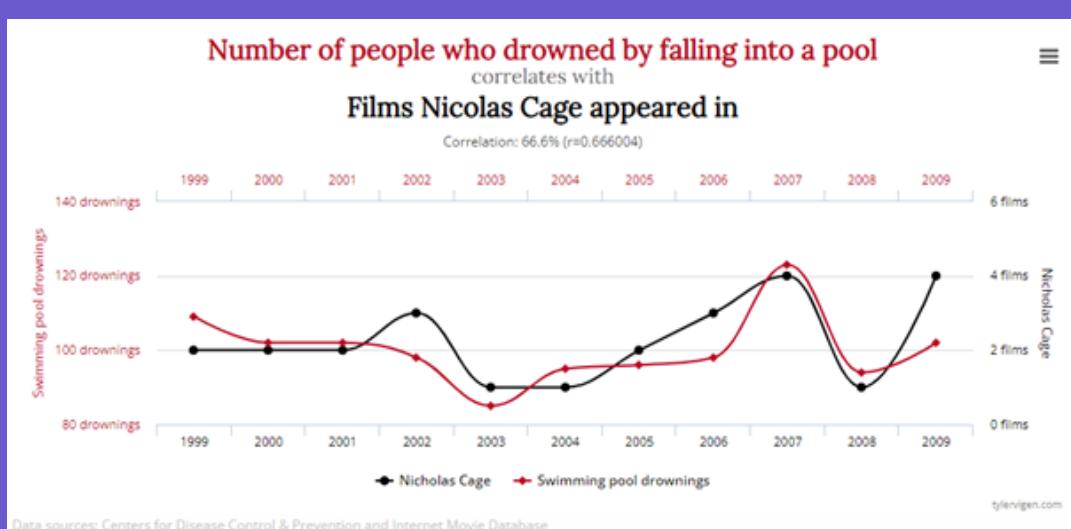
Although researchers in psychology know that correlation does not imply causation, many journalists do not. One website about correlation and causation, http://jonathan.mueller.faculty.noctrl.edu/100/correlation_or_causation.htm, links to dozens of media reports about real biomedical and psychological research. Many of the headlines suggest that a causal relationship has been demonstrated when a careful reading of the articles shows that it has not because of the directionality and third-variable problems.

One such article is about a study showing that children who ate candy every day were more likely than other children to be arrested for a violent offense later in life. But could candy really "lead to" violence, as the headline suggests? What alternative explanations can you think of for this statistical relationship? How could the headline be rewritten so that it is not misleading?

Explanation: However, the correlation between ice cream sales and drowning incidents does not imply causation. In reality, both variables are influenced by a third factor: temperature. Warmer temperatures during the summer months lead to an increase in ice cream sales as people seek ways to cool down. Simultaneously, warm weather encourages people to engage in water-related activities such as swimming, which increases the risk of drowning incidents.

Conclusion: In this example, while there is a correlation between ice cream sales and drowning incidents, it would be incorrect to conclude that one directly causes the other. Instead, both variables are influenced by a common factor, temperature, demonstrating that correlation does not imply causation. To establish causation, further research using experimental designs or controlling for other variables would be necessary.

As you have learned by reading this book, there are various ways that researchers address the directionality and third-variable problems. The most effective is to conduct an experiment. For example, instead of simply measuring how much people exercise, a researcher could bring people into a laboratory and randomly assign half of them to run on a treadmill for 15 minutes and the rest to sit on a couch for 15 minutes. Although this seems like a minor change to the research design, it is extremely important. Now if the exercisers end up in more positive moods than those who did not exercise, it cannot be because their moods affected how much they exercised (because it was the researcher who used random assignment to determine how much they exercised). Likewise, it cannot be because some third variable (e.g., physical health) affected both how much they exercised and what mood they were in. Thus, experiments eliminate the directionality and third-variable problems and allow researchers to draw firm conclusions about causal relationships.



Some other excellent and amusing examples of spurious correlations can be found at <http://www.tylervigen.com>.

Personality Research

Personality research is a branch of psychology that focuses on understanding individual differences in patterns of thinking, feeling, and behaving. It is essentially research on the common kinds of participant variables we have considered as extraneous variables in experimental control and design. Personality researchers seek to explore and explain the characteristics that make each person unique, how these traits develop over time, and how they influence various aspects of life, such as relationships, career choices, and mental health.

The Big Five personality traits, also known as the Five Factor Model (FFM), are five broad dimensions of personality that are commonly used to describe and measure individual differences in personality. These traits have been identified through quantitative analysis of relationships among items from questionnaires asking about behavior, preferences and tendencies. The Big Five traits are:

- Openness to Experience: This trait reflects the extent to which individuals are open-minded, curious, imaginative, and receptive to new experiences. People high in openness tend to be creative, adventurous, and intellectually curious, while those low in openness may be more traditional, conservative, and resistant to change.
- Conscientiousness: Conscientiousness refers to the degree of organization, responsibility, dependability, and self-discipline exhibited by individuals. People high in conscientiousness are diligent, organized, and goal-oriented, while those low in conscientiousness may be more careless, impulsive, and prone to procrastination.
- Extraversion: Extraversion reflects the level of sociability, assertiveness, energy, and positive emotionality displayed by individuals. People high in extraversion are outgoing, talkative, and energetic, enjoying social interactions and seeking stimulation. In contrast, those low in extraversion tend to be introverted, reserved, and prefer solitary activities.
- Agreeableness: Agreeableness refers to the degree of warmth, kindness,

empathy, cooperation, and altruism exhibited by individuals. People high in agreeableness are compassionate, trusting, and cooperative, valuing harmony and interpersonal relationships. Individuals low in agreeableness may be more competitive, skeptical, and less concerned with others' welfare.

- Neuroticism (or Emotional Stability): Neuroticism reflects the tendency to experience negative emotions such as anxiety, depression, moodiness, and vulnerability to stress. Individuals high in neuroticism are prone to worry, rumination, and emotional instability, while those low in neuroticism tend to be emotionally resilient, calm, and emotionally stable.

Measures on these five traits are considered relatively stable over time and across situations, forming the core structure of personality. They provide a comprehensive framework for understanding and measuring individual differences in personality and have been widely used in research across various fields, including psychology, sociology, organizational behavior, and personality assessment.

Research based on these characteristics is necessarily non-experimental as participants cannot be assigned to conditions. Research in this area investigates the ways in which personality traits influence behavior, attitudes, and interpersonal relationships. They examine how different personality traits are associated with various outcomes, such as academic and occupational success, health behaviors, leadership styles, and susceptibility to mental health disorders. Personality disorders are characterized by enduring patterns of maladaptive behavior, thoughts, and emotions that cause significant distress or impairment in functioning. Researchers seek to understand the underlying mechanisms, etiology, and effective treatments for these disorders.

Personality research also is used in studies of cross-cultural and gender differences that appear in personality traits and patterns. Researchers investigate how cultural norms, values, and socialization practices shape personality development and expression across different societies and cultural groups. This approach can often lead to research questions based on

hypotheses about nature versus nurture. That is, the degree to which some aspects of behavior might be innate and observed differences arise from genetic characteristics that vary across individuals. These questions are often contrasted with hypotheses that behavioral tendencies occur due to differences in experiences that might shape behavior over time.

Nature/nurture questions can easily become very contentious and become debates about core, immutable differences in behavior or cognition between people. It is important to remember in these debates that the research tools being applied to these questions are necessarily non-experimental (usually correlational) and the same careful inference framework need to be applied. As we have seen in several examples above, many spurious correlations occur in the world and many observed correlations occur due to unexpected third variables. While these are usually easy to spot as obvious errors in the amusing examples presented, many people overlook the inference problem when evaluating the same claims, for example, about observed differences between men and women.

Key Takeaways

- Non-experimental research is research that lacks the manipulation of an independent variable.
- There are two broad types of non-experimental research. Correlational research that focuses on statistical relationships between variables that are measured but not manipulated; and observational research in which participants are observed and their behavior is recorded without the researcher interfering or manipulating any variables.
- In general, experimental research is high in internal validity, correlational research is low in internal validity, and quasi-experimental research is in between.
- Correlational research involves measuring two variables and assessing the relationship between them, with no manipulation of an independent variable.
- Correlation does not imply causation. A statistical relationship between two variables, X and Y, does not necessarily mean that X causes Y. It is also possible that Y causes X, or that a third variable, Z, causes both X and Y.
- While correlational research cannot be used to establish causal relationships between variables, correlational research does allow researchers to achieve many other important objectives (establishing reliability and validity, providing converging evidence, describing relationships, and making predictions)

Exercises

1. Why are experimental studies higher in internal validity?
2. Why are nonexperimental studies higher in external validity?

For each of the following 3 examples, indicate what kinds of non-experimental design would be used to study the phenomenon and also outline a 2-group experimental approach that could answer a question that might be inspired by the non-experimental work.

- 3a. Pushing ahead in line
- 3b. Increased crime rates during higher temperature months
- 3c. Studiousness in college students

For each of the following statistical relationships, decide whether the directionality problem is present and think of at least one plausible third variable.

- 4a. People who eat more lobster tend to live longer.
- 4b. People who exercise more tend to weigh less.
- 4c. College students who drink more alcohol tend to have poorer grades.

15 Surveys & Instruments

Shortly after the terrorist attacks in New York City and Washington, DC, in September of 2001, researcher (Lerner, Gonzalez, Small & Fischhoff, 2003) reported an Internet-based survey of nearly 2,000 American teens and adults ranging in age from 13 to 88. They asked participants about their reactions to the attacks and for their judgments of various terrorism-related and other risks. Among the results were that the participants tended to overestimate most risks, that females did so more than males, and that there were no differences between teens and adults. The most interesting result, however, had to do with the fact that some participants were “primed” to feel anger by asking them what made them angry about the attacks and by presenting them with a photograph and audio clip intended to evoke anger. Others were primed to feel fear by asking them what made them fearful about the attacks and by presenting them with a photograph and audio clip intended to evoke fear. As the researchers hypothesized, the participants who were primed to feel anger perceived less risk than the participants who had been primed to feel fear—showing how risk perceptions are strongly tied to specific emotions.

The study by Lerner and her colleagues is an example of survey research in psychology—the topic of this chapter. We begin with an overview of survey research, including its definition, some history, and a bit about who conducts

it and why. We then look at survey responding as a psychological process and the implications of this for constructing good survey questionnaires. Finally, we consider some issues related to actually conducting survey research, including sampling the participants and collecting the data.

Learning Objectives

1. Define what survey research is, including its two important characteristics.
2. Describe several different ways that survey research can be used and give some examples.
3. Explain what a **context effect** is and give some examples.
4. Define sampling bias in general and non-response bias in particular. List some techniques that can be used to increase the response rate and reduce non-response bias.
5. Define **instrument reliability**, including the different types and how they are assessed.
6. Define **instrument validity**, including the different types and how they are assessed.

What Is Survey Research?

Survey research is a quantitative and qualitative method with two important characteristics. First, the variables of interest are measured using self-reports (using questionnaires or interviews). In essence, survey researchers ask their participants (who are often called respondents in survey research) to report directly on their own thoughts, feelings, and behaviors. Second, considerable attention is paid to the issue of sampling. In particular, survey researchers have a strong preference for large random samples because they provide the most accurate estimates of what is true in the population. In fact, survey research may be the only approach in psychology in which random sampling is routinely used. Beyond these two characteristics, almost anything goes in survey research. Surveys can be long or short. They can be conducted in person, by telephone, through the mail, or over the Internet. They can be about voting intentions, consumer preferences, social attitudes, health, or anything else that it is possible to ask people about and receive meaningful answers. Although survey data are often analyzed using statistics, there are many questions that lend themselves to more qualitative analysis.

Most survey research is non-experimental. It is used to describe single variables (e.g., the percentage of voters who prefer one presidential candidate or another, the prevalence of schizophrenia in the general population, etc.) and also to assess statistical relationships between variables (e.g., the relationship between income and health). But surveys can also be used within experimental research. The study by Lerner and her colleagues is a good example. Their use of self-report measures and a large national sample identifies their work as survey research. But their manipulation of an independent variable (anger vs. fear) to assess its effect on a dependent variable (risk judgments) also identifies their work as experimental.

History and Uses of Survey Research

Survey research may have its roots in English and American “social surveys”

conducted around the turn of the 20th century by researchers and reformers who wanted to document the extent of social problems such as poverty (Converse, 1987). By the 1930s, the US government was conducting surveys to document economic and social conditions in the country. The need to draw conclusions about the entire population helped spur advances in sampling procedures. At about the same time, several researchers who had already made a name for themselves in market research, studying consumer preferences for American businesses, turned their attention to election polling. A watershed event was the presidential election of 1936 between Alf Landon and Franklin Roosevelt. A magazine called Literary Digest conducted a survey by sending ballots (which were also subscription requests) to millions of Americans. Based on this “straw poll,” the editors predicted that Landon would win in a landslide. At the same time, the new pollsters were using scientific methods with much smaller samples to predict just the opposite—that Roosevelt would win in a landslide. In fact, one of them, George Gallup, publicly criticized the methods of Literary Digest before the election and all but guaranteed that his prediction would be correct. And of course, it was, demonstrating the effectiveness of careful survey methodology (We will consider the reasons that Gallup was right later in this chapter). Gallup’s demonstration of the power of careful survey methods led later researchers to local, and in 1948, the first national election survey by the Survey Research Center at the University of Michigan. This work eventually became the American National Election Studies (<https://electionstudies.org/>) as a collaboration of Stanford University and the University of Michigan, and these studies continue today.

From market research and election polling, survey research made its way into several academic fields, including political science, sociology, and public health—where it continues to be one of the primary approaches to collecting new data. Beginning in the 1930s, psychologists made important advances in questionnaire design, including techniques that are still used today, such as the Likert scale (defined below). Survey research has a strong historical association with the social psychological study of attitudes, stereotypes, and prejudice. Early attitude researchers were also among the first psychologists

to seek larger and more diverse samples than the convenience samples of university students that were routinely used in psychology (and still are).

Survey research continues to be important in psychology today. For example, survey data have been instrumental in estimating the prevalence of various mental disorders and identifying statistical relationships among those disorders and with various other factors. The National Comorbidity Survey is a large-scale mental health survey conducted in the United States (see <http://www.hcp.med.harvard.edu/ncs>). In just one part of this survey, nearly 10,000 adults were given a structured mental health interview in their homes in 2002 and 2003. The table below presents results on the lifetime prevalence of some anxiety, mood, and substance use disorders. Obviously, this kind of information can be of great use both to basic researchers seeking to understand the causes and correlates of mental disorders as well as to clinicians and policymakers who need to understand exactly how common these disorders are.

Table of Some Lifetime Prevalence Results from the National Comorbidity Survey. Note that the lifetime prevalence of a disorder is the percentage of people in the population that develop that disorder at any time in their lives.

Lifetime prevalence			
Disorder	Total	Female	Male
Generalized anxiety disorder	5.7	7.1	4.2
Obsessive-compulsive disorder	2.3	3.1	1.6
Major depressive disorder	16.9	20.2	13.2
Bipolar disorder	4.4	4.5	4.3
Alcohol abuse	13.2	7.5	19.6
Drug abuse	8.0	4.8	11.6

And as the opening example makes clear, survey research can even be used as a data collection method within experimental research to test specific hypotheses about causal relationships between variables. Such studies, when conducted on large and diverse samples, can be a useful supplement to laboratory studies conducted on university students. Survey research is thus a flexible approach that can be used to study a variety of basic and applied

research questions.

Constructing Surveys

The heart of any survey research project is the survey itself. Although it is easy to think of interesting questions to ask people, constructing a good survey is not easy at all. The problem is that the answers people give can be influenced in unintended ways by the wording of the items, the order of the items, the response options provided, and many other factors. At best, these influences add noise to the data. At worst, they result in systematic biases and misleading results. In this section, therefore, we consider some principles for constructing surveys to minimize these unintended effects and thereby maximize the reliability and validity of respondents' answers.

Before looking at specific principles of survey construction, it will help to consider survey responding as a psychological process.

Consider, for example, the following questionnaire item:

How many alcoholic drinks do you consume in a typical day?

- a lot more than average
- somewhat more than average
- average
- somewhat fewer than average
- a lot fewer than average

Although this item at first seems straightforward, it poses several difficulties for respondents. First, they must interpret the question. For example, they must decide whether "alcoholic drinks" include beer and wine (as opposed to just hard liquor) and whether a "typical day" is a typical weekday, typical weekend day, or both. Even though Chang and Krosnick (2003) found that asking about "typical" behavior has been shown to be more valid than

asking about “past” behavior, their study compared “typical week” to “past week” and may be different when considering typical weekdays or weekend days). Once respondents have interpreted the question, they must retrieve relevant information from memory to answer it. But what information should they retrieve, and how should they go about retrieving it? They might think vaguely about some recent occasions on which they drank alcohol, they might carefully try to recall and count the number of alcoholic drinks they consumed last week, or they might retrieve some existing beliefs that they have about themselves (e.g., “I am not much of a drinker”). Then they must use this information to arrive at a tentative judgment about how many alcoholic drinks they consume in a typical day. For example, this mental calculation might mean dividing the number of alcoholic drinks they consumed last week by seven to come up with an average number per day. Then they must format this tentative answer in terms of the response options actually provided. In this case, the options pose additional problems of interpretation. For example, what does “average” mean, and what would count as “somewhat more” than average? Finally, they must decide whether they want to report the response they have come up with or whether they want to edit it in some way. For example, if they believe that they drink a lot more than average, they might not want to report that for fear of looking bad in the eyes of the researcher, so instead, they may opt to select the “somewhat more than average” response option.

From this perspective, what at first appears to be a simple matter of asking people how much they drink (and receiving a straightforward answer from them) turns out to be much more complex. Measurement of alcohol use is a good example of where a stronger methodological approach is to use an established “instrument” for which existing studies of reliability and validity are available. This may also highlight important theoretical questions such as the difference in alcohol use scales aimed at quantifying behavior versus those specifically aimed at identifying possible abuse or addiction disorders (e.g., Greenfield, 2000)

Context Effects on Survey Responses

Again, this complexity can lead to unintended influences on respondents' answers. These are often referred to as context effects because they are not related to the content of the item but to the context in which the item appears (Schwarz & Strack, 1990). For example, there is an item-order effect when the order in which the items are presented affects people's responses. One item can change how participants interpret a later item or change the information that they retrieve to respond to later items. Strack, Martin & Schwarz (1988) asked college students about both their general life satisfaction and their dating frequency. When the life satisfaction item came first, the correlation between the two was only $-.12$ (very weak), suggesting that the two variables are only weakly related. But when the dating frequency item came first, the correlation between the two was $.66$ (strongly correlated), suggesting that those who date more have a strong tendency to be more satisfied with their lives. Reporting the dating frequency first made that information more accessible in memory so that they were more likely to base their life satisfaction rating on it.

The response options provided can also have unintended effects on people's responses (Schwarz, 1999). For example, when people are asked how often they are "really irritated" and given response options ranging from "less than once a year" to "more than once a month," they tend to think of major irritations and report being irritated infrequently. But when they are given response options ranging from "less than once a day" to "several times a month," they tend to think of minor irritations and report being irritated frequently. People also tend to assume that middle response options represent what is normal or typical. So, if they think of themselves as normal or typical, they tend to choose middle response options. For example, people are likely to report watching more television when the response options are centered on a middle option of 4 hours than when centered on a middle option of 2 hours. To mitigate against order effects, rotate questions and response items when there is no natural order. Counterbalancing or randomizing the order of presentation of the questions in online surveys are good practices for survey

questions and can reduce response order effects that show that among undecided voters, the first candidate listed in a ballot receives a 2.5% boost simply by virtue of being listed first!

Writing Survey Items

Questionnaire items can be either open-ended or closed-ended. Open-ended items simply ask a question and allow participants to answer in whatever way they choose. The following are examples of open-ended questionnaire items.

- *"What is the most important thing to teach children to prepare them for life?"*
- *"Please describe a time when you were discriminated against because of your age."*
- *"Is there anything else you would like to tell us about?"*

Open-ended items are useful when researchers do not know how participants might respond or when they want to avoid influencing their responses. Open-ended items tend to be used in qualitative research or in the early stages of a research project.

Closed-ended items ask a question and provide a set of response options for participants to choose from. The alcohol item just mentioned is an example, as are the following:

How old are you?

Under 18

18 to 34

35 to 49

50 to 70

Over 70

On a scale of 0 (no pain at all) to 10 (worst pain ever experienced), how

much pain are you in right now?

Have you ever in your adult life been depressed for a period of 2 weeks or more? Yes No

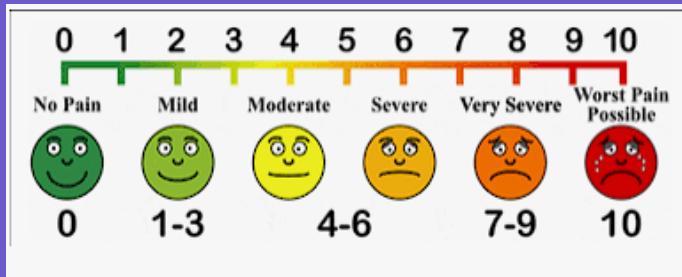
Closed-ended items are used when researchers have a good idea of the different responses that participants might make. They are quantitative in nature, so they are also used when researchers are interested in a well-defined variable or construct such as participants' level of agreement with some statement, perceptions of risk, or frequency of a particular behavior. Closed-ended items are more difficult to write because they must include an appropriate set of response options. However, they are relatively quick and easy for participants to complete. They are also much easier for researchers to analyze because the responses can be easily converted to numbers and entered into a spreadsheet. For these reasons, closed-ended items are much more common.

All closed-ended items include a set of response options from which a participant must choose. For categorical variables like sex, race, or political party preference, the categories are usually listed and participants choose the one (or ones) to which they belong. For quantitative variables, a rating scale is typically provided. A rating scale is an ordered set of responses that participants must choose from. Figure 7.2 shows several examples. The number of response options on a typical rating scale ranges from three to 11—although five and seven are probably most common. Five-point scales are best for unipolar scales where only one construct is tested, such as frequency (Never, Rarely, Sometimes, Often, Always). Seven-point scales are best for bipolar scales where there is a dichotomous spectrum, such as liking (Like very much, Like somewhat, Like slightly, Neither like nor dislike, Dislike slightly, Dislike somewhat, Dislike very much). For bipolar questions, it is useful to offer an earlier question that branches them into an area of the scale; if asking about liking ice cream, first ask "Do you generally like or dislike ice cream?" Once the respondent chooses like or dislike, refine it by offering them relevant choices from the seven-point scale. Branching improves both reliability and validity (Krosnick & Berent, 1993). Although you

often see scales with numerical labels, it is best to only present verbal labels to the respondents but convert them to numerical values in the analyses. Avoid partial labels or length or overly specific labels. In some cases, the verbal labels can be supplemented with (or even replaced by) meaningful graphics.

Example Scales

Here are a few examples of scales with more creative use of graphics or layout used for different kinds of survey responses.



Statement

Academic detailing is a useful form of education that aligns providers' prescribing behavior with evidence-based practice.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	2	3	4	5

How satisfied are you with *

	Very Unsatisfied	Unsatisfied	Neutral	Satisfied	Very Satisfied
Purchase	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Service	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Company Overall	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

1.) On a scale of „no itch“ (left) to „worst imaginable itch“ (right), how was

Please mark a position between 0 and 10 that best represents your itch with a cross on the line below.

...your itch, on average, in the past 24 hours?	
...your worst itch in the past 24 hours?	

Likert Scales

In reading about psychological research, you are likely to encounter the term Likert scale. Although this term is sometimes used to refer to almost any rating scale (e.g., a 0-to-10 life satisfaction scale), it has a much more precise meaning.

In the 1930s, researcher Rensis Likert (pronounced LICK-ert) created a new approach for measuring people's attitudes (Likert, 1932). It involves presenting people with several statements—including both favorable and unfavorable statements—about some person, group, or idea. Respondents then express their agreement or disagreement with each statement on a 5-point scale: Strongly Agree, Agree, Neither Agree nor Disagree, Disagree, Strongly Disagree. Numbers are assigned to each response and then summed across all items to produce a score representing the attitude toward the person, group, or idea. For items that are phrased in an opposite direction (e.g., negatively worded statements instead of positively worded statements), reverse coding is used so that the numerical scoring of statements also runs in the opposite direction. The entire set of items came to be called a Likert scale.

Thus, unless you are measuring people's attitude toward something by assessing their level of agreement with several statements about it, it is best to avoid calling it a Likert scale. You are probably just using a "rating scale."

Writing Effective Items

We can now consider some principles of writing questionnaire items that minimize unintended context effects and maximize the reliability and validity of participants' responses. A rough guideline for writing questionnaire items is provided by the **BRUSO** model (Peterson, 2000). An acronym, BRUSO stands for "brief," "relevant," "unambiguous," "specific," and "objective." Effective questionnaire items are brief and to the point. They avoid long, overly technical, or unnecessary words. This brevity makes them easier

for respondents to understand and faster for them to complete. Effective questionnaire items are also relevant to the research question. If a respondent's sexual orientation, marital status, or income is not relevant, then items on them should probably not be included. Again, this makes the questionnaire faster to complete, but it also avoids annoying respondents with what they will rightly perceive as irrelevant or even "nosy" questions. Effective questionnaire items are also unambiguous; they can be interpreted in only one way. Part of the problem with the alcohol item presented earlier in this section is that different respondents might have different ideas about what constitutes "an alcoholic drink" or "a typical day." Effective questionnaire items are also specific so that it is clear to respondents what their response should be about and clear to researchers what it is about. A common problem here is closed-ended items that are "double barreled." They ask about two conceptually separate issues but allow only one response. For example, "Please rate the extent to which you have been feeling anxious and depressed." This item should probably be split into two separate items—one about anxiety and one about depression. Finally, effective questionnaire items are objective in the sense that they do not reveal the researcher's own opinions or lead participants to answer in a particular way. The table below shows some examples of poor and effective questionnaire items based on the BRUSO criteria. The best way to know how people interpret the wording of the question is to conduct a pilot test and ask a few people to explain how they interpreted the question.

Table: BRUSO Model of Writing Effective Questionnaire Items with Examples

Criterion	Poor	Effective
B—Brief	“Are you now or have you ever been the possessor of a firearm?”	“Have you ever owned a gun?”
R—Relevant	“What is your sexual orientation?”	Do not include this item unless it is clearly relevant to the research
U—Unambiguous	“Are you a gun person?”	“Do you currently own a gun?”
S—Specific	“How much have you read about the new gun control measure and sales tax?”	“How much have you read about the new sales tax?”
O—Objective	“How much do you support the new gun control measure?”	“What is your view of the new gun control measure?”

For closed-ended items, it is also important to create an appropriate response scale. For categorical variables, the categories presented should generally be mutually exclusive and exhaustive. Mutually exclusive categories do not overlap. For a religion item, for example, the categories of Christian and Catholic are not mutually exclusive but Protestant and Catholic are mutually exclusive. Exhaustive categories cover all possible responses. Although Protestant and Catholic are mutually exclusive, they are not exhaustive because there are many other religious categories that a respondent might select: Jewish, Hindu, Buddhist, and so on. In many cases, it is not feasible to include every possible category, in which case an Other category, with a space for the respondent to fill in a more specific response, is a good solution. If respondents could belong to more than one category (e.g., race), they should be instructed to choose all categories that apply.

For rating scales, five or seven response options generally allow about as much precision as respondents are capable of. However, numerical scales with more options can sometimes be appropriate. For dimensions such as attractiveness, pain, and likelihood, a 0-to-10 scale will be familiar to many

respondents and easy for them to use. Regardless of the number of response options, the most extreme ones should generally be “balanced” around a neutral or modal midpoint. An example of an unbalanced rating scale measuring perceived likelihood might look like this:

Unlikely | Somewhat Likely | Likely | Very Likely | Extremely Likely

A balanced version might look like this:

Extremely Unlikely | Somewhat Unlikely | As Likely as Not | Somewhat Likely | Extremely Likely

Note, however, that a middle or neutral response option does not have to be included. Researchers sometimes choose to leave it out because they want to encourage respondents to think more deeply about their response and not simply choose the middle option by default. However, including middle alternatives on bipolar dimensions can be used to allow people to choose an option that is neither.

Formatting the Survey

Writing effective items is only one part of constructing a survey. For one thing, every survey should have a written or spoken introduction that serves two basic functions (Peterson, 2000). One is to encourage respondents to participate in the survey. In many types of research, such encouragement is not necessary either because participants do not know they are in a study (as in naturalistic observation) or because they are part of a subject pool and have already shown their willingness to participate by signing up and showing up for the study. Survey research usually catches respondents by surprise when they answer their phone, go to their mailbox, or check their e-mail—and the researcher must make a good case for why they should agree to participate. Thus, the introduction should briefly explain the purpose of the survey and its importance, provide information about the sponsor of the

survey (university-based surveys tend to generate higher response rates), acknowledge the importance of the respondent's participation, and describe any incentives for participating.

The second function of the introduction is to establish informed consent. Remember that this involves describing to respondents everything that might affect their decision to participate. This includes the topics covered by the survey, the amount of time it is likely to take, the respondent's option to withdraw at any time, confidentiality issues, and so on. Written consent forms are not always used in survey research (when the research is of minimal risk and completion of the survey instrument is often accepted by the IRB as evidence of consent to participate), so it is important that this part of the introduction be well documented and presented clearly and in its entirety to every respondent.

The introduction should be followed by the substantive questionnaire items. But first, it is important to present clear instructions for completing the questionnaire, including examples of how to use any unusual response scales. Remember that the introduction is the point at which respondents are usually most interested and least fatigued, so it is good practice to start with the most important items for purposes of the research and proceed to less important items. Items should also be grouped by topic or by type. For example, items using the same rating scale (e.g., a 5-point agreement scale) should be grouped together if possible to make things faster and easier for respondents. Demographic items are often presented last because they are least interesting to participants but also easy to answer in the event respondents have become tired or bored. Of course, any survey should end with an expression of appreciation to the respondent.

Conducting Surveys

Surveys are famously sensitive to the sampling methodology used to recruit participants into the research study. The sampling methods discussed in Chapter 11 provide the technical terms for a variety of approaches to this challenge. The main concern with being able to draw robust inferences from surveys arise from problems of **sampling bias**.

Probability sampling was developed in large part to address the issue of sampling bias. Sampling bias occurs when a sample is selected in such a way that it is not representative of the entire population and therefore produces inaccurate results. This bias was the reason that the Literary Digest straw poll was so far off in its prediction of the 1936 presidential election. The mailing lists used came largely from telephone directories and lists of registered automobile owners, which over-represented wealthier people, who were more likely to vote for Landon. Gallup was successful because he knew about this bias and found ways to sample less wealthy people as well.

There is one form of sampling bias that even careful random sampling is subject to. It is almost never the case that everyone selected for the sample actually responds to the survey. Some may have died or moved away, and others may decline to participate because they are too busy, are not interested in the survey topic, or do not participate in surveys on principle. If these survey non-responders differ from survey responders in systematic ways, then this difference can produce **non-response bias**. For example, in a mail survey on alcohol consumption, researcher Vivienne Lahaut and colleagues found that only about half the sample responded after the initial contact and two follow-up reminders (Lahaut, Jansen, van de Mheen, Garretsen, 2002). The danger here is that the half who responded might have different patterns of alcohol consumption than the half who did not, which could lead to inaccurate conclusions on the part of the researchers. So to test for non-response bias, the researchers later made unannounced visits to the homes of a subset of the non-responders—coming back up to five times if they did not find them at home. They found that the original non-responders

included an especially high proportion of abstainers (nondrinkers), which meant that their estimates of alcohol consumption based only on the original responders were too high.

Although there are methods for statistically correcting for non-response bias, they are based on assumptions about the non-responders—for example, that they are more similar to late responders than to early responders—which may not be correct. For this reason, the best approach to minimizing non-response bias is to minimize the number of non-responders—that is, to maximize the response rate. There is a large research literature on the factors that affect survey response rates (Groves et al., 2004). In general, in-person interviews have the highest response rates, followed by telephone surveys, and then mail and Internet surveys. Among the other factors that increase response rates are sending potential respondents a short pre-notification message informing them that they will be asked to participate in a survey in the near future and sending simple follow-up reminders to non-responders after a few weeks. The perceived length and complexity of the survey can also make a difference, which is why it is important to keep survey questionnaires as short, simple, and on topic as possible. Finally, offering an incentive—especially cash—is a reliable way to increase response rates. However, ethically, there are limits to offering incentives that may be so large as to be considered coercive.

The four main ways to conduct surveys are through in-person interviews, by telephone, through the mail, and over the internet. As with other aspects of survey design, the choice depends on both the researcher's goals and the budget. In-person interviews have the highest response rates and provide the closest personal contact with respondents. Personal contact can be important, for example, when the interviewer must see and make judgments about respondents, as is the case with some mental health interviews. But in-person interviewing is by far the most costly approach. Telephone surveys have lower response rates and still provide some personal contact with respondents. They can also be costly but are generally less so than in-person interviews. Traditionally, telephone directories have provided fairly

comprehensive sampling frames. However, this trend is less true today as more people choose to only have cell phones and do not install land lines that would be included in telephone directories. Mail surveys are less costly still but generally have even lower response rates—making them most susceptible to non-response bias.

Not surprisingly, internet surveys are becoming more common. They are increasingly easy to construct and use. Although initial contact can be made by mail with a link provided to the survey, this approach does not necessarily produce higher response rates than an ordinary mail survey. A better approach is to make initial contact by email with a link directly to the survey. This approach can work well when the population consists of the members of an organization who have known email addresses and regularly use them (e.g., a university community). For other populations, it can be difficult or impossible to find a comprehensive list of email addresses to serve as a sampling frame. Alternatively, a request to participate in the survey with a link to it can be posted on websites known to be visited by members of the population. But again it is very difficult to get anything approaching a random sample this way because the members of the population who visit the websites are likely to be different from the population as a whole. However, internet survey methods are in rapid development. Because of their low cost, and because more people are online than ever before, internet surveys are likely to become the dominant approach to survey data collection in the near future.

Surveys as Research Instruments

There are a wide variety of established questionnaires that have been developed and extensively tested and used across studies to establish a reliable and valid measure of a specific set of underlying constructs. A well-established survey measure with a strong history of use in research will often be referred to as a research instrument. These will be published as a list of the questions used in the measure and a scoring system for combining

responses into a single quantitative score (as in the Self-Esteem measure described in Chapter 2). There are a large number of these established measures you may encounter in reading the research literature.

Some of these existing measures, particularly those that have applications in clinical psychology, are proprietary. This means that a publisher owns the rights to them and that you would have to purchase them. These include many standard intelligence tests, the Beck Depression Inventory, and the Minnesota Multiphasic Personality Inventory (MMPI). Details about many of these measures and how to obtain them can be found in other reference books, including *Tests in Print* and the *Mental Measurements Yearbook*. There are also tools implemented in assessment technology like tablet computers that provide reliable, consistent assessments like the NIH ToolBox of neurobehavioral assessments.

In planning research, it is generally a good idea to use an existing measure that has been used successfully in previous research instead of attempting to develop your own. Among the advantages are that (a) you save the time and trouble of creating your own, (b) there is already some evidence that the measure is valid (if it has been used successfully), and (c) your results can more easily be compared with and combined with previous results. In fact, if there already exists a reliable and valid measure of a construct, other researchers might expect you to use it unless you have a good and clearly stated reason for not doing so.

If you choose to use an existing measure, you may still have to choose among several alternatives. You might choose the most common one, the one with the best evidence of reliability and validity, the one that best measures a particular aspect of a construct that you are interested in or even the one that would be easiest to use. For example, the Ten-Item Personality Inventory (TIPI) is a self-report questionnaire that measures all the Big Five personality dimensions with just 10 items (Gosling, Rentfrow & Swann, 2003). It is not as reliable or valid as longer and more comprehensive measures, but a researcher might choose to use it when testing time is severely limited.

Creating a New Measure

Instead of using an existing measure, you might want to create your own. Perhaps there is no existing measure of the construct you are interested in or existing ones are too difficult or time-consuming to use. Or perhaps you want to use a new measure specifically to see whether it works in the same way as existing measures—that is, to evaluate convergent validity. In this section, we consider some general issues in creating new measures that apply equally to self-report, behavioral, and physiological measures.

First, be aware that most new measures in psychology are really variations of existing measures, so you should still look to the research literature for ideas. Perhaps you can modify an existing questionnaire, create a paper-and-pencil version of a measure that is normally computerized (or vice versa), or adapt a measure that has traditionally been used for another purpose. For example, the famous Stroop task (Stroop, 1935)—in which people quickly name the colors that various color words are printed in—has been adapted for the study of social anxiety. People high in social anxiety are slower at color naming when the words have negative social connotations such as “stupid” (Amir, Freshman, & Foa, 2002).

When you create a new measure, you should strive for simplicity, aiming to keep the measure brief to avoid boring or frustrating your participants to the point that their responses start to become less reliable and valid. The need for brevity, however, needs to be weighed against the fact that it is nearly always better for a measure to include multiple items rather than a single item. There are two reasons for this. One is a matter of content validity. Multiple items are often required to cover a construct adequately. The other is a matter of reliability. People’s responses to single items can be influenced by all sorts of irrelevant factors—misunderstanding the particular item, a momentary distraction, or a simple error such as checking the wrong response option. But when several responses are summed or averaged, the effects of these irrelevant factors tend to cancel each other out to produce more reliable scores. When using multiple items this way, there will typically be a way to

combine them into a single overall score by summing or averaging.

The method of scoring the items is often part of the procedure for implementing the new survey questions. Much of this will use ideas from previous discussions of experimental control, avoiding bias in responses and considering demand characteristics on participants. Although informed consent requires telling participants what they will be doing, it does not require revealing your hypothesis or other information that might suggest to participants how you expect them to respond. A questionnaire designed to measure financial responsibility need not be titled "Are You Financially Responsible?" It could be titled "Money Questionnaire" or have no title at all. Finally, the effects of your expectations can be minimized by arranging to have the measure administered by a helper who is "blind" or unaware of its intent or of any hypothesis being tested. Regardless of whether this is possible, you should standardize all interactions between researchers and participants—for example, by always reading the same set of instructions word for word.

When using questionnaires that ask about sensitive or personal questions, methods can be used to guarantee participants' anonymity and make clear to them that you are doing so. If you are testing them in groups, be sure that they are seated far enough apart that they cannot see each other's responses. You can even allow them to seal completed questionnaires into individual anonymous envelopes or put them into a drop box where they immediately become mixed with others' questionnaires.

Evaluating a Measure

Every new measure needs to be thoroughly evaluated in terms of its **reliability** and **validity**. These terms are used here in the same spirit as when they are applied to research design. A reliable measure should produce similar results when used multiple times. A valid measure is thought to be an effective assessment of the intended construct.

Instrument reliability requires that it be consistent when used to measure a construct. Psychologists consider three types of consistency: over time (test-retest reliability), across items (internal consistency), and across different researchers (inter-rater reliability).

Test-retest reliability is the extent to which the measure is consistent across time. For example, intelligence is generally thought to be consistent across time. A person who is highly intelligent today will be highly intelligent next week. This means that any good measure of intelligence should produce roughly the same scores for this individual next week as it does today. Clearly, a measure that produces highly inconsistent scores over time cannot be a very good measure of a construct that is supposed to be consistent.

Assessing test-retest reliability requires using the measure on a group of people at one time, using it again on the same group of people at a later time, and then looking at the test-retest correlation between the two sets of scores. This is typically done by computing the correlation coefficient between tests (Chapter 16).

Again, high test-retest correlations make sense when the construct being measured is assumed to be consistent over time, which is the case for intelligence, self-esteem, and the Big Five personality dimensions. But other constructs are not assumed to be stable over time. The very nature of mood, for example, is that it changes. A measure of mood that produced a low test-retest correlation over a period of a month would not be a cause for concern.

Another kind of reliability is **internal consistency**, which is the consistency of people's responses across the items on a multiple-item measure. In general, all the items on such measures are supposed to reflect the same underlying construct, so people's scores on those items should be correlated with each other. On the Rosenberg Self-Esteem Scale, people who agree that they are a person of worth should tend to agree that they have a number of good qualities. If people's responses to the different items are not correlated with each other, then it would no longer make sense to claim that they are all measuring the same underlying construct. This is as true for behavioral

and physiological measures as for self-report measures. For example, people might make a series of bets in a simulated game of roulette as a measure of their level of risk seeking. This measure would be internally consistent to the extent that individual participants' bets were consistently high or low across trials.

Like test-retest reliability, internal consistency can only be assessed by collecting and analyzing data. One approach is to look at a split-half correlation. This involves splitting the items into two sets, such as the first and second halves of the items or the even- and odd-numbered items. Then a score is computed for each set of items, and the relationship between the two sets of scores is examined. A more elaborate version of this is to use a statistic called Cronbach's α (the Greek letter alpha). Conceptually, α is the mean of all possible split-half correlations for a set of items. For example, there are 252 ways to split a set of 10 items into two sets of five. Cronbach's α would be the mean of the 252 split-half correlations. Note that this is not how α is actually computed, but it is a correct way of interpreting the meaning of this statistic. Again, a value of +.80 or greater is generally taken to indicate good internal consistency.

Many behavioral measures involve significant judgment on the part of an observer or a rater. **Inter-rater reliability** is the extent to which different observers are consistent in their judgments. For example, if you were interested in measuring university students' social skills, you could make video recordings of them as they interacted with another student whom they are meeting for the first time. Then you could have two or more observers watch the videos and rate each student's level of social skills. To the extent that each participant does, in fact, have some level of social skills that can be detected by an attentive observer, different observers' ratings should be highly correlated with each other. Inter-rater reliability would also have been measured in Bandura's Bobo doll study. In this case, the observers' ratings of how many acts of aggression a particular child committed while playing with the Bobo doll should have been highly positively correlated. Interrater reliability is often assessed using Cronbach's α when the judgments

are quantitative or an analogous statistic called Cohen's κ (the Greek letter kappa) when they are categorical.

Instrument Validity

Validity is the extent to which the scores from a measure represent the construct they are intended to. But how do researchers make this judgment? We have already considered one factor that they take into account—reliability. When a measure has good test-retest reliability and internal consistency, researchers should be more confident that the scores represent what they are supposed to. There has to be more to it, however, because a measure can be extremely reliable but have no validity whatsoever. As an absurd example, imagine someone who believes that people's index finger length reflects their self-esteem and therefore tries to measure self-esteem by holding a ruler up to people's index fingers. Although this measure would have extremely good test-retest reliability, it would have absolutely no validity. The fact that one person's index finger is a centimeter longer than another's would indicate nothing about which one had higher self-esteem.

Discussions of validity usually divide it into several distinct "types." But a good way to interpret these types is that they are other kinds of evidence—in addition to reliability—that should be taken into account when judging the validity of a measure. Here we consider three basic kinds: face validity, content validity, and criterion validity.

Face validity is the extent to which a measurement method appears "on its face" to measure the construct of interest. Most people would expect a self-esteem questionnaire to include items about whether they see themselves as a person of worth and whether they think they have good qualities. So a questionnaire that included these kinds of items would have good face validity. The finger-length method of measuring self-esteem, on the other hand, seems to have nothing to do with self-esteem and therefore has poor face validity. Although face validity can be assessed quantitatively—for example, by having a large sample of people rate a measure in terms of whether it appears to measure what it is intended to—it is usually assessed

informally.

Face validity is at best a very weak kind of evidence that a measurement method is measuring what it is supposed to. One reason is that it is based on people's intuitions about human behavior, which are frequently wrong. It is also the case that many established measures in psychology work quite well despite lacking face validity. The Minnesota Multiphasic Personality Inventory-2 (MMPI-2) measures many personality characteristics and disorders by having people decide whether each of over 567 different statements applies to them—where many of the statements do not have any obvious relationship to the construct that they measure. For example, the items "I enjoy detective or mystery stories" and "The sight of blood doesn't frighten me or make me sick" both measure the suppression of aggression. In this case, it is not the participants' literal answers to these questions that are of interest, but rather whether the pattern of the participants' responses to a series of questions matches those of individuals who tend to suppress their aggression.

Content validity is the extent to which a measure "covers" the construct of interest. For example, if a researcher conceptually defines test anxiety as involving both sympathetic nervous system activation (leading to nervous feelings) and negative thoughts, then his measure of test anxiety should include items about both nervous feelings and negative thoughts. Or consider that attitudes are usually defined as involving thoughts, feelings, and actions toward something. By this conceptual definition, a person has a positive attitude toward exercise to the extent that they think positive thoughts about exercising, feels good about exercising, and actually exercises. So to have good content validity, a measure of people's attitudes toward exercise would have to reflect all three of these aspects. Like face validity, content validity is not usually assessed quantitatively. Instead, it is assessed by carefully checking the measurement method against the conceptual definition of the construct.

Criterion validity is the extent to which people's scores on a measure are correlated with other variables (known as criteria) that one would expect

them to be correlated with. For example, people's scores on a new measure of test anxiety should be negatively correlated with their performance on an important school exam. If it were found that people's scores were in fact negatively correlated with their exam performance, then this would be a piece of evidence that these scores really represent people's test anxiety. But if it were found that people scored equally well on the exam regardless of their test anxiety scores, then this would cast doubt on the validity of the measure.

A criterion can be any variable that one has reason to think should be correlated with the construct being measured, and there will usually be many of them. For example, one would expect test anxiety scores to be negatively correlated with exam performance and course grades and positively correlated with general anxiety and with blood pressure during an exam. Or imagine that a researcher develops a new measure of physical risk taking. People's scores on this measure should be correlated with their participation in "extreme" activities such as snowboarding and rock climbing, the number of speeding tickets they have received, and even the number of broken bones they have had over the years. When the criterion is measured at the same time as the construct, criterion validity is referred to as concurrent validity; however, when the criterion is measured at some point in the future (after the construct has been measured), it is referred to as predictive validity (because scores on the measure have "predicted" a future outcome).

Criteria can also include other measures of the same construct. For example, one would expect new measures of test anxiety or physical risk taking to be positively correlated with existing established measures of the same constructs. This is known as convergent validity. The use of convergent validity is obviously challenging for cases where a complete new measure is being developed. In some cases, the approach is used of showing that the measure does not capture a construct measured by a different instrument.

Discriminant validity, on the other hand, is the extent to which scores on a measure are not correlated with measures of variables that are conceptually distinct. For example, self-esteem is a general attitude toward the self that is fairly stable over time. It is not the same as mood, which is how good or bad

one happens to be feeling right now. So people's scores on a new measure of self-esteem should not be very highly correlated with their moods. If the new measure of self-esteem were highly correlated with a measure of mood, it could be argued that the new measure is not really measuring self-esteem; it is measuring mood instead.

When they created the Need for Cognition Scale, Cacioppo and Petty also provided evidence of discriminant validity by showing that people's scores were not correlated with certain other variables. For example, they found only a weak correlation between people's need for cognition and a measure of their cognitive style—the extent to which they tend to think analytically by breaking ideas into smaller parts or holistically in terms of "the big picture." They also found no correlation between people's need for cognition and measures of their test anxiety and their tendency to respond in socially desirable ways. All these low correlations provide evidence that the measure is reflecting a conceptually distinct construct.

It should be clear that establishing a new measure that has high reliability and robust validity is typically the product of an extensive program of psychological research. The recommendation here to generally prefer using existing, published measures is to take advantage of the work that previous researchers have already invested in doing this. At the same time, simple rating scales presented via surveys are often highly effective methods for collecting information about participants attitudes, opinions or intentions. These less formal approaches typically have very high face validity but the lack of prior research means that it may be necessary to consider limitations of the research of this kind of measure together with the standard concerns about sampling and possible bias from non-responders.

Key Takeaways

- Survey research features the use of self-report measures on carefully selected samples. It is a flexible approach that can be used to study a wide variety of basic and applied research questions.
- Survey research has its roots in applied social research, market research, and election polling. It has since become an important approach in many academic disciplines, including political science, sociology, public health, and, of course, psychology.
- Survey research involves asking respondents to self-report on their own thoughts, feelings, and behaviors.
- Most survey research is non-experimental in nature (it is used to describe variables or measure statistical relationships between variables) but surveys can also be used to measure dependent variables in true experiments.
- Responding to a survey item is itself a complex cognitive process that involves interpreting the question, retrieving information, making a tentative judgment, putting that judgment into the required response format, and editing the response.
- Survey responses are subject to numerous context effects due to question wording, item order, response options, and other factors. Researchers should be sensitive to such effects when constructing surveys and interpreting survey results.
- Survey items are either open-ended or closed-ended. Open-ended items simply ask a question and allow respondents to answer in whatever way they want. Closed-ended items ask a question and provide several response options that respondents must choose from.
- Use verbal labels instead of numerical labels although the responses can be converted to numerical data in the analyses.
- According to the BRUSO model, questionnaire items should be brief, relevant, unambiguous, specific, and objective.
- Survey research usually involves probability sampling, in which each member of

the population has a known probability of being selected for the sample. Types of probability sampling include simple random sampling, stratified random sampling, and cluster sampling.

- Sampling bias occurs when a sample is selected in such a way that it is not representative of the population and therefore produces inaccurate results. The most pervasive form of sampling bias is non-response bias, which occurs when people who do not respond to the survey differ in important ways from people who do respond. The best way to minimize non-response bias is to maximize the response rate by prenotifying respondents, sending them reminders, constructing questionnaires that are short and easy to complete, and offering incentives.
- Surveys can be conducted in person, by telephone, through the mail, and on the internet. In-person interviewing has the highest response rates but is the most expensive. Mail and internet surveys are less expensive but have much lower response rates. Internet surveys are likely to become the dominant approach because of their low cost.

Exercises

- Discussion: Think of a question that each of the following professionals might try to answer using survey research.
 - a social psychologist
 - an educational researcher
 - a market researcher who works for a supermarket chain
 - the mayor of a large city
 - the head of a university police force
- Discussion: Write a survey item and then write a short description of how someone might respond to that item based on the cognitive model of survey responding (or choose any item on the Rosenberg Self-Esteem Scale at <http://www.bsos.umd.edu/socy/research/rosenberg.htm>).
- Practice: Write survey items for each of the following general questions. In some cases, a series of items, rather than a single item, might be necessary.
 - How much does the respondent use Facebook?
 - How much exercise does the respondent get?
 - How likely does the respondent think it is that the incumbent will be re-elected in the next presidential election?
 - To what extent does the respondent experience “road rage”?
- Discussion: If possible, identify an appropriate sampling frame for each of the following populations. If there is no appropriate sampling frame, explain why.
 - students at a particular university
 - adults living in the state of Washington
 - households in Pullman, Washington
 - people with low self-esteem

16 Statistics 3

In the previous two discussions of statistical tools for psychological science, we reviewed the main workhorses of experimental research: the t-test and the ANOVA. Most designs that have one or more variables manipulated across two (or more) levels with a measured operational definition of the dependent variable will be analyzed with those tools. Here we will review two additional simple statistical methods that can be applied to relationships between experimental variables when it is not the case that we have a categorical independent variable and a measured dependent variable.

The correlation analysis is a mathematical tool for quantifying the reliability of the relationship between two continuous variables. It is very commonly used in non-experimental designs, although less so in experimental research. It does serve as an introduction point to a set of more complex approaches that are used in non-experimental research. At the end of the chapter we will very briefly review a set of more complex statistical tools that build on linear regression models with the observation that correlation is the simplest possible version of linear regression. Detailed discussion of the use and application of these tools are beyond the scope of this text but are core to complex quantitative techniques used to help draw conclusions about causality between variables in non-experimental design.

We will also describe a non-parametric statistical tool that is applied to

cases where the dependent variable is a categorical outcome rather than a continuous measure. The **chi-squared** χ^2 analysis assesses the reliability of the effect of manipulating an independent variable on the rates of occurrence of a dependent variable with two distinct outcomes. The name of this analysis references the statistical parameter used to characterize the strength of the effect of the IV on the DV, just as in our other statistical tools, except that in this case the statistical parameter is the Greek letter X, pronounced ki. This analysis method is presented as an introduction to a class of analysis tools for this kind of outcome data. As with general linear regression models, we will not provide a thorough exploration of these more complex tools here.

Learning Objectives

1. Understand when and how to use a correlation analysis
2. Interpret and understand ranges for Pearson's and Spearman's r
3. Understand how correlational data is displayed in a scatter plot diagram
4. Understand when and how to do a χ^2 analysis with categorical data
5. Understand data presented in rates tables
6. Understand how to draw inferences from χ^2 analysis.

Correlation Analysis

For non-experimental research, simple percentages may be computed to describe the percentage of people who engaged in some behavior or held some belief. But more commonly, non-experimental research involves computing the correlation between two variables. A **correlation coefficient** describes the strength and direction of the relationship between two variables. The values of a correlation coefficient can range from -1.00 (the strongest possible negative relationship) to $+1.00$ (the strongest possible positive relationship). A value of 0 means there is no relationship between the two variables. Positive correlation coefficients indicate that as the values of one variable increase, so do the values of the other variable. A good example of a positive correlation is the correlation between height and weight, because as height increases weight also tends to increase. Negative correlation coefficients indicate that as the value of one variable increase, the values of the other variable decrease. An example of a negative correlation is the correlation between stressful life events and happiness; because as stress increases, happiness is likely to decrease.

The phrase **Correlation is not Causality** is common and important to psychological science. The challenges this creates for validity of conclusions was reviewed in detail in Chapter 14. Here we consider the statistical tools used to assess reliability of the relationship between two variables. As

always, a reliable relationship does not automatically imply the conclusion is valid. We might observe a very robust relationship between variable X and Y and still not know for sure if X caused Y, Y caused X or some third variable Z caused the observed relationship between X and Y.

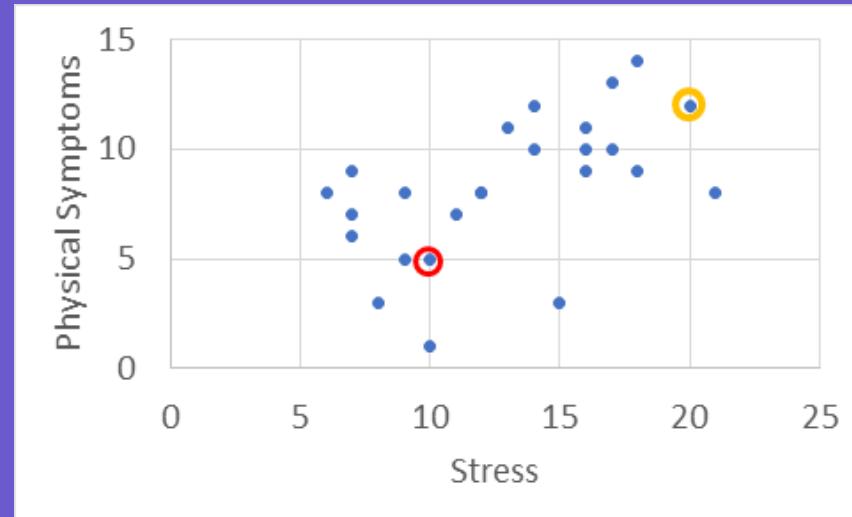
Mathematically, a correlation analysis simply assesses the strength of the relationship between two continuous variables. In theory, if the independent variable was manipulated across a continuous range by the experimenter, we could draw perfectly reasonable causal inference from a correlation coefficient. In practice, this is vanishingly rare, so it is generally safe to use the heuristic that if you see a correlation coefficient reflecting a correlation analysis, the research to which it is being applied is likely also correlational research and non-experimental.

Correlations Between Quantitative Variables

Correlations between quantitative variables are often presented using **scatterplots**. An example is shown on the right based on hypothetical data on the relationship between the amount of stress people are under and the number of physical symptoms they have. Each point in the scatterplot represents one person's score on both variables. For example, the circled point in red represents a person whose stress score was 10 and who had five physical symptoms. The orange circled point is a participant with a stress score of 20 and twelve physical symptoms. Taking all the points into account, one can see that people under more stress tend to have more physical symptoms. This is a good example of a positive relationship, in which higher scores on one variable tend to be associated with higher scores on the other. In other words, they move in the same direction, either both up or both down. A negative relationship is one in which higher scores on one variable tend to be associated with lower scores on the other. In other words, they move in opposite directions. There is a negative relationship between stress and immune system functioning, for example, because higher stress is associated with lower immune system functioning.

Scatter Plots

The graph type shown to the right is a *scatterplot*. This is a graph where each data point is shown on (x,y) coordinates. This type of data visualization is notable for making



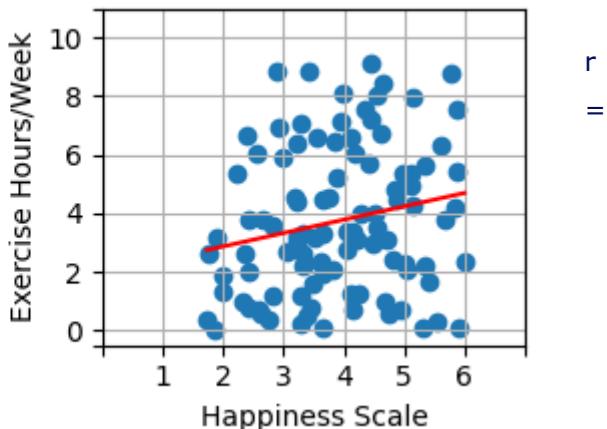
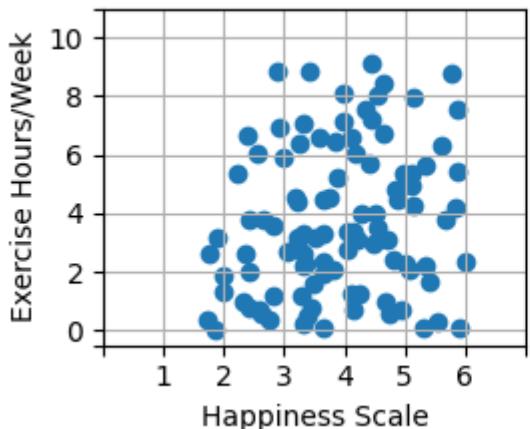
visible every point in the dataset, which can be messy but is very useful for spotting outlier points and getting a sense of the overall relationship between the two measures. Many tools also allow a quick addition of a *trendline* to help visualize the linear relationship.

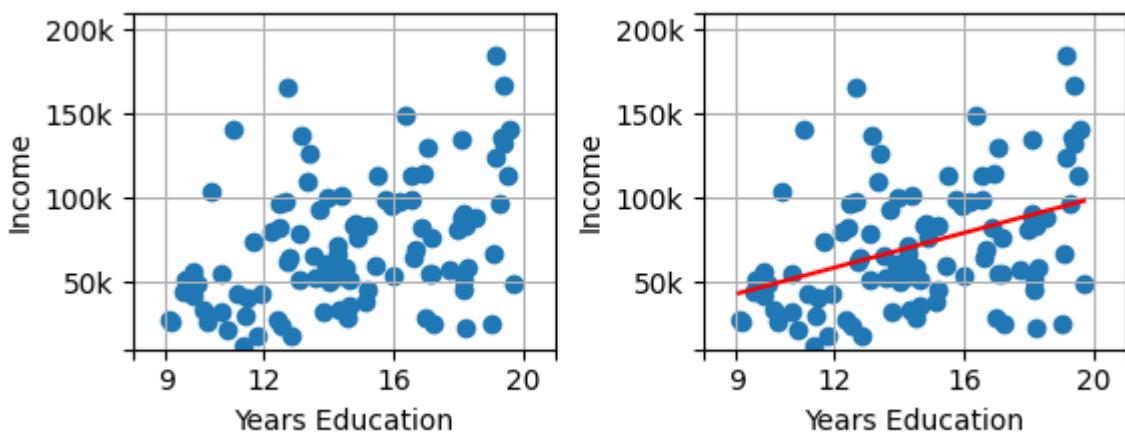
The strength of a correlation between quantitative variables is typically measured using a statistic called **Pearson's Correlation Coefficient** (or **Pearson's r**). Pearson's r ranges from -1.00 (the strongest possible negative relationship) to $+1.00$ (the strongest possible positive relationship). A value of 0 means there is no relationship between the two variables. When Pearson's r is 0 , the points on a scatterplot form a shapeless "cloud." As its value moves toward -1.00 or $+1.00$, the points come closer and closer to falling on a single straight line. Correlation coefficients near $\pm .10$ are considered small, values near $\pm .30$ are considered medium, and values near $\pm .50$ are considered large. Notice that the sign of Pearson's r is unrelated to its strength. Pearson's r values of $.30$ and $-.30$, for example, are equally strong; it is just that one represents a moderate positive relationship and the other a moderate negative relationship. With the exception of reliability coefficients, most correlations that we find in psychology are small or

moderate in size. The scatterplot above has a correlation coefficient between the hypothetical data of 0.55, a fairly strong positive relationship where physical symptoms (y-axis) go up as stress (x-axis) go up.

An analysis that produces a correlation coefficient is expressed with the statistical parameter, r , which like other statistical parameters (t , F) reflects the strength of the relationship between the variables. It is also associated with a p-value, which always has the same definition, the probability of observing this relationship by chance if the null hypothesis was correct. For a correlation analysis, the null hypothesis is that there is no relationship between variables which would produce an $r = 0.00$.

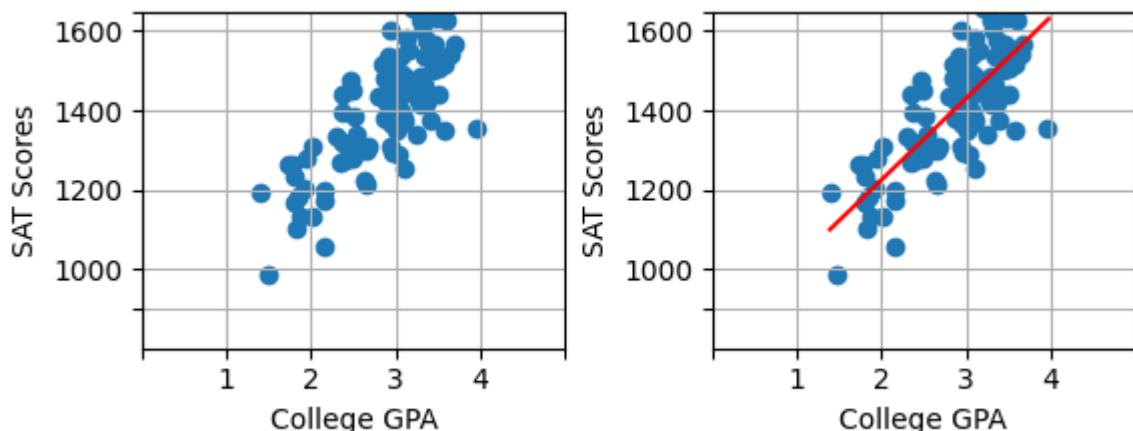
In this simulated example, we see a relationship between hours per week of exercise and scores on a happiness questionnaire. The correlation shown here is $r = 0.20$. The left panel shows the scatterplot of all the data (100 simulated values). With such a weak relationship, it is difficult to see that happiness is going up very slightly with increased exercise. On the right panel, a trendline has been added which shows the average increase. For this kind of analysis, it should be noted that the weakness of this correlated relationship is affected by the number of points that are very far from the trendline. Although there is a tendency for points to be higher on the right side of the graph, the relationship is not very strong, although it does meet the reliability criterion of $p < .05$.

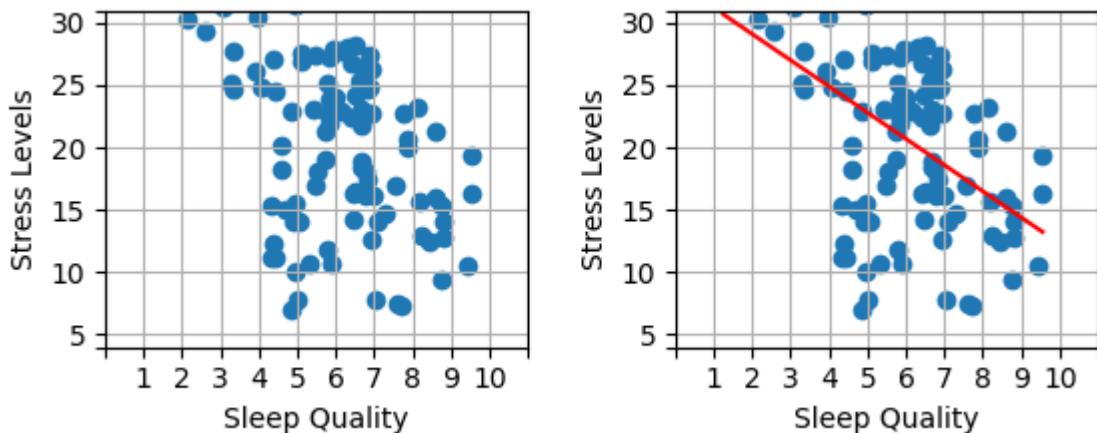




In the second example above, we have a simulated relationship between years of education and income after school. On the left panel, the relationship is now more easily visible, reflecting the fact that this relationship is producing a correlation of $r = 0.4$. The relationship is even easier to see on the right panel with the added trendline. You might note that the slopes of the line is not that different from the previous example, but the dots (which are each one participant) are more closely clustered around the line.

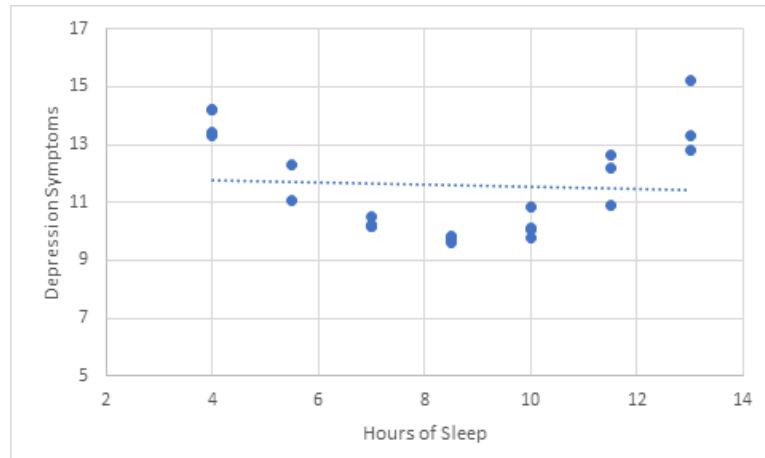
The example below shows a much stronger relationship between college grades and standardized tests taken before starting at school. These simulated data show a very strong relationship, $r = 0.80$.





The previous three examples all showed positive relationships between the two variables assessed. In this example, we see a negative correlation reflecting that sleep quality decreases with higher stress levels. These simulated data illustrate a robust negative relationship, $r = -0.5$.

There are two common situations in which the value of Pearson's r can be misleading. Pearson's r is a good measure only for linear relationships, in which the points are best approximated by a straight line. It is not a good measure for curvilinear relationships, in which the points are better approximated by a curved line. The figure below, for example, shows a hypothetical relationship between the amount of sleep people get per night and their level of depression.



In this example, the line that best approximates the points is a U-shaped curve because people who get about eight hours of sleep tend to be the least depressed. Those who get too little sleep and those who get too much sleep tend to be more depressed. Even though the figure shows a fairly systematic relationship between depression and sleep, Pearson's r would be close to zero because the points in the scatterplot are not well fit by a single straight line (flat trend line shown). This means that it is important to make a scatterplot and confirm that a relationship is approximately linear before using Pearson's r . Curvilinear relationships are fairly common in psychology. The technique for measuring them extends the idea of the linear trendlines seen on the scatterplots above to curvilinear lines defined by polynomials, which goes beyond the scope of the tools discussed here.

Another common situations in which the value of Pearson's r can be misleading is when one or both of the variables have a limited range in the sample relative to the population. This problem is referred to as restriction of range. Assume, for example, that there is a strong negative correlation between people's age and their enjoyment of hip hop music as shown by the scatterplot across age ranges from 18 to 80. However, if data were collected from a restricted range sample, e.g., 18 to 24, the relationship might not be visible. This is yet another example of why we cannot confidently draw conclusions from null results. It is also a reminder that calculation of a correlation coefficient based on Pearson's r depends on having data sampled across a reasonably wide range and also assumes that the distribution of both the x and y variables are roughly normal (following a Gaussian distribution).

A tool to be aware for conditions in which the data are not normally distributed is the **Spearman's rank correlation**. This also results in calculating an r statistic that acts just like the Pearson's correlation. Spearman's correlation can be used when the observed data has a number of notable outliers that would not be expected in a normally distributed dataset. This correlation coefficient is calculated based on ranking the data such that the lowest value is recoded as 1 and each higher value is one more so that the highest value in the data set is the number of total participants.

This reduces the distorting impact of extreme outliers that can reduce the effectiveness of a more typical Person's correlation. An example of where this tool can be used effectively is in the analysis of reaction time (RT) data where most of the responses cluster around some average speed but there are a few extremely slow responses (a very common shape of RT data). This produces a highly skewed distribution that is not Gaussian (normal). A rank correlation, Spearman's, analysis enables analysis of these types of data without problems caused by the violation of the assumption of normality. As with all tools that allow us to carry out statistical analysis when data violate assumptions of normality, this should be used with caution and some thought towards why the data are not normally distributed.

Analysis of Categorical Data

In our standard model of experimental design, we use a manipulated (experimental) independent variable and measure a dependent variable. The IV (or factors) typically have a small number of levels, often 2, among which participants are assigned. We can think of our IV as being defined by a categorical variable in that participants are assigned to one condition or the other. The DV is a continuous variable that we can then look for differences in the average score across conditions. The correlation analysis described above is one variation from this model where both the IV and DV are continuous measures. It can also be the case that the DV sometimes needs to be a categorical variable.

The canonical examples of categorical variables can be captured in the memorable phrase, *you can't be a little bit pregnant or a little bit dead*. These are events for which there are only two outcomes: you are, or you aren't. Measures of these kinds of variables are 'binary' in that there are two possibilities. It is also possible to have categorical variables for which there are more than two alternatives. In general, if the alternatives can be ordered in a systematic, ranked way, these will often be coded as a familiar continuous variable. But there are plenty of cases where there is a range of

options that are each independent choices. For example, one might look at some aspect of high school education and what college within Northwestern a student applied with the possible outcomes being WCAS, McCormick, Medill, or the Bienen School of Music. Here the outcome variable is categorical across four possibilities.

These approaches can very well be experimental, and the same concepts drive our ability to draw inferences from the data: did the IV affect the DV? But now we need a statistical tool that allows us to characterize how the shift in categorical choices was affected by the manipulated IV. The general approach for analyzing these data is to organize the outcomes into a **contingency table**.

As an example, consider a non-experimental study that asked athletes if they generally stretch before exercising and if they have had an injury in the past year. Whether or not they stretch is a categorical variable with two possibilities: yes/no. The same is true for whether they have had an injury in the past year: yes or no. Suppose we had data from 800 athletes. We could organize the results in the following contingency table.

	Injury	No Injury	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

The bolded data in the table reflects the actual data counts and the rightmost and bottom columns are sums of the data above. Looking at the outcome counts, we can see that the number of people with injuries is much lower in the group that stretches (55 versus 231), but we should also note that there are different numbers of participants in the stretch/no-stretch conditions (350 versus 450). To correct for this, we should calculate **rates** for all the conditions here to see if there is evidence the injury rate is different for the two stretching conditions. The rate of injury for the stretches condition is $55/350 = 15.7\%$. The rate of injury for the not stretching condition is $231/450$

= 51.3%. That is clearly a lot higher, but what we have done so far is effectively calculated the descriptive statistics for a categorical design, which are usually rates at which the outcome occurred. To draw conclusions about whether stretching causes a lower rate of injuries, we need a statistical test to identify if this difference is statistically reliable. As with all other statistical approaches, we want the familiar p-value, which is always the probability of observing this pattern of data under the null hypothesis regardless of the kind of data being evaluated. The null hypothesis here is that stretching or not are associated with injuries at the roughly the same rate overall and the observed different was somehow just luck.

There are a variety of ways to analyze a contingency table, but we will focus on one tool that is fairly flexible across common experimental designs, the chi-squared analysis or χ^2 . The X is the Greek letter 'chi' (pronounced ki) and the analysis is variously referred to by the Greek letter, the term *chi-squared* or the mixed term *chi²*. Conceptually, this analysis is based on looking at the difference between the **observed rate** of occurrence from the **expected rate** of occurrence in each of the four cells of the contingency table. The expected rate is the rate of occurrence under the null hypothesis that stretching does not matter.

It is useful to note that the injury rate in the dataset does not have to be 50% just because there are two possibilities. We can estimate the average injury rate by looking at the total number of injuries over all the participants ignoring the stretching condition (just as we did with marginal means for main effects in ANOVA). There were a total of 286 injuries out of the 800 participants, which is a rate of 35.8%. The χ^2 formula is essentially telling us that if the average rate of injury is 35.8%, what are the odds that one group would exhibit a 15.7% rate and the other group a 51.3% rate. The number of participants is critical for this calculation, so it is actually done with the numbers expected in each cell not from the rates themselves. The χ^2 is the sum of the difference between the Expected number of participants in each cell (under the null) and the Observed number of participants. For the stretches condition, at a 35.8% injury rate, we should have seen ~125

injuries, yet we only saw 55. χ^2 is actually calculated as the square of this difference divided by the expected value and so the sum of this across the four cells is:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

For this example, we would obtain a χ^2 value of 108.7, which is exceedingly improbable and since $p < .05$ (the same criterion as always), we can reject the null hypothesis and conclude that stretching did reliably affect the injury rate in this data collection sample.

Nothing about the calculation would be different if this was an experimental study where athletes were assigned to a stretching or no stretching condition and then injury rate was assessed afterwards. Of course, that would be a problematically unethical study since we would have assigned participants to a condition that we thought might cause them to become injured.

The method applies to any design where the DV is a categorical variable and works well even if there are more than two possible outcomes. The general approach is to organize the data into a contingency table of counts of each of the combinations of IV level and DV outcomes. From this, calculate the rates to be able to describe the data as descriptive statistics. The calculation of a χ^2 analysis is simple enough that you could do it with a calculator or spreadsheet, but you can also simply use standard analysis programs like R which also helpfully provide the direct estimate of the p-value.

Extending this methodology to multiple independent variables is somewhat tricky and requires digging into a set of analytical tools known as **non-parametric statistics**. These tools are useful for categorical data and situations where assumptions of normality are not met in the observed data but are not very frequently used in basic experimental science.

Advanced Statistics

Here we will briefly review a set of more complex analytic methods to provide some familiarity with the names of these approaches. We will not attempt to apply these methods in this class, but you may see references to these approaches in published research reports that you encounter as part of your background reading. These fairly common but more complex analyses include **regression**, **general linear models**, and **factor analysis**. We will also introduce the tool of meta-analysis, which aims to aggregate information from multiple published studies to help better quantify effect sizes and reliability across similar experimental approaches.

Relationships Among Multiple Variables

A common extension of research based on correlations between a single continuous independent variable and a continuous dependent variable is to collect data from a number of IVs and combine these to understand their relationship to the DV. The relationship between these IVs and the DV is assessed via a regression analysis, within a general framework of an approach called general linear modeling (GLM). The approach of using a GLM is the basis of the vast majority statistical modeling of complex datasets except for a handful of special and interesting cases where non-linear relationships are evaluated (although if you work in this area, you learn a variety of approaches for transforming non-linear data to be suitable for linear modeling).

The correlation analysis described above is a simplified case of regression where there is one predictor (IV) variable and one outcome (DV) variable. The trendlines shown in the correlation examples are calculated by a linear regression. Whereas in correlation, we get a statistical parameter r , GLM analysis will often report a statistic, r^2 , which has a useful verbal description of **percentage of variance** accounted for in the data. This is, in fact, the same r statistic we use for correlation analysis squared and can help understand what a correlation analysis is telling us. The idea behind the

phrase *percent of variance accounted for* is that the data have variance which results from a range of factors related to extraneous variables, participant variances and measurement error. Our predictor variable (IV) accounts for some of the variability in the measured variable (DV) but we acknowledge that there are many other sources of variance not accounted for. In the example of a fairly strong correlation of $r = 0.5$, $r^2 = 0.25$ meaning a quarter of the variance is accounted for and three quarters (75%) is still unknown. So even in the case of a strong relationship, there's still a lot we have not captured in the statistical model. For milder correlations in the range of 0.3, we would be happy to account for 10% of the variance. The r^2 statistic is generally reported when using any GLM (regression) analysis. Many ANOVA programs are using a GLM behind the scenes (ANOVA is also a special case of the broader regression approach) and may report the r^2 statistic as well. If so, you can use this heuristic to get a sense of how systematic your data are by how much of the variance is accounted for in your model.

Another important use of regression analysis is to explore possible causal relationships among variables even when these measures are collected in a correlational research style (without manipulating the IV). The primary way of doing this is through the statistical control of potential third variables. Instead of controlling these variables through random assignment or by holding them constant as in an experiment, the researcher instead measures them and includes them in the statistical analysis called **partial correlation**. Using this technique, researchers can examine the relationship between two variables, while statistically controlling for one or more potential third variables. You will typically see this described as *controlling for*, as in X appears to cause a change in Y controlling for Z in an analysis where the effect of Z are attempted to be controlled for to try to give an independent view of how X affects Y.

For example, assume a researcher was interested in the relationship between watching violent television shows and aggressive behavior but she was concerned that socioeconomic status (SES) might represent a third variable that is driving this relationship. In this case, she could conduct a study in

which she measures the amount of violent television that participants watch in their everyday life, the number of acts of aggression that they have engaged in, and their SES. She could first examine the correlation between violent television viewing and aggression. Let's say she found a correlation of +.35, which would be considered a moderate sized positive correlation. Next, she could use partial correlation to reexamine this relationship after statistically controlling for SES. This technique would allow her to examine the relationship between the part of violent television viewing that is independent of SES and the part of aggressive behavior that is independent of SES. If she found that the partial correlation between violent television viewing and aggression while controlling for SES was +.34, that would suggest that the relationship between violent television viewing and aggression is largely independent of SES (i.e., SES is not a third variable driving this relationship). On the other hand, if she found that after statistically controlling for SES the correlation between violent television viewing and aggression dropped to +.03, then that would suggest that SES is indeed a third variable that is driving the relationship. If, however, she found that statistically controlling for SES reduced the magnitude of the correlation from +.35 to +.20, then this would suggest that SES accounts for some, but not all, of the relationship between television violence and aggression. It is important to note that while partial correlation provides an important tool for researchers to statistically control for third variables, researchers using this technique are still limited in their ability to arrive at causal conclusions because this technique does not take care of the directionality problem and there may be other third variables driving the relationship that the researcher did not consider and statistically control.

If you continue to study more complex research methods beyond this class, you will encounter regression analysis, ANCOVA (analysis of covariance), MANOVA (multiple variable ANOVA), logistic regression, and even more complex tools such as structural equation modeling. Many of these are used in non-experimental research studies to try to increase the confidence in drawing causal conclusions from datasets where the independent variables cannot be manipulated (e.g., epidemiology, economics). It may be useful to

you then to know that these are all founded on the GLM core and developing an understanding of this multiple regression approach will help you grasp a wide range of more specialized tools used for those areas of research.

Multiple Dependent Variables

In the development of novel survey instruments, it is common to collect a set of dependent variables and look for relationships among these. This can be used to calculate both convergent and discriminant validity of the scale. For example, when Cacioppo & Petty (1982) first reported the Need for Cognition Scale—a measure of the extent to which people like to think and value thinking—they used it to measure the need for cognition for a large sample of college students, along with three other variables: intelligence, socially desirable responding (the tendency to give what one thinks is the “appropriate” response), and dogmatism. The results of this study are summarized in a correlation matrix (below) showing the correlation (Pearson’s r) between every possible pair of variables in the study.

Correlation Matrix Showing Correlations Among the Need for Cognition and Three Other Variables Based on Research by Cacioppo and Petty (1982)

	Need for cognition	Intelligence	Social desirability	Dogmatism
Need for cognition	--			
Intelligence	+.39	--		
Social desirability	+.08	+.02	--	
Dogmatism	-.27	-.23	+.03	--

For example, the correlation between the need for cognition and intelligence was +.39, the correlation between intelligence and socially desirable responding was +.02, and so on. (Only half the matrix is filled in because the other half would contain exactly the same information. Also, because the

correlation between a variable and itself is always +1.00, these values are replaced with dashes throughout the matrix.) In this case, the overall pattern of correlations was consistent with the researchers' ideas about how scores on the need for cognition should be related to these other constructs.

A more complex statistical technique for evaluating relationships among a large number of variables is the approach of **factor analysis**. Factor analysis attempts to organize the observed data as arising from a smaller number of predictor variables than were originally used. As an example, researchers Rentfrow & Gosling (2008) asked more than 1,700 university students to rate how much they liked 14 different popular genres of music. They then submitted these 14 variables to a factor analysis, which identified four distinct underlying factors. The researchers called them Reflective and Complex (blues, jazz, classical, and folk), Intense and Rebellious (rock, alternative, and heavy metal), Upbeat and Conventional (country, soundtrack, religious, pop), and Energetic and Rhythmic (rap/hip-hop, soul/funk, and electronica). The underlying idea is that the rating of the blues, jazz, classical and folk music tended to be similar to each other, so these are reduced to one underlying factor. Note that this analysis does not tell you anything about what the factor means, which is often considered a weakness of this approach, leaving it up to the authors to decide to describe this cluster as *Reflected and Complex*. The analysis provides a table of *factor loadings* (right) which indicate how well each of the observed measures relates to the inferred cluster (factor).

The way to read a factor loadings table is to treat these as similar to a correlation. Participants who liked Blues tended to like other music that was scored as highly related to the Reflected and Complex factor: Jazz, Classical and Folk. This technique is often used to distill a large dataset with many measures into a description based on a smaller number of underlying factors.

**Table of Factor Loadings of
14 Music Genres studied by
survey of undergraduate students**

Genre	Reflective and Complex	Intense and Rebellious	Upbeat and Conventional	Energetic and Rhythmic
Blues	.85	.01	-.09	.12
Jazz	.83	.04	.07	.15
Classical	.66	.14	.02	-.13
Folk	.64	.85	-.04	-.07
Rock	.17	.85	-.04	-.07
Alternative	.02	.80	.13	.04
Heavy Metal	.07	.75	-.11	.04
Country	-.06	.05	.72	-.03
Sound tracks	.01	.04	.70	.17
Religious	.23	-.21	.64	-.01
Pop	-.20	.06	.59	.45
Rap/hip-hop	-.19	-.12	.17	.79
Soul/funk	.39	-.11	.11	.69
Electronica/dance	-.02	.15	-.01	.60

Survey data analyzed through Four Varimax-Rotated Principal Components.
Based on Research by Rentfrow and Gosling (2003).

Meta-Analysis

The technique of **meta-analysis** describes a method for aggregating data across many published studies. As researchers have become more concerned about replicability and reliability of research, this tool has become increasingly popular to draw broad conclusions across many published studies rather than rely on individual reported experiments. The approach is typically

carried out in a series of steps. First, the set of relevant published research studies is identified through a systematic literature search. Then for each publication, the reported effect size is identified, sometimes together with other methodological elements to quantify the quality of the study. Studies with larger samples or lower overall variance are likely to be reporting higher quality results. The effect sizes across many studies are then combined to identify the central tendency of the magnitude of the effect of the IV on the DV. This can simply be the mean of the effect sizes, but it can also be a weighted average to give more impact from the higher quality research. This approach provides an excellent way to get a good estimate of the true size of the effect of the IV on the DV.

Meta-analysis can also be used to look at the distribution of results, which can help identify outlier studies with findings that look very different from what should be similar reports. The distribution of effect sizes across studies can also be used to identify publication bias or whether the tendency across studies supports the null hypothesis that the IV does not affect the DV. These can be very helpful approaches in controversial research areas where findings do not consistently replicate. Comparison of methodology across findings can also help identify what extraneous variables might be affecting outcomes that were imperfectly controlled in individual studies.

The primary challenge of meta-analysis is that it requires having a lot of published research available to aggregate. As a result it can only be applied in research areas where several researchers have published related results. Good estimation of effect size would be very valuable for research in newer areas so that power analysis can be used to plan samples for research, but meta-analysis does not help here. The desire for more findings to incorporate in meta-analysis has led some researchers to call for relaxing the p-value criteria for reliability to allow studies with marginally reliable findings or even null findings to be incorporated into this approach.

Exercises

Question 1. A colleague helping with a research project comes to tell us in that the latest statistical analysis, they found an $r = 2.30$? Should we be excited?

Question 2. A college administrator is choosing between 2 admissions tests, one that correlates with subsequent college performance $r = -.54$, and one for which, $r = +.45$. Which test should be preferred and why?

Question 3. Of all the statistical tests we have covered in class so far, explain which is appropriate for each of the following hypothetical designs and why:

- Older people (age in years) are found to exercise less (in hours per week) than middle aged people
- At a car dealership, people planning to buy a car are either shown a documentary on climate change or not and then they are scored as to whether they purchased an electric vehicle or gas vehicle
- Hospitalization rate (percent of people hospitalized) due to COVID infection differ in counties that voted red (Republican) or blue (Democrat) in the 2020 election
- The relationship of first year salary post-graduation with average GPA in sophomore year during college
- Number of social media posts made by people who either took research methods in college or did not
- An analysis of scores on the Oxford Happiness Scale based on participants diet (vegetarian/not) and whether they have a Facebook account.

17 Qualitative Research

The history of psychology is often described as starting with William James, an American philosopher, who wrote **The Principles of Psychology** in 1890 that essentially established psychology as a field separate from natural philosophy. Although there was a developing tradition in experimental psychology being developed by German academics, especially Wilhelm Wundt, around the same time, James' work is notable for relying primarily on the technique of **introspection**.

Introspection largely relies on self-reflection and attempting to document the process of psychological experience by focus on subjective experience. While the overview of psychology that James produced using this method still contains a very effective overview of many aspects of the field, introspection as a tool became somewhat disfavored in the early part of the 20th century. It quickly became clear that much of the process of cognition that drives behavior is not particularly accessible to description based on subjective experience. Introspection also does not allow itself to rigorous study as there is no way to resolve differences in self-reflection across people.

From a methodological standpoint, introspection, observational and other qualitative approaches were replaced by **behaviorism**, championed by BF Skinner in the 1930s. The experimental tools used in this approach were early versions of the basic methodology described here but were often mainly

applied to animal studies. Research with experimental animals avoids some of the challenges in developing operational definitions of complex human thought or behavior by keeping the dependent variables to simple and easily measured responses in animals.

The **cognitive revolution** in psychology in the 1960s can be seen as bringing human participants into experimental methodology. Research from that time and since has recognizable methodology of random assignment to different levels of an independent variable and subsequent measurement of a dependent variable. At around the same time, tools for more detailed physiological study of animal models led to the development of the field of **neuroscience**, which capitalized on a lot of the methodologies developed by behaviorists combined with sophisticated biological manipulations and measures. In Chapter 20, we will review some methods of **cognitive neuroscience** that attempts to bring those ideas back to the study of human participants again.

Within the framework presented here for psychological science, we can consider James' initial work with introspection to provide a set of terms for the constructs of psychological science. Introspection is too subjective for effective science, but we can approach these constructs by creating operational definitions to be used in rigorous scientific methods. Even in doing so, we should acknowledge that there are many areas of psychology in which we are still identifying the key constructs to be studied. In these areas, tools known as **qualitative research**, which rely more on observation, case studies, collecting stories, and structured interviews. In some cases, these data can be used to drive quantitative measures for research and in others, the observations themselves can illustrate directions in which experimental methods might be applied in the future.

Learning Objectives

1. List several ways in which qualitative research differs from quantitative research in psychology.
2. Describe the strengths and weaknesses of qualitative research in psychology compared with quantitative research.
3. Give examples of qualitative research in psychology.
4. List the various types of observational research methods and distinguish between each.
5. Describe the strengths and weakness of each observational research method.

What Is Qualitative Research?

This textbook is primarily about quantitative research, in part because most studies conducted in psychology are quantitative in nature. Quantitative researchers typically start with a focused research question or hypothesis, collect a small amount of numerical data from a large number of individuals, describe the resulting data using statistical techniques, and draw general conclusions about some large population. Although this method is by far the most common approach to conducting empirical research in psychology, there is an important alternative called qualitative research. Qualitative research originated in the disciplines of anthropology and sociology but is also used to study psychological topics. Qualitative researchers generally begin with a less focused research question, collect large amounts of relatively “unfiltered” data from a relatively small number of individuals, and describe their data using nonstatistical techniques, such as grounded theory, thematic analysis, critical discourse analysis, or interpretative phenomenological analysis. They are usually less concerned with drawing general conclusions about human behavior than with understanding in detail the experience of their research participants.

Consider, for example, a study by researcher Per Lindqvist and his colleagues,

who wanted to learn how the families of teenage suicide victims cope with their loss (Lindqvist, Johansson, Karlsson, 2008). They did not have a specific research question or hypothesis, such as, What percentage of family members join suicide support groups? Instead, they wanted to understand the variety of reactions that families had, with a focus on what it is like from their perspectives. To address this question, they interviewed the families of 10 teenage suicide victims in their homes in rural Sweden. The interviews were relatively unstructured, beginning with a general request for the families to talk about the victim and ending with an invitation to talk about anything else that they wanted to tell the interviewer. One of the most important themes that emerged from these interviews was that even as life returned to "normal," the families continued to struggle with the question of why their loved one committed suicide. This struggle appeared to be especially difficult for families in which the suicide was most unexpected.

The Purpose of Qualitative Research

The strength of quantitative research is its ability to provide precise answers to specific research questions and to draw general conclusions about human behavior. This method is how we know that people have a strong tendency to obey authority figures, for example, and that female undergraduate students are not substantially more talkative than male undergraduate students. But while quantitative research is good at providing precise answers to specific research questions, it is not nearly as good at generating novel and interesting research questions. Likewise, while quantitative research is good at drawing general conclusions about human behavior, it is not nearly as good at providing detailed descriptions of the behavior of particular groups in particular situations. And quantitative research is not very good at communicating what it is actually like to be a member of a particular group in a particular situation.

But the relative weaknesses of quantitative research are the relative strengths of qualitative research. Qualitative research can help researchers to generate

new and interesting research questions and hypotheses. The research of Lindqvist et al (2008), for example, suggests that there may be a general relationship between how unexpected a suicide is and how consumed the family is with trying to understand why the teen committed suicide. This relationship can now be explored using quantitative research. But it is unclear whether this question would have arisen at all without the researchers sitting down with the families and listening to what they themselves wanted to say about their experience. Qualitative research can also provide rich and detailed descriptions of human behavior in the real-world contexts in which it occurs. Among qualitative researchers, this depth is often referred to as “thick description” (Geertz, 1973). Similarly, qualitative research can convey a sense of what it is actually like to be a member of a particular group or in a particular situation—what qualitative researchers often refer to as the “lived experience” of the research participants. Lindqvist and colleagues, for example, describe how all the families spontaneously offered to show the interviewer the victim’s bedroom or the place where the suicide occurred—revealing the importance of these physical locations to the families. It seems unlikely that a quantitative study would have discovered this detail.

Some contrasts between qualitative and quantitative research

Qualitative	Quantitative
In-depth information about relatively few people	Less depth information with larger samples
Conclusions are based on interpretations drawn by the investigator	Conclusions are based on statistical analyses
Global and exploratory	Specific and focused

Data Collection in Qualitative Research

Data collection approaches in qualitative research are quite varied and can involve naturalistic observation, participant observation, archival data, artwork, and many other things. But one of the most common approaches, especially for psychological research, is to conduct interviews. Interviews in qualitative research can be unstructured—consisting of a small number of general questions or prompts that allow participants to talk about what is of interest to them—or structured, where there is a strict script that the interviewer does not deviate from. Most interviews are in between the two and are called semi-structured interviews, where the researcher has a few consistent questions and can follow up by asking more detailed questions about the topics that come up. Such interviews can be lengthy and detailed, but they are usually conducted with a relatively small sample. The unstructured interview was the approach used by Lindqvist and colleagues in their research on the families of suicide victims because the researchers were aware that how much was disclosed about such a sensitive topic should be led by the families, not by the researchers.

Another approach used in qualitative research involves small groups of people who participate together in interviews focused on a particular topic or issue, known as focus groups. The interaction among participants in a focus group can sometimes bring out more information than can be learned in a one-on-one interview. The use of focus groups has become a standard technique in business and industry among those who want to understand consumer tastes and preferences. The content of all focus group interviews is usually recorded and transcribed to facilitate later analyses. However, we know from social psychology that group dynamics are often at play in any group, including focus groups, and it is useful to be aware of those possibilities. For example, the desire to be liked by others can lead participants to provide inaccurate answers that they believe will be perceived favorably by the other participants. The same may be said for personality characteristics. For example, highly extroverted participants can sometimes dominate discussions within focus groups.

Data Analysis in Qualitative Research

Although quantitative and qualitative research generally differ along several important dimensions (e.g., the specificity of the research question, the type of data collected), it is the method of data analysis that distinguishes them more clearly than anything else. To illustrate this idea, imagine a team of researchers that conducts a series of unstructured interviews with people recovering from alcohol use disorder to learn about the role of their religious faith in their recovery. Although this project sounds like qualitative research, imagine further that once they collect the data, they code the data in terms of how often each participant mentions God (or a “higher power”), and they then use descriptive and inferential statistics to find out whether those who mention God more often are more successful in abstaining from alcohol. Now it sounds like quantitative research. In other words, the quantitative-qualitative distinction depends more on what researchers do with the data they have collected than with why or how they collected the data.

But what does qualitative data analysis look like? Just as there are many ways to collect data in qualitative research, there are many ways to analyze data. One general approach called **grounded theory** (Glaser & Strauss, 1967) was developed within the field of sociology in the 1960s and has gradually gained popularity in psychology. In qualitative research using grounded theory, researchers start with the data and develop a theory or an interpretation that is *grounded in* those data. They do this analysis in stages. First, they identify ideas that are repeated throughout the data. Then they organize these ideas into a smaller number of broader themes. Finally, they write a theoretical narrative—an interpretation of the data in terms of the themes that they have identified. This theoretical narrative focuses on the subjective experience of the participants and is usually supported by many direct quotations from the participants themselves.

As an example, consider a study Abrams & Curran (2009), who used the grounded theory approach to study the experience of postpartum depression symptoms among low-income mothers. Their data were the result of

unstructured interviews with 19 participants with the observed broad themes below in a table. In their research report, they provide numerous quotations from their participants, such as this one from "Destiny:"

Well, just recently my apartment was broken into and the fact that his Medicaid for some reason was canceled so a lot of things was happening within the last two weeks all at one time. So that in itself I don't want to say almost drove me mad but it put me in a funk....Like I really was depressed. (p. 357)

Table of Themes and Repeating Ideas in a Study of Postpartum Depression Among Low-Income Mothers. Based on Research by Abrams and Curran (2009).

Theme	Repeating ideas
Ambivalence	"I wasn't prepared for this baby," "I didn't want to have any more children."
Caregiving overload	"Please stop crying" "I need a break" "I can't do this anymore."
Juggling	"No time to breathe" "Everyone depends on me" "Navigating the maze"
Mothering alone	"I really don't have any help" "My baby has no father."
Real-life worry	"I don't have any money" "Will my baby be OK?" "It's not safe here"

Their theoretical narrative focused on the participants' experience of their symptoms, not as an abstract *affective disorder* but as closely tied to the daily struggle of raising children alone under often difficult circumstances.

The Quantitative-Qualitative Debate

Given their differences, it may come as no surprise that quantitative and qualitative research in psychology and related fields do not coexist in complete harmony. Some quantitative researchers criticize qualitative methods on the grounds that they lack objectivity, are difficult to evaluate in terms of reliability and validity, and do not allow generalization to people or situations other than those actually studied. At the same time, some qualitative researchers criticize quantitative methods on the grounds that they overlook the richness of human behavior and experience and instead answer simple questions about easily quantifiable variables.

However many researchers agree that the two approaches can and should be combined into what has come to be called mixed-methods research (Todd, Nerlich, McKeown, Clarke, 2004). One approach to combining quantitative and qualitative research is to use qualitative research for hypothesis generation and quantitative research for hypothesis testing. Again, while a qualitative study might suggest that families who experience an unexpected suicide have more difficulty resolving the question of why, a well-designed quantitative study could test a hypothesis by measuring these specific variables in a large sample. A second approach to combining quantitative and qualitative research is referred to as triangulation. The idea is to use both quantitative and qualitative methods simultaneously to study the same general questions and to compare the results. If the results of the quantitative and qualitative methods converge on the same general conclusion, they reinforce and enrich each other. If the results diverge, then they suggest an interesting new question: Why do the results diverge and how can they be reconciled?

Using qualitative research can often help clarify quantitative results via triangulation. Trenor, Yu, Waight, Zerda, and Sha (2008) investigated the experience of female engineering students at a university. In the first phase, female engineering students were asked to complete a survey, where they rated a number of their perceptions, including their sense of belonging. Their results were compared across the student ethnicities, and statistically, the

various ethnic groups showed no differences in their ratings of their sense of belonging. One might look at that result and conclude that ethnicity does not have anything to do with one's sense of belonging. However, in the second phase, the authors also conducted interviews with the students, and in those interviews, many minority students reported how the diversity of cultures at the university enhanced their sense of belonging. Without the qualitative component, we might have drawn the wrong conclusion about the quantitative results. This example shows how qualitative and quantitative research work together to help us understand human behavior.

What Is Observational Research?

The term observational research is used to refer to several different types of non-experimental studies in which behavior is systematically observed and recorded. The goal of observational research is to describe a variable or set of variables. More generally, the goal is to obtain a snapshot of specific characteristics of an individual, group, or setting. As described previously, observational research is non-experimental because nothing is manipulated or controlled, and as such we cannot arrive at causal conclusions using this approach. The data that are collected in observational research studies are often qualitative in nature but they may also be quantitative or both (mixed-methods). There are several different types of observational methods that will be described below.

Naturalistic Observation

Naturalistic observation is an observational method that involves observing people's behavior in the environment in which it typically occurs. Thus, naturalistic observation is a type of field research (as opposed to a type of laboratory research). Jane Goodall's famous research on chimpanzees is a classic example of naturalistic observation. Dr. Goodall spent three decades observing chimpanzees in their natural environment in East Africa. She

examined such things as chimpanzee's social structure, mating patterns, gender roles, family structure, and care of offspring by observing them in the wild. However, naturalistic observation could more simply involve observing shoppers in a grocery store, children on a school playground, or psychiatric inpatients in their wards. Researchers engaged in naturalistic observation usually make their observations as unobtrusively as possible so that participants are not aware that they are being studied. Such an approach is called disguised naturalistic observation. Ethically, this method is considered to be acceptable if the participants remain anonymous and the behavior occurs in a public setting where people would not normally have an expectation of privacy. Grocery shoppers putting items into their shopping carts, for example, are engaged in public behavior that is easily observable by store employees and other shoppers. For this reason, most researchers would consider it ethically acceptable to observe them for a study. On the other hand, one of the arguments against the ethicality of the naturalistic observation of "bathroom behavior" discussed earlier in the book is that people have a reasonable expectation of privacy even in a public restroom and that this expectation was violated.

In cases where it is not ethical or practical to conduct disguised naturalistic observation, researchers can conduct undisguised naturalistic observation where the participants are made aware of the researcher presence and monitoring of their behavior. However, one concern with undisguised naturalistic observation is reactivity. Reactivity refers to when a measure changes participants' behavior. In the case of undisguised naturalistic observation, the concern with reactivity is that when people know they are being observed and studied, they may act differently than they normally would. This type of reactivity is known as the Hawthorne effect. For instance, you may act much differently in a bar if you know that someone is observing you and recording your behaviors and this would invalidate the study. So disguised observation is less reactive and therefore can have higher validity because people are not aware that their behaviors are being observed and recorded. However, we now know that people often become used to being observed and with time they begin to behave naturally in the researcher's

presence. In other words, over time people habituate to being observed. Think about reality shows like Big Brother or Survivor where people are constantly being observed and recorded. While they may be on their best behavior at first, in a fairly short amount of time they are flirting, having sex, wearing next to nothing, screaming at each other, and occasionally behaving in ways that are embarrassing.

Participant Observation

Another approach to data collection in observational research is **participant observation**. In participant observation, researchers become active participants in the group or situation they are studying. Participant observation is very similar to naturalistic observation in that it involves observing people's behavior in the environment in which it typically occurs. As with naturalistic observation, the data that are collected can include interviews (usually unstructured), notes based on their observations and interactions, documents, photographs, and other artifacts. The only difference between naturalistic observation and participant observation is that researchers engaged in participant observation become active members of the group or situations they are studying. The basic rationale for participant observation is that there may be important information that is only accessible to, or can be interpreted only by, someone who is an active participant in the group or situation. Like naturalistic observation, participant observation can be either disguised or undisguised. In disguised participant observation, the researchers pretend to be members of the social group they are observing and conceal their true identity as researchers.

In a famous example of disguised participant observation, Leon Festinger and his colleagues infiltrated a doomsday cult known as the Seekers, whose members believed that the apocalypse would occur on December 21, 1954. Interested in studying how members of the group would cope psychologically when the prophecy inevitably failed, they carefully recorded the events and reactions of the cult members in the days before and after the supposed end

of the world. Unsurprisingly, the cult members did not give up their belief but instead convinced themselves that it was their faith and efforts that saved the world from destruction. Festinger and his colleagues later published a book about this experience, which they used to illustrate the theory of cognitive dissonance (Festinger, Riecken, Schachter, 1956).

In contrast with undisguised participant observation, the researchers become a part of the group they are studying and they disclose their true identity as researchers to the group under investigation. Once again there are important ethical issues to consider with disguised participant observation. First no informed consent can be obtained and second deception is being used. The researcher is deceiving the participants by intentionally withholding information about their motivations for being a part of the social group they are studying. But sometimes disguised participation is the only way to access a protective group (like a cult). Further, disguised participant observation is less prone to reactivity than undisguised participant observation.

Rosenhan's study (1973) of the experience of people in a psychiatric ward would be considered disguised participant observation because Rosenhan and his pseudopatients were admitted into psychiatric hospitals on the pretense of being patients so that they could observe the way that psychiatric patients are treated by staff. The staff and other patients were unaware of their true identities as researchers.

Another example of participant observation comes from a study by sociologist Amy Wilkins on a university-based religious organization that emphasized how happy its members were (Wilkins, 2008). Wilkins spent 12 months attending and participating in the group's meetings and social events, and she interviewed several group members. In her study, Wilkins identified several ways in which the group "enforced" happiness—for example, by continually talking about happiness, discouraging the expression of negative emotions, and using happiness as a way to distinguish themselves from other groups.

One of the primary benefits of participant observation is that the researchers are in a much better position to understand the viewpoint and experiences

of the people they are studying when they are a part of the social group. The primary limitation with this approach is that the mere presence of the observer could affect the behavior of the people being observed. While this is also a concern with naturalistic observation, additional concerns arise when researchers become active members of the social group they are studying because that they may change the social dynamics and/or influence the behavior of the people they are studying. Similarly, if the researcher acts as a participant observer there can be concerns with biases resulting from developing relationships with the participants. Concretely, the researcher may become less objective resulting in more experimenter bias.

Structured Observation

Another observational method is **structured observation**. Here the investigator makes careful observations of one or more specific behaviors in a particular setting that is more structured than the settings used in naturalistic or participant observation. Often the setting in which the observations are made is not the natural setting. Instead, the researcher may observe people in the laboratory environment. Alternatively, the researcher may observe people in a natural setting (like a classroom setting) that they have structured some way, for instance by introducing some specific task participants are to engage in or by introducing a specific social situation or manipulation.

Structured observation is very similar to naturalistic observation and participant observation in that in all three cases researchers are observing naturally occurring behavior; however, the emphasis in structured observation is on gathering quantitative rather than qualitative data. Researchers using this approach are interested in a limited set of behaviors. This allows them to quantify the behaviors they are observing. In other words, structured observation is less global than naturalistic or participant observation because the researcher engaged in structured observations is interested in a small number of specific behaviors. Therefore, rather than recording everything that happens, the researcher only focuses on very specific behaviors of interest.

As an example, researchers Robert Kraut and Robert Johnston wanted to study bowlers' reactions to their shots, both when they were facing the pins and then when they turned toward their companions (Kraut & Johnston, 1979). But what "reactions" should they observe? Based on previous research and their own pilot testing, Kraut and Johnston created a list of reactions that included "closed smile," "open smile," "laugh," "neutral face," "look down," "look away," and "face cover" (covering one's face with one's hands). The observers committed this list to memory and then practiced by coding the reactions of bowlers who had been videotaped. During the actual study, the observers spoke into an audio recorder, describing the reactions they observed. Among the most interesting results of this study was that bowlers rarely smiled while they still faced the pins. They were much more likely to smile after they turned toward their companions, suggesting that smiling is not purely an expression of happiness but also a form of social communication.

In another example (this one in a laboratory environment), Dov Cohen and his colleagues had observers rate the emotional reactions of participants who had just been deliberately bumped and insulted by a confederate after they dropped off a completed questionnaire at the end of a hallway. The confederate was posing as someone who worked in the same building and who was frustrated by having to close a file drawer twice in order to permit the participants to walk past them (first to drop off the questionnaire at the end of the hallway and once again on their way back to the room where they believed the study they signed up for was taking place). The two observers were positioned at different ends of the hallway so that they could read the participants' body language and hear anything they might say. Interestingly, the researchers hypothesized that participants from the southern United States, which is one of several places in the world that has a "culture of honor," would react with more aggression than participants from the northern United States, a prediction that was in fact supported by the observational data (Cohen, Nisbett, Bowdle, Schwarz, 1996).

When the observations require a judgment on the part of the observers—as in

the studies by Kraut and Johnston and Cohen and his colleagues—a process referred to as coding is typically required. Coding generally requires clearly defining a set of target behaviors. The observers then categorize participants individually in terms of which behavior they have engaged in and the number of times they engaged in each behavior. The observers might even record the duration of each behavior. The target behaviors must be defined in such a way that guides different observers to code them in the same way. Researchers are expected to demonstrate the interrater reliability of their coding procedure by having multiple raters code the same behaviors independently and then showing that the different observers are in close agreement. Kraut and Johnston, for example, video recorded a subset of their participants' reactions and had two observers independently code them. The two observers showed that they agreed on the reactions that were exhibited 97% of the time, indicating good interrater reliability.

One of the primary benefits of structured observation is that it is far more efficient than naturalistic and participant observation. Since the researchers are focused on specific behaviors this reduces time and expense. Also, often times the environment is structured to encourage the behaviors of interest which again means that researchers do not have to invest as much time in waiting for the behaviors of interest to naturally occur. Finally, researchers using this approach can clearly exert greater control over the environment. However, when researchers exert more control over the environment it may make the environment less natural which decreases external validity. It is less clear for instance whether structured observations made in a laboratory environment will generalize to a real-world environment. Furthermore, since researchers engaged in structured observation are often not disguised there may be more concerns with reactivity.

Quantifying Qualitative Data

From any qualitative source, from observations or interviews, behaviors can be coded, scored and counted. This process essentially turns the data

obtained in a nominally qualitative method into quantitative data that can be used to support statistical analysis using the tools we have covered previously in this text. Doing so requires establishing a coding system targeting the behavior in question and using that to calculate a quantitative measure of that behavior, even something as simple as counting occurrences.

For example, a researcher studying relationship satisfaction among married couples might observe in open-ended interviews that couples having difficulty tend to show non-verbal expressions of contempt, such as eye-rolling, during discussions of dispute. This observational research might lead to the hypothesis that the rate of these expressions of contempt is correlated with scores of overall marriage satisfaction. Frequencies of these kinds of expressions could then be counted from recorded interactions and correlated with scores of relationship satisfaction. This would allow a quantitative evaluation of the strength of the association between these two measures. Of course, this would still be non-experimental research as neither of the variables evaluated were manipulated directly by the experimenter.

Qualitative research often focuses on observations in very unstructured contexts to generate hypotheses. This approach may be followed by other non-experimental approaches based on more qualitative measures related to the initial observations. Because of the challenges inherent to drawing conclusions from non-experimental studies, these studies might be followed up by experimental methods with carefully controlled conditions and random assignment to levels of a manipulated independent variable. Note that at each step along this kind of research plan, we might lose some aspects of external validity in exchange for greater internal validity and confidence in the conclusions obtained.

Key Takeaways

- Qualitative research is an important alternative to quantitative research in psychology. It generally involves asking broader research questions, collecting more detailed data (e.g., interviews), and using non-statistical analyses.
- Many researchers conceptualize quantitative and qualitative research as complementary and advocate combining them. For example, qualitative research can be used to generate hypotheses and quantitative research to test them.
- There are several different approaches to observational research including naturalistic observation, participant observation, structured observation, case studies, and archival research.
- Naturalistic observation is used to observe people in their natural setting; participant observation involves becoming an active member of the group being observed; structured observation involves coding a small number of behaviors in a quantitative manner; case studies are typically used to collect in-depth information on a single individual; and archival research involves analyzing existing data.

Exercises

Question 1. Qualitative research is often used in a research program as a first or preliminary study before research with experimental methods and manipulated variables. Why is this?

Question 2. Why might we use a qualitative research technique to study parent-child interactions and aggressive behavior in preschoolers? Outline what such a study might look like.

Question 3. In a diary study of conflict resolution in relationships, the research protocol could involve daily documentation of any conflict experiences and how they were or were not resolved. What kinds of events might be coded for and counted in a study of this type?

18 Ethics 2: RCR

In 1998 a medical journal called *The Lancet* published an article of interest to many psychologists. The researchers claimed to have shown a statistical relationship between receiving the combined measles, mumps, and rubella (MMR) vaccine and the development of autism—suggesting furthermore that the vaccine might even cause autism. One result of this report was that many parents decided not to have their children vaccinated, which of course put them at higher risk for measles, mumps, and rubella. However, follow-up studies by other researchers consistently failed to find a statistical relationship between the MMR vaccine and autism—and it is widely accepted now in the scientific community that there is no relationship. In addition, several more serious problems with the original research were uncovered. Among them were that the lead researcher stood to gain financially from his conclusions because he had patented a competing measles vaccine. He had also used biased methods to select and test his research participants and had used unapproved and medically unnecessary procedures on them. In 2010 *The Lancet* retracted the article, and the lead researcher's right to practice medicine was revoked (Burns, 2010).

However, the idea that there were scientific concerns about the safety of vaccines persisted in popular understanding and became a key part of skepticism about new mRNA vaccines that protect against the COVID-19

virus. Millions of people subsequently refused to use the vaccine and as a result many thousands of people died unnecessarily. The damage caused by the original unethical report is likely to be higher than any other instance of scientific fraud in human history.

In the first chapter on research ethics, the focus was how to carry out scientific research in a manner that follows current expectations for best practice. We might summarize the core idea as “be nice to research participants.” Treat them with respect and design research to have value for the world (beneficence) and to make these benefits broadly available (justice). These ideas are then reflected in research processes related to obtaining informed consent and working with the Institutional Review Board as an external monitor of regulatory compliance with best practices.

The area described as **Responsible Conduct of Research (RCR)** reflects carrying out the scientific process in a fair and ethical manner related to the integrity of research and fairness in assigning credit to the researchers involved in the research process. The capitalization of the term is due to fairly recent changes in training of scientists requiring explicit engagement with these issues, especially at the student level (graduate students and undergraduate researchers). The focus on these specific topics has been driven by funding agencies (NIH, NSF) who seek to improve the reliability and quality of the research process.

Responsible Conduct of Research

The core ideas can be expressed simply: Don't Lie, Cheat or Steal. The main topic of RCR is to detail how these kindergarten ethics ideas apply to research processes in psychological science. Compliance with RCR principles is aimed to maintain the highest level of integrity in research processes so that the scientific community can rely on and trust the results of our scientific work. We have previously discussed the problem of the Type 1 error, a false claim of an effect among variables in research that turns out to be inaccurate. This can happen due to poor design or unexpected problems with extraneous variables. It can also happen due to an integrity violation where researchers do not follow best RCR practices. False claims that affect people's behavior based on the study can have substantial negative effects on society, such as with the example of vaccines. Unfortunately, identifying and attempting to retract false claims appears to have a side effect of reducing non-scientists confidence in science overall, which can also lead to massive societal costs, for example due to skepticism about climate change.

Part of the motivation to increase awareness and training in RCR was due to the acknowledgment of external pressures on scientists engaged with the research processes. Successful science can produce substantial rewards for researchers including employment, promotion and access to research funding. Carrying out an unsuccessful research project is then costly in both time and opportunity loss that could reduce access to these rewards. The main content of this text is to illustrate the methods for carrying out research with the most rigor and care possible, but under external pressures, some researchers have failed to adhere to rigorous methods leading to results that are incorrect and/or retracted.

RCR Failure: Fraud

An obvious way that integrity can be violated in research practice is the wholesale fabrication of data, which is then presented as if it were properly collected. This is fairly rare, although not unheard of, in scientific work as the majority of modern scientific work is done in teams of researchers. This type of violation is so blatant and obvious that it is unlikely not to be known to other members of a collaborative research team. Purely fraudulent findings are also unlikely to be useful as part of a research program to drive subsequent research, which is often an important part of the general operation of a research group. Being found responsible for research fraud also effectively ends a scientist's research career, leading to immediate dismissal from the university or institution at which they work and a future bar on any external funding support.

RCR Failure: Falsification

A more pernicious issue in maintaining rigor in research processes relates to **falsification** of research findings, which covers a range of inappropriate data handling methods that lead to presentation of a false conclusion. The simplest of these is to exclude data collected that does not support the researcher's hypothesis. If we hypothesize that experimental condition A leads to higher scores on our DV than condition B, we can simply exclude all the low scorers from condition A (or high scores from B) and obtain an apparently statistically reliable result. This "data selection" is an obvious failure of research integrity and is treated in the same manner as wholesale fabrication.

There are subtler ways that aspects of falsification can creep into research. Scoring of subjective ratings of performance might not be done in a completely blind manner. Performance hints or clues could be given to participants in one condition. Bias could be covertly embedded in the task instructions or context. Participant recruiting could embed bias in assignment

to conditions if not done properly randomly. These subtler issues are seen as problematic due to being difficult or nearly impossible to detect in the report of a completed research project. They could even be created accidentally by researchers who are simply so focused on research success that they implicitly deviate from best practice. A goal of RCR training for all lab personnel is to make everybody on the research team aware of these potential failures to provide checks on both their own work and the work of their collaborators.

The process of publishing research depends on peer-review of the methods, results and conclusions of a research project. Unfortunately, data handling problems are not visible to a peer reviewer, so this process does not effectively protect against RCR problems. In many modern journals, authors are encouraged or even required to publish their dataset in a publicly available location to support their results. However, even this may not protect against data selection if the publicly available data has already had the inconsistent data excluded.

A university or large research institution will typically have an office charged with evaluating scientific processes to assure compliance with best practices in research integrity, an **Office of Research Integrity**. If a concern is raised about a specific researcher or team, this office is charged with investigating and determining if an integrity violation has occurred. This investigation generally takes the form of an audit of research practices, review of raw data, preliminary analysis and evaluation of as many steps of the core research process as possible. These investigations can be complex and time-consuming as active research processes among a team of collaborators can often be fast-moving and sometimes important decisions are made quickly without immediate realization of how consequential they are.

Various recommendations have been put forth to improve the general process of research to maintain the highest levels of compliance with best practice. Some of these involve slowing the pace of research. For example, **pre-registration** of all research studies by reporting methodology, recruiting and planned analysis in advance of formal data collection. These obviously

improve integrity but unfortunately can actually exacerbate the problem of external pressure to produce successful results by putting methodologically rigorous labs at a disadvantage in competition with labs that move faster and less carefully. Another approach is to document and record as much of the research process as possible so that if a question about integrity is raised later, an audit can verify if a problem occurs. One way to accomplish is to work as if there were cameras recording every aspect of the research process in the laboratory so that everybody possibly biased decision about participant exclusion, assignment to conditions or scoring could be evaluated later.

RCR Failures: Plagiarism and Research Privacy

One of the challenges with maintaining a robust record of all research practices is that some of the external pressure on scientists comes from competition to obtain an important discovery first. This problem is most evident in research fields like drug discovery where establishing the effectiveness of a new pharmacological agent can produce a patentable discover worth as much as a billion dollars. Large financial rewards for discoveries are exceedingly rare in psychological science but being the first to discover, name or characterize a novel aspect of psychology can be very rewarding in career advancement and scientific fame. As a result, cases do occur where multiple labs are considering very similar hypotheses and essentially racing to complete and publish their research project first. The lab that wins the race will accomplish a high-impact publication and lasting credit for the idea whereas the lab that finishes second will lament having gotten scooped.

For research being carried out in a context of this kind of competition, researchers will often work with a high degree of privacy about the research being carried out. Unfortunately, protecting research methods by operating in secrecy does not generally support the ability to provide oversight of those research practices. Thus, the areas with the most pressure to produce are also often the most difficult to verify. This is a difficult problem to solve,

and the most common current approach depends entirely on training of all research staff.

This issue is where the concept of **plagiarism** plays a more visible role in ethical research practice. The most common form of the issue of plagiarism familiar to students has to do with copying another's words or ideas and claiming them as your own. This is rare in scientific publication as the majority of the scientific record is easily available to all. In addition, best practices in science are to thorough review background research and cite the relevant research to support the latest findings. Once findings are published, the problem of claiming credit for another researchers' idea is rare.

However, before publication of a novel finding, there is a risk of another research group finding out about a novel methodological approach to research and then appropriating this idea for their own without credit. This is an obvious integrity violation that can be difficult to deal with.

A particularly famous example historically is the famous work of Francis Crick and James Watson to identify the double-helix structure of the DNA molecule. They had been working on this problem for some time, as had several other large research labs, all of which were working secretly in order to be first to solve the problem. Crick and Watson were also serving as reviews for research grants and in that role saw preliminary data obtained by Rosalind Franklin using x-ray crystallography of DNA that was consistent with the idea of a double helix. That directed their subsequent work to show that DNA was constructed that way, which led to substantial scientific fame and a Nobel prize. It was not until decades later that Franklin's contribution to this discovery was fully appreciated and properly acknowledged. The ethical issues in this case are quite complex as there is no doubt that Crick and Watson developed completely novel methods and tools to come to their conclusion. When their Nobel prize was awarded, Franklin had passed away due to an unfortunately young case of ovarian cancer and since that award is not given posthumously, she would not likely have been included in any event. There are reports that she was even offered authorship on the original paper but declined (her own findings were published simultaneously).

However, it is also clear that use of pre-publication data from another research lab without their knowledge or permission is clearly not following best practice for ethical research.

RCR: Authorship

Credit for scientific findings is generally reflected in participating in authoring the scientific report. Surprisingly, the rules for who is officially an author on a published report are not completely clear and consistent across all domains of science, or even all subdomains within psychology. The APA provides guidance that researchers who provide **substantial intellectual contribution** to a project should receive authorial credit. However, *substantial intellectual contribution* is not defined. Cases where a key idea is appropriated from another research group and used without credit are clear violations of this policy.

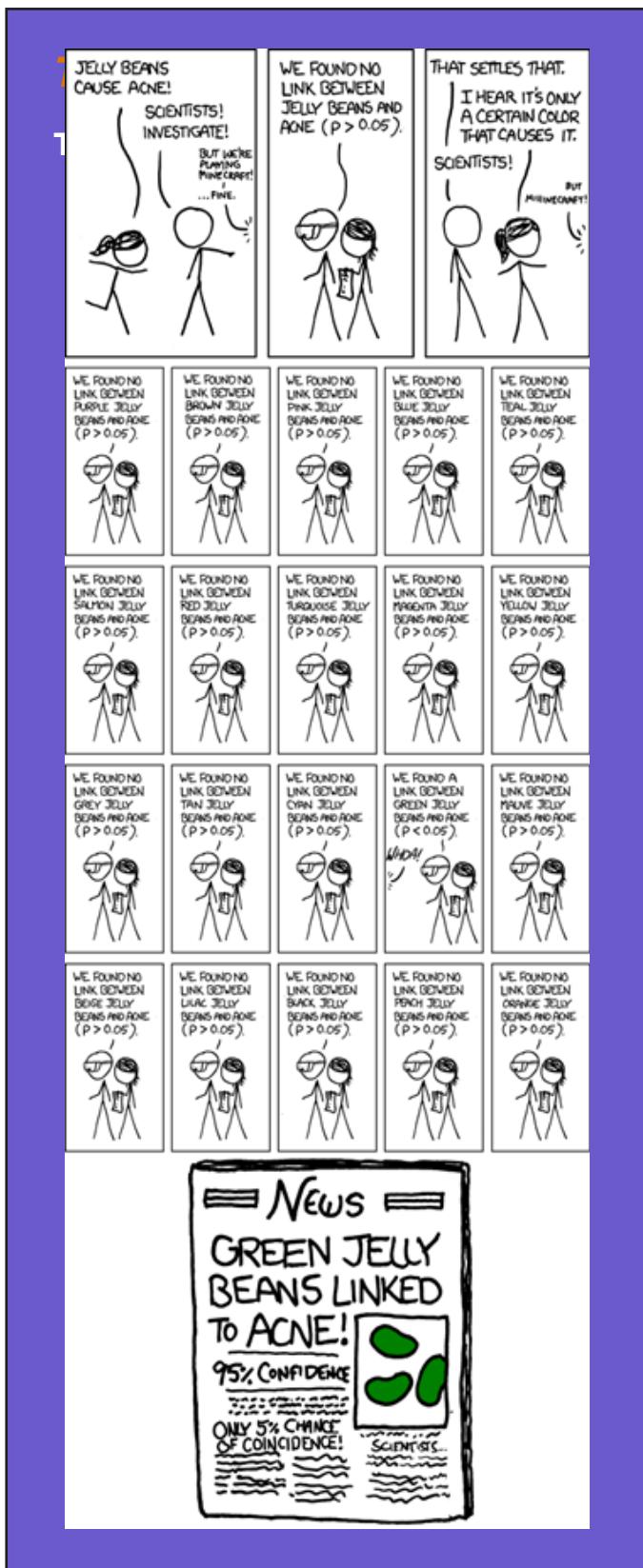
More complicated are questions about a research result that comes out of a lab with a number of staff members, graduate students and/or undergraduate student researchers. Some levels of participation in the project are considered to merit co-authorship, but in other cases, there might merely be a mention in the Acknowledgments section of the published paper. There are no fixed rules for how this is decided, and guidelines vary across individual laboratories and also across different subdomains of research.

Even trickier is the question of who takes on the coveted role of lead or first author of a manuscript reporting the results of a project done by a collaborating group. Again, there are no fixed rules for this, but it is generally up to the leader of the research team to set a clear policy that is known to all research personnel in advance and to follow that policy without bias or favoritism. The research team leader is known as the PI for Principal Investigator (almost always by the acronym to avoid the common misuse of the homonym *principle*). In most university laboratories, the PI is the professor who runs the lab, even if the research is in practice generally organized by a co-investigator such as a post-doctoral researcher (person

with a Ph.D. but is hired by the PI not the university) or a junior faculty member. Difficult situations can emerge where there is interpersonal conflict among research team members, especially if a graduate student and supervising faculty member have differing opinions about credit. One goal of improving RCR guidelines is to avoid conflict around this issue and minimize or eliminate conditions where a lead researcher appropriates credit for work done by laboratory members

RCR: Analytic Flexibility

An issue that has gotten increased attention in modern scientific practice related to questions about reproducibility of research is the potential problems associated with a flexible approach to data analysis. In its more extreme forms, this produces the phenomenon known as **p-hacking** where a large number of analytic approaches



are tried until one of them produces a result that meets the traditional criterion of $p < .05$. To understand the math behind this, remember that the p -value associated with an analysis is technically the probability of observing the data pattern under the null hypothesis. That is the same as saying that if there was no effect and all the DV measurements will random numbers drawn from the same distribution, what are the odds it would accidentally look like there was an effect? The standard criterion of .05 is effectively a 1:20 odds measure. So, if we had a true null effect but ran a study 20 times, we might accidentally observe what looks like a real result. If you are familiar with conditional probability math, you'll know that it's actually only about 65% likely to get a false positive in this case, but the broader point that multiple rounds of data collection or multiple analysis techniques essentially mean that we are not effectively meeting the standard criterion.

There are statistical tools for correctly adjusting complex analysis of data where a large number of multiple comparisons are being considered. An example of this is where we might have a large number of measures on a large-scale survey and search for all possible relationships among all the variables. The conjoint probability of a false finding across tests is no longer .05 and tools such as **Bonferroni correction** should be applied to maintain rigor. The integrity concern arises from the fact that it cannot be easily determined from a published finding if the authors attempted a wide range of very different types of analysis but only published the one that produced an effect. A related phenomenon informally called the file-drawer effect is where a research group might have attempted a study multiple times and discarded all the data from null findings and selected only an unusual successful version of the experiment to present.

The real problem with issues related to analytic flexibility is that the normal process of experimental research often requires researchers to approach their project with a certain amount of creativity and flexibility. In particularly novel research, it can take several attempts to construct the best operational definition of a tricky construct, which will produce a series of unsuccessful experiments before a successful one. In other cases, the DV will not exhibit

the exact effect predicted but another approach appears to show a robust finding (e.g., a recoding of the DV, or a related measure). As a result, researchers working with a high degree of rigor and integrity can still end up questioning their own processes as to whether they have accidentally biased their results through flexible procedures.

The idea of pre-registering experimental design and analysis plans is aimed at avoiding false positive reporting through analytic flexibility. This is highly effective, but has the unfortunate side effect of potentially eliminating report of interesting but unexpected findings that would be identified through more creative methods. Where possible, the best practice is to internally replicate an unexpected finding, that is, publish a report with multiple studies showing a consistent pattern. If flexibility was required for the first finding but it is a true effect, it should replicate in subsequent research. This approach is not easily applied in all areas of psychological science, however. Studies with expensive imaging methods, based on limited patient samples, or unique large-scale population surveys cannot be quickly or easily repeated. In those cases, considered analysis plans are important for the researchers and in addition, interested readers of the reports should likely look for parallel similar findings or subsequent research replications.

RCR: Conflict of Interest

In standard reviews of RCR-related issues, the potential for problems related to **conflict of interest** is also covered. This is relatively rare in experimental psychology research but is a general concern in areas such as health, education, marketing, and other applied research domains. The core of the idea is simple. If a physician is contracted to carry out part of an efficacy study on a new pharmacological agent that is potentially worth a billion dollars if the drug reliable works, that research is being done under extreme external pressure. If the lead researcher has a financial interest in the company making the drug, great care needs to be taken to avoid any possible implicit bias in research procedures that might undercut the validity of the

findings. In these cases, extremely rigorous double-blind procedures are used with active external oversight to guarantee integrity. And unfortunately, there are still cases found later where these processes failed.

Most psychological science research does not assess hypotheses with large financial implications. These studies do sometimes have implications for political policies, educational practices or some health interventions. In particular, this can occur when research funding is provided by a private foundation that seeks to advance a specific idea or agenda. To maintain integrity, researchers are required to disclose all financial support to university oversight and to the journal editors when research is published. This allows identification of possible conflicts of interest where, for example, research funding was contingent on a successful finding, which is not a good circumstance to foster best practices for integrity.

Tension between Ethics and Science

Now that the basic structure of best practice for ethical research (Chapter 8) and ethical conduct of science (above) has been reviewed, it is useful to note a basic tension between best practices in research ethics and maximizing the internal validity of experimental research. It may already be clear that ethical conflict in psychological research is unavoidable. Because there is little, if any, psychological research that is completely risk-free, there will almost always be a conflict between risks and benefits. Research that is beneficial to one group (e.g., the scientific community) can be harmful to another (e.g., the research participants), creating especially difficult tradeoffs. We have also seen that being completely truthful with research participants can make it difficult or impossible to conduct scientifically valid studies on important questions.

Of course, many ethical conflicts are fairly easy to resolve. Nearly everyone would agree that deceiving research participants and then subjecting them to physical harm would not be justified by filling a small gap in the research literature. But many ethical conflicts are not easy to resolve, and competent and well-meaning researchers can disagree about how to

resolve them. Consider, for example, an actual study on “personal space” conducted in a public men’s room (Middlemist, Knowles, & Matter, 1976). The researchers secretly observed their participants to see whether it took them longer to begin urinating when there was another man (a confederate of the researchers) at a nearby urinal. While some critics found this to be an unjustified assault on human dignity (Koocher, 1977), the researchers had carefully considered the ethical conflicts, resolved them as best they could, and concluded that the benefits of the research outweighed the risks (Middlemist, Knowles, & Matter, 1977). For example, they had interviewed some preliminary participants and found that none of them was bothered by the fact that they had been observed.

The point here is that although it may not be possible to eliminate ethical conflict completely, it is possible to deal with it in responsible and constructive ways. In general, this means thoroughly and carefully thinking through the ethical issues that are raised, minimizing the risks, and weighing the risks against the benefits. It also means being able to explain one’s ethical decisions to others, seeking feedback on them, and ultimately taking responsibility for them.

Several of the sources of risks to participants are both very common and very mild. Keeping participants blind to the experimental conditions has an element of deception built into the design. Maintaining a signed consent form and records of compensation (payment) to participants creates a privacy risk that somebody may discover they participated in the research study. The time spent completing the procedure is an inconvenience and even survey completion or simple cognitive tasks can be either cognitively challenging or boring. For these common design elements, providing advance information through the informed consent process and fairly compensating participants for their time addresses these mild issues.

However, it is worth acknowledging areas where the best experimental research practices become impossible due to ethical concerns. A number of these arise from research based on **interventions** aimed to provide a benefit to the participant as well as advancing science. A clear example is research

aimed to establish the effectiveness of life-saving medical interventions. When research on drugs to treat and cure AIDS were being developed, it quickly became clear that while the best research design is a double-blind randomized clinical trial using a placebo, that meant condemning the control group participants to poor health outcomes or even death. The same is true for treatments of life-threatening cancer.

In these cases, it was decided that control groups would not be included in research designs and the treatment condition, which would be all participants, would be compared to population based outcomes of the disease. This is a weaker research practice because the sampling may well be biased depending on availability of the clinical trial and who enrolls is being compared with a broader population sample. However, it was decided that the ethical issues created for researchers and participants if there was true random assignment to control conditions far exceeded the added scientific value. And it might also be noted that maximally effective research practices are primarily needed to detect subtle effect sizes. If we optimistically hope that these medical treatments are having a large effect on health improvements or preventing death, a technically weaker research approach is still enough to establish efficacy.

The same concern about access to a benefit comes up in other areas of public policy research where social or educational interventions are deployed to help a segment of the population. In order to show a causal relationship of the intervention to improved outcomes, the intervention needs to be unavailable to a control group who do not benefit. For interventions that are expensive to implement, the need to show efficacy results in a need to be slightly unfair to participants in the control group. A standard practice in these studies is to make the intervention available to the control group after assessing the effectiveness of the intervention.

The problem of assignment to the control group was also a persistent issue with scientific research throughout the COVID-19 pandemic. The ability to truly establish the quantitative effectiveness of mask wearing on slowing the spread of the virus would have required random assignment of participants

to non-mask-wearing conditions, which might have been life threatening. Instead, research in this area depended on non-experimental observations across mask wearing conditions that occurred based on personal choice or local culture. This led to widely varying estimates of mask effectiveness and criticism of the scientists attempting to study this. The problem was compounded by the fact that well-done science is not necessarily a fast process and there were demands for a more rapid answer. In addition, the estimates of mask effectiveness indicated it was not a huge effect at the individual level, although small effects over a large number of people showed robust positive epidemiological effects once that data could be collected.

HIPAA and PHI

Research related to COVID epidemiology also exposed an aspect of health-based research familiar to scientists who work in medical contexts (e.g., Neuropsychology, Chapter 20) known as HIPAA. **HIPAA stands for the Health Insurance Portability and Accountability Act**, a law passed to protect research participants from negative consequences associated with potential loss of privacy in health research. The insurance related issue was the fact that health insurance was once provided by companies who could deny coverage based on health status, known as pre-existing conditions. As a consequence, if you were in a clinical trial testing an AIDS treatment, you could potentially lose your health insurance for either having or being at risk for AIDS. HIPAA added oversight to handling of **Protected Health Information (PHI)** to improve the ethical practice of health-based science which affects a great deal of medical research and some psychological studies.

In the normal course of research, scientists are expected to protect the privacy of participants as much as possible but in standard practice this can not be done perfectly, especially with respect to the fact of participation. Participants might be seen coming to the research laboratory by others. While approved research staff are charged with carrying out research

procedures, other people around the lab or research collaborators might become aware of details. Oversight of research practices with respect to informed consent or financial compensation methods may leave information about participation available. If research processes have need for keeping information about PHI, HIPAA applies to the research protocol and requires a series of improved information protection procedures aimed to reduce risk for participants. A full discussion of these methods is beyond the scope of this text but awareness of these special cases highlights more challenging elements of ethical research practice.

A common discussion/misconception related to the COVID pandemic might be useful as illustration. Asking people about their COVID vaccination status was in some cases thought to be a violation of HIPAA due to the need to reveal what would normally be PHI. However, self-disclosing PHI is not covered by HIPAA. It would only be relevant if someone were to access your medical records directly without your consent. Asking somebody about PHI might be rude, but it not technically an ethical violation. It should also be noted that ethical issues like this are never thought to have clear black-and-white answers in all cases. It would not be unreasonable for a decision to be made that prioritizes public health and the spread of disease over individual privacy. This kind of issue is also made complex by technological advance. At the time of this writing, it is not clear how ethical practice related to PHI/HIPAA is applied to online systems that give external proof of vaccine status that appears to derive directly from medical records. While these would seem to be in a gray area, they have been widely accepted and seen as valuable, likely meaning that the broader understanding of how PHI is handled will continue to evolve and regulatory guidelines will continue to be modified and improved.

Data Sharing

For sensitive data like PHI, researchers need to work carefully within guidelines in order to share data with collaborators. In general, data sharing is done by first de-identifying data, which is to remove all information in the

data records that would link performance data back to the specific participant in the study. In many cases, this is as easy as coding data by participant id and avoiding the use of name (or email) in data records. If data cannot be effectively de-identified, then a research sharing agreement is written and reviewed by the IRB to evaluate any risks associated with possible privacy exposure.

Identification of participants is another area where technological advances have led to changes in oversight procedures. Many years ago, researchers might freely share biological specimens from human participants research with other labs. However, the advent of DNA sequencing means that blood or tissue samples can be analyzed in a way that reveals the original participant and now must be evaluated for privacy risk. It has been suggested that the same aspect may be true of some neuroimaging data, i.e., that brain images might be uniquely identifying eventually even though the tools for this do not yet exist. Machine learning techniques may also somebody be able to recover identity from extensive survey data if enough relatively individual data has been collected. In most cases, the privacy risk is minimal but data sharing should always be handled carefully to maintain compliance with best practice.

Waiver of consent

In some field research studies, research practice requires data collection without the ability to first provide informed consent. This can be approved by the IRB through a request for a **waiver of consent**. This process is also used in cases of severe privacy risk, such as data collection about criminal or high-risk behavior. If participants are potentially asked about prior behavior that could have immediate legal consequences, carrying out the research requires absolute privacy protection. In this case, the written consent process can be waived to ensure that no method of tracing the data back to the participant exists. This is obviously a highly specialized case and one that it very carefully and extensively reviewed. One aspect of the complexity of this process is that although ethical practice might require protection of all

research records, it cannot be guaranteed that a legal process such as a court order might override institutional preferences. Ethical practice guidelines from the IRB, the university or even federal funding agencies do not have legal standing to stop a court order or warrant.

Mandatory Reporting

A fairly recent example of tension between legal understanding of ethical practice and science is recent decisions requiring some research personnel to be **mandatory reporters** for some kinds of observations. The most common situation for this is in developmental research with young children. If research personnel suspect that child abuse is occurring, they are required to report this to institutional authority for investigation. This appears to be a privacy violation for the child and their family, yet the decision was made that the need to protect children overrides the privacy concern. This policy has additional consequences such as the need for training of research staff in these laboratories to appropriately decide when reporting of suspicion is required.

A related issue arises in studies of mental health among adolescents, even university students. Measures related to clinical depression have to be used carefully in practice as there may be institutional or legal policies in place requiring intervention for adolescents at risk. If answers to a mental health survey indicate a potential for self-harm, it may be necessary to have trained and qualified mental health professionals available for immediate participant support. A counter-intuitive consequence of this policy is that questions related to high levels of risk (e.g., self-harm) are often practically removed from mental health surveys if the research team does not have access to adequate support services. That is, knowing that a participant is at risk and not acting is deemed to be a worse situation than not asking questions that would identify the risk. This is clearly an example where scientific practice and ethics have fallen into a complex gray area without an obvious solution.

Summary

In most psychological research, effective compliance with best practices for ethical research are straightforward and easy to carry out. Showing respect for persons, ensuring voluntary cooperation, properly protecting privacy and data are usually easy to implement. Specialized procedures and training become critical for working in some select, more challenging subdomains. Awareness of the more complex issues in these domains is sometimes useful for understanding the long lists of questions asked in the process of obtaining approval for research from the IRB. Their processes have to be effective for all research and allow for determination of what the risks levels are without any assumptions about the usual case of minimal risk associated with much psychological science.

Key takeaways

- Everybody on the research team needs to be informed about standard practices and policies related to the responsible conduct of research to ensure fair and accurate scientific processes
- Data need to be handled with extreme care to avoid any possibilities of bias in selection that would weaken inferences from analysis
- Citation of others' influencing research is standard and appropriate. Use or reference to others' unpublished work should not be done.
- Authorship rules for publication vary across laboratories but generally adhere to a policy of authorship to individuals providing a significant intellectual contribution.
- Any information about participants related to personal or private health information must be handled carefully.
- Intervention research poses additional challenges to ethical science and is done in close collaboration with the IRB

Exercises

Question 1. Suppose a friend doing a research project said to you, "I'm sure my hypothesis is correct, so I'll just give my participants a hint here and there to make sure the data come out properly."

1a. What kind of RCR violation is this?

1b. What methodological approach should be used by this research to avoid this problem?

Question 2. In a study of marriage relationships, a researcher discovers that one of the participants in the study is an acquaintance and has indicated 'yes' to the question of having had an extra-marital affair. The researcher is debating whether they should inform the participants' spouse.

Question 3. A researcher doing a study on academic performance of students who have been diagnosed with ADHD. Some of the potential participants have refused to answer the question about their diagnosis, so the researcher contacts their medical provides without the participant's knowledge. What kind of research ethics violation is being considered here?

Question 4. Give an example of a research study not included in the chapter or which it would be unethical to include a placebo/control group.

19 Applied Research

Scientific research that attempts to maximize the external validity of findings works to apply the theories, methodologies and results outside the well-controlled laboratory environment. Like with non-experimental designs, giving up experimental control poses challenges for maximizing the internal validity of the scientific process. However, in many applied research studies, there can still be the key manipulated independent variable that allows for strong causal conclusions about how that manipulation affects the measured dependent variable.

In this chapter we will review three related types of research approaches that are all examples of quasi-experimental designs that mix elements of controlled laboratory designs with real world contexts. **Field research** typically refers to procedures testing psychological hypotheses in the conditions in which they are thought to apply. The term field research can also be applied to non-experimental approaches based on observation but when an environmental variable is surreptitiously controlled by a research team, this approach is very similar to experimental research within the laboratory.

Intervention research describes practical applications of research aimed at improving conditions or outcomes for populations related to the participants recruited into the study. These research projects are often very important

for policy questions and how organizations or institutions allocate resources. At their core, the basic methodology often depends on the simple designs described in Chapters 2-9 but ethical considerations about fair treatment of participants often require adding substantial complexity to the research process.

The term **translational research** is often used in medical or public health contexts to describe an attempt to bring a basic scientific finding, e.g., from biology, into practical medical applications aimed at improving health. In these areas, the term often serves to make a distinction among types of research where **basic science** is laboratory work aimed at understanding a core mechanism with an idea that this finding will later be used in translational science to provide some benefit. While these terms are used more in related fields like neuroscience and public health than psychological science, there is the same distinction between the value of well-controlled laboratory basic science with optimal internal validity and the messier process of applying a finding outside the lab with better external validity.

One of the major concerns with applied research is fitting this kind of science into the ethical framework for research. One of the benefits of laboratory work is that participants are aware they are voluntarily opting into the research process, can provide informed consent and choose whether they wish to continue through the study. Many of these factors are more difficult to achieve in applied research, necessitating careful discussions with the IRB providing oversight to ensure that the research properly balances the risks to participants and the benefits of the science. In addition, many kinds of intervention or translational research involve manipulating a factor that is thought to be a benefit to the participants. This is challenging to do and follow normal practice about justice, being fair to all participants. If the intervention being studied is effective, the control comparison condition is being unfairly deprived of this benefit. If the intervention is ineffective, the participants receiving it might be treated unfairly. However, there are a handful of common methodological approaches that minimize this issue as much as possible.

Example: Laptops in the classroom

Many teachers have been concerned that the use of laptop computers in classroom environments are not conducive to effective learning. Although laptops can help with notetaking, they also allow the possibility that students' attention might become distracted from the content being covered and interfering with learning. In laboratory environments, there is a robust literature on the phenomenon of *dual-tasking* that documents how attempting to do two cognitive tasks simultaneously affects performance on each task. There is also some research on a concept called *desirable difficulty* that suggests that memory might be better for material that one works a little harder to process. In a laboratory setting, taking notes manually (handwriting) has been shown to lead to better memory than typing (which is theoretically easier). However, neither of these phenomena have been convincingly demonstrated to lead to worse outcomes in a practical classroom environment.

Consideration of the challenges embedded in taking on a rigorous field research study to quantitatively test the effect of laptops illustrates why this study is deceptively difficult to carry out. On the surface this looks like a simple two-group design. Randomly assign half the students in some sample to having laptops available in class and the other half to relying solely on paper and pencil. This would be simple operational definition of our independent variable. As a dependent variable, we might be tempted to simply use final grade in the class at the end of the term as a measured operational definition of learning effectiveness.

As simple as this looks, as soon as we consider constructing a procedure to implement such a design, we run into problems. First, one of the concerns about laptops in the classroom is that they may actually impose their distracting effects on students sitting nearby. If somebody sitting in front of you is viewing non-class material, you might find yourself in an imposed dual-tasking condition where that information is competing with attention to the intended class content. As a result, creating properly controlled access to

laptops or freedom from their distraction might require assigning participants to conditions at the class level, i.e., one section of a class has laptop use allowed and another does not. However, if the two sections are not taught by the same instructor, we have an important uncontrolled extraneous variable (instructor) that is unfortunately confounded with our IV.

Further, if we pause to consider some of the other extraneous variables that might affect our study, we realize that our hypothesis may depend on the class content being studied. Humanities classes, STEM classes, small seminars and large lecture hall classes might all be affected by possible laptop distractions in different ways. Certain teaching styles might be more or less affected. It will also be difficult or impossible to control any possible bias due to expectancy effects from the students or the teacher, who will certainly be aware of the condition being imposed on the participants. Even though we have tried to conceptualize this research as a field study to improve external validity, all the choices about which classes are recruited into participation end up potentially limiting the applicability of our research.

Recruiting participants in a classroom context is also a potentially thorny issue. Students should feel free to take classes at their own choice, so it does not appear that any recruiting procedure is necessarily purely voluntary. And participants should always feel free to withdraw from the study, so the possibility of bias due to students being more likely to drop a class where they are unable to use their laptops for notetaking. If the intervention is effective, we also have to acknowledge that students who were required to take a class with laptops in use might have essentially have had their grades lowered by random assignment to conditions, which is generally considered unethical practice.

The main point of this example is to consider what seems like a fairly simple and straightforward two-group experimental design and realize that as an intervention study, it is exceptionally difficult to carry out a study with a very high level of internal validity. As a result, it is likely that any general policy decisions would necessarily rely on non-experimental methods (observation, correlational research) or laboratory studies that do not test the effect of the

intervention in a real context. In general with interventions, The elements that can not be approached with the best experimental design principles do not mean that valid inferences can not be drawn, but that our critical evaluation skills for identifying alternate explanations need to be applied to these studies. At the same time, a well-constructed and well-executed intervention design should be appreciated for the methodological care and effort that needs to be invested to make this type of research effective.

Quasi-Experimental Design

Quasi-experimental designs blend elements of experimental and non-experimental research designs. The goal in this approach to research is to try to capture aspects of the higher external validity of non-experimental designs, but still try to support strong causal statements about the relationship between the independent variable(s) and dependent variable.

A laboratory study is a study that is conducted in the laboratory environment where principles of experimental control such as constancy and counter-balancing can be used. In contrast, a field study is a study that is conducted in the real-world, in a natural environment. This increases external validity but at some cost to internal validity. Field studies can be real experimental designs with manipulated independent variables. In theory this allows for stronger causal inferences than non-experimental designs. However, the lack of full control over extraneous variables makes these less reliable and more vulnerable to Type 2 errors where the data are too noisy to reliably support strong conclusions.

Internal validity challenges in field research come from the same sources as all other experimental research, mainly arising from accidentally confounded variables or participant selection bias. In laboratory research, extraneous variables can be identified, and an experimental procedure constructed to minimize their effects (constancy) or at least de-confound these (counter-balance). In field research, there is often a very limited ability to employ these techniques, so it is more likely that conclusions will have to be

considered in the context of alternate explanations (like in non-experimental research).

Laboratory experiments that are designed properly typically have high internal validity. Internal validity refers to the degree to which we can confidently infer a causal relationship between variables. When we conduct an experimental study in a laboratory environment we have very high internal validity because we manipulate one variable while controlling all other outside extraneous variables. When we manipulate an independent variable and observe an effect on a dependent variable and we control for everything else so that the only difference between our experimental groups or conditions is the one manipulated variable then we can be quite confident that it is the independent variable that is causing the change in the dependent variable. In contrast, because field studies are conducted in the real-world, the experimenter typically has less control over the environment and potential extraneous variables, and this decreases internal validity, making it less appropriate to arrive at causal conclusions.

But there is typically a trade-off between internal and external validity. External validity simply refers to the degree to which we can generalize the findings to other circumstances or settings, like the real-world environment. When internal validity is high, external validity tends to be low; and when internal validity is low, external validity tends to be high. So laboratory studies are typically low in external validity, while field studies are typically high in external validity. Since field studies are conducted in the real-world environment it is far more appropriate to generalize the findings to that real-world environment than when the research is conducted in the more artificial sterile laboratory.

There are also field studies which are non-experimental in nature because nothing is manipulated. These will typically rely on observational methods and produce observed correlations between variables that are measured in non-laboratory contexts.

Extraneous variables in Field Research

In a field research protocol, the procedure is designed to be carried out in the real world. How are participants going to be identified to “recruit” into the study? In many cases they might not even know they are in the study, raising questions about ethics that must be addressed. Beyond ethics, the implications of the recruiting method may have important implications for the generalizability of results if there are any concerns that the recruiting method accidentally biases the sample.

Many field research procedures effectively require experimenters to become actors to carry out the procedures of the study as if it were a script. The validity of the research can end up depending on the acting ability of the experimenters and/or confederates. Field research procedures can also put a lot of pressure on observational techniques that require scoring aspects of behavior that can have a significant subjective component. Where possible, standard techniques should be used such as multiple raters and assessing consistency, or to structure the procedure so that the raters are blind to the experimental condition. These are not always possible and can weaken the internal validity of the results.

The process of “recruiting” participants can also be highly sensitive to the place, time and date of planned experimental data collection in field research. These can inadvertently incorporate bias into the recruiting procedures that needs to be considered when interpreting the results. Recruiting bias typically affects the external validity of the research with the idea participants found at a different time or place might have behaved differently. While field research is generally higher in external validity, recruiting processes can pose a challenge.

Ethics in Field Research

Field research has several specific characteristics of ethical concerns that occur frequently. The most obvious is that data is being collected from

participants without their awareness that they are in a research study. That is an elevated level of deception that is almost always inherent in this type of research that needs to be evaluated by the IRB before data collection can occur.

Field research also often involves areas of psychology that are sensitive since these are the kinds of topics for which behavior can be markedly different in the laboratory. In a laboratory setting, when participants are asked about attitudes related to stereotypes or high-risk behavior, their responses may incorporate expectations about what they believe the experimenter thinks is most socially appropriate. Collecting data on these ideas in the field may provide a much more direct insight into what people actually do rather than allowing them to report what they think is commonly considered acceptable. This exacerbates the issues related to deception and lack of clarity about informed consent. However, since these are often questions of great scientific interest, they can certainly be judged to be ethical in that the benefit of the research outweighs the costs to the participant in risk, or embarrassment.

Intervention research

A great deal of psychological science research aims to further our understanding of human behavior and the underlying cognitive processes. Studies that have a robust theoretical foundation and are designed to add new facts to what is known are termed “basic science” research. However, many psychological science research areas of study are aimed more directly at identifying ideas that might immediately and positively impact the world which are described as “applied research.” These ideas are not necessarily mutually exclusive, but in practice, basic science tends to be carried out within the laboratory with maximum control and the best internal validity. Applied research is typically done as field research where there is a controlled independent variable that involves the intervention.

In medical research, studies that aim to establish the effectiveness of a new drug are done as intervention studies. These follow many of the same

methodological issues, but often do not involve psychological measures. In addition, many of these studies are done in the context of very specific methodological practices related to ethics and rigor (e.g., double-blind methods) that are not always available in psychological studies. Research on improvements in educational practice captures more of the challenges of psychological research as there are important questions about efficacy of interventions that we want to observe in the context they would be used, e.g., a classroom. But in this work, it is challenging to implement all the preferred aspects of experimental control that we would want in laboratory research.

Basic Intervention Designs

In a **one-group post test only design**, a treatment is implemented (or an independent variable is manipulated) and then a dependent variable is measured once after the treatment is implemented. Imagine, for example, a researcher who is interested in the effectiveness of an anti-drug education program on elementary school students' attitudes toward illegal drugs. The researcher could implement the anti-drug program, and then immediately after the program ends, the researcher could measure students' attitudes toward illegal drugs.

This is the weakest type of quasi-experimental intervention design. A major limitation to this design is the lack of a control or comparison group. There is no way to determine what the attitudes of these students would have been if they hadn't completed the anti-drug program. Despite this major limitation, results from this design are frequently reported in the media and are often misinterpreted by the general population. For instance, advertisers might claim that 80% of women noticed their skin looked bright after using Brand X cleanser for a month. If there is no comparison group, then this statistic means little to nothing.

In a **one-group pretest-post-test design**, the dependent variable is measured once before the treatment is implemented and once after it is implemented. Let's return to the example of a researcher who is interested

in the effectiveness of an anti-drug education program on elementary school students' attitudes toward illegal drugs. The researcher could measure the attitudes of students at a particular elementary school during one week, implement the anti-drug program during the next week, and finally, measure their attitudes again the following week. The pretest-posttest design is much like a within-subjects experiment in which each participant is tested first under the control condition and then under the treatment condition. It is unlike a within-subjects experiment, however, in that the order of conditions is not counterbalanced because it typically is not possible for a participant to be tested in the treatment condition first and then in an "untreated" control condition. This is basically a non-experimental approach but gets used fairly often because of the ease of implementation.

The standard approach to ruling out the threats to internal validity in one-group designs is by designing the research protocol to include a **control group**, one that does not receive the treatment effect. A control group would be subject to the same threats from history, maturation, testing, instrumentation, regression to the mean, and spontaneous remission and so would allow the researcher to measure the actual effect of the treatment (if any). Of course, including a control group would mean that this is no longer a one-group design.

It can be difficult to identify an appropriate control task. The best control tasks manage demand characteristics for the participants. For example, a study on improving educational outcomes using intense extra training with personal tutors is going to have the issue that participants receiving tutoring are aware that something is being done to help them. If the control condition is simply additionally assigned, optional reading, the control condition may not respond to the post-test with the same expectations. However, a really well-matched control condition can be difficult to construct in many cases. As a result, it is not that uncommon to use the simplest alternative and simply not intervene. This is sometimes called a *wait-list control* condition to reflect the fact that recruiting is done among people who are seeking the benefit of the intervention, who are then randomly assigned to receive the intervention

or not. Those not receiving the intervention are designated the waitlist control condition.

Once a control condition has been identified this approach can be used as a simple two-group design comparing the intervention to the control condition. It can also be combined with a pre/post design where the DV is assessed both before and after either the intervention or control condition. Note that in this approach, the statistical tool that would be used to measure the intervention effect would be an ANOVA as this is now a 2x2 design. A successful intervention would cause an interaction where the difference between the first and second assessments would be larger for the group that received the intervention.

Using a control condition increases the internal validity of the design at the cost of an ethical problem. Participants being recruited into an intervention design study are often seeking help and randomly assigning them to the waitlist control condition is not entirely fair. A method for mitigating this problem is to use a **phased design** such as an **AB** or **ABA** design. In these designs, interventions alternate with baseline control conditions. The advantage of this approach is that the comparison group can alternate with the opposite conditions. The simplest design of this type is to use a waitlist control condition in the first phase and then in the second phase, apply the intervention to the control group. Contrasts on the DV at the end of the first phase are used as the core statistical test, but to manage ethical concerns the control group then obtains the potential benefit of the intervention in the second phase.

A more complex variation of the phased designs is to combine pre/post testing and multiple intervention phases into a **crossover interaction design**. In this design, all participants receive a pre-test at recruitment. They are then randomly assigned to intervention or control condition in the first phase. Then everybody receives a post-test. Then the intervention/control conditions are swapped in the second phase, which is then followed by a second post-test. While this is potentially complex and time-consuming to implement, a series of potentially interesting, related hypothesis are tested.

At the first post-test, differences in performance or differences in improvement from the pre-test across groups provide a measure of the intervention effectiveness. This is usually the primary goal of the research study. In addition, performance from the first post-test to the second for the intervention group provides a measure of whether the benefit of the intervention fades over time. This is very commonly an important question in interventions, not only whether they help but whether the benefits persist. Performance of the group that gets the intervention on the second phase can be compared as a pre/post design that effectively replicates the pre/post design of the early intervention group. This embedded replication, if successful, helps increase confidence in the efficacy of the intervention. And since both groups eventually get exposure to the intervention, ethical issues caused by random assignment to conditions are mitigated.

Longitudinal Research

Intervention research is often done over a more extended time period than our simpler experimental designs run in an hour or two in a laboratory. This feature requires consideration of some history effects like those that apply to within-participant designs. One such effect is the phenomenon of regression to the mean, which is a natural consequence of variability in human performance. A well-known example was reported in discussions with instructors who provided feedback to pilots who felt that it was most important to harshly criticize poor performance and ineffective to compliment good performance. If some of the daily variation in performance is simply due to random variance, today's poor performers are likely to improve to perform near the mean on the next session. At the same time, better performers will fall back to their average performance. The tendency for performance to center around a mean level plus variability leads to the phenomenon of regression to the mean. In this case, today's top performers are most likely to do nominally worse next time, and the bottom performers ought to do better even without a robust learning effect. This effect can look a lot like a response to feedback, as in this example, but with the data

described here, it is impossible to tell which account for the finding.

A related idea in clinical or medical intervention research is the idea of spontaneous remission, which is the rate of improvement that is not influenced by an intervention. This idea is another rationale for placebo control groups. Participants in the control group may improve due to demand characteristics (knowing they are in a study) or may simply improve due to other effects related to the passage of time. Intervention research aimed at treatment very rarely relies on one-group designs due to the risk of improvement being due to effects unrelated to the intervention.

Mortality Effects

The term **mortality effect** refers to a challenge to the internal validity of a study based on participants dropping out of the study. In this context mortality often does not refer to death of participants but can simply reflect a decision not to continue with participation in the research project. Since all participation in research should be voluntary, if one of the conditions is seen as unpleasantly onerous and not worth the time required or compensation offered, participants have the right to opt out of the study. If one of the experimental conditions experiences a higher rate of dropout, then this can bias the subsequent comparison between groups.

To illustrate this problem, consider a simple research design comparing the effect of reward and punishment on learning. Participants in the reward condition are asked to do a moderately challenging perceptual task and are rewarded with 50 cents for each correct response. A separate group of participants attempts to learn the same task but are punished with a mild electric shock for each incorrect response. If the study runs over several sessions and several days, a very plausible outcome is that the participants receiving shocks will exhibit higher levels of performance. However, without detailed information about participant dropout, we can draw no conclusion about the relative value of reward versus punishment. It is very likely in this simple design that participants in the punishment group who were poor at

learning the task would drop out of the study. If more of the poor learners drop out of the punishment than the reward group, any differences at the end of the study could be entirely due to differential mortality across groups.

This effect does not only happen in intervention research but can occur with any longitudinal design where data is collected across many days, sessions or even weeks or years. Research using this approach will generally need to report detailed information on participant retention across the study and compare this across all independent variables to try to rule out this problem. Intervention research with a challenging intervention can be weakened by participants opting out of the study because the intervention is unpleasant (e.g., diet or exercise interventions). Even participants in the control condition can opt out of participation if they become aware they are temporarily in a control or placebo condition. In some designs, recruiting and compensation practices are designed explicitly with reducing participant dropout by including compensation bonuses for completing the study. These techniques are always reviewed carefully by the IRB for any accidental coercion to participation that they might imply.

Mortality effects in research are closely related to the idea of survivorship bias in which inferences about data available at the end of a study or other ongoing process can be distorted by the missing data unavailable at the end. The most famous example of this was an analysis of military airplane damage done in World War II on planes that returned from combat. The key inference was that the undamaged parts of the planes that returned were likely the most vulnerable and critical regions based on the inference that the planes that had been hit in those regions failed to return. This idea has been raised as an issue in highly competitive fields (business, athletics, acting) where studying only people who have succeeded may not reveal critical differences that actually predict success for which information from less successful people needs to be included, essentially a sampling issue in the language of experimental research.

Maintaining Good Scientific Practice

Intervention research is also an area where challenges can arise related to conflict of interest and experimenter bias and can make maintenance of best rigor practices difficult. Because intervention research is challenging to implement, it is generally done with a strong hypothesis and expectation about the outcome by the research staff. Research procedures should be designed with attention to the possibility of bias in scoring the dependent variable or any aspect of implementing the conditions manipulated for the independent variable(s). In addition, large scale intervention research is resource intensive to do effectively, often meaning that a funding agency is supporting the research, usually with the idea of wanting to find effective interventions to address social problems. This can create the appearance of conflict of interest in that successful research can have fairly immediate direct benefits for researchers.

The existence of a potential conflict of interest does not mean that research cannot be carried out with a high degree of rigor and integrity. Standard practice is to disclose all information related to possible conflicts of interest so that reports of results can be interpreted in that context. Results in this context may be considered with a heightened degree of skepticism, but should not be treated as if they cannot be accurate. As with all research, confidence in conclusions accrues over time, replication and extension to novel operational definitions, samples and research contexts.

Key Takeaways

- Applied research methodologies use research methodologies outside the well-controlled laboratory environment.
- Field research aims to strengthen the external validity of scientific findings by establishing that the independent variable affects behavior in real-world contexts.
- Intervention research is designed to test manipulations that are hypothesized to have positive benefits for participants. These studies are often part of translational research programs to bring basic science to practice.
- These quasi-experimental designs involve the manipulation of an independent variable in order to draw causal conclusions. However, there are greater challenges from uncontrolled extraneous variables, and may not be able to fully use random assignment to conditions or counterbalancing. For these reasons, quasi-experimental research is generally higher in internal validity than non-experimental studies but lower than true experiments.
- Intervention research frequently uses within-participant methods to help control for participant variables and particular care must be taken to address the usual history effects and also any influence of the intervention on participant dropout or mortality effects.
- Ethical practice in intervention research requires using designs that provide fair access to the beneficial intervention to all participants whenever possible. This often leads to phased designs or a crossover intervention design approach.
- Ethical practice in field studies often does not permit the standard recruiting and informed consent process for participants. This kind of research needs to be carefully reviewed by the IRB to ensure that the research still meets expectations for balancing risk and benefits of research.

Exercises

TBA

20 Development and Neuropsychology

The patient known by the initial H.M. is one of the most famous case studies in psychology. H.M. suffered from severe epilepsy, a syndrome noted for severe and unpredictable seizures, that was intractable to all available medical treatment. Epileptic seizures arise from neural dysfunction that is often associated with a specific brain region, or foci, that sets off a brain-wide electrical storm. Treatment is usually pharmacological, but in severe cases where drug treatments are found to be ineffective, surgical resection (removal) of the foci region is sometimes done to reduce or eliminate subsequent seizure activity. The surgical approach can be very effective as the dysfunctional region is typically not contributing to general cognitive function in any meaningful way, so removal has no long-lasting effects on any general cognitive process. However, the surgery has to be done carefully to avoid removing any cortical brain regions that might still be contributing to cognition, known as *eloquent* cortex. Epileptic foci often occur in or near parts of the brain associated with memory so modern treatment approaches include a long period of cortical mapping to attempt to identify all the regions near the epileptic foci that need to be left intact.

H.M.'s surgical treatment predated the modern understanding of critical

brain regions for specific cognitive functions and the tools to do very careful mapping of still-functioning cortical regions had not yet been developed. In fact, at the time, the prevailing theory of cortical function was one of *equipotentiality*, the idea that all brain regions contributed together to complex cognitive functions such as memory. It was not thought that any one region of the brain would support a function as complex as memory, although there were some hints from other neuropsychological studies that there might be more *modular* organization to cognitive function in the brain.

To treat H.M., Scoville & Milner (1957) described a large, bilateral resection of cortical regions around the medial temporal lobe of his brain. For the epilepsy and subsequent seizure activity, the treatment was a success, however it soon became clear that H.M.'s cognitive function had been substantially altered in an unexpected way. H.M. exhibited a pattern of memory impairment that came to be called anterograde amnesia. The key feature of this syndrome is the inability to form new memories. Previously acquired knowledge and memories were intact, but nothing new could be added to his memory store. Almost all other aspects of general cognitive function were completely intact. He could carry on conversation, carry out problem solving and decision-making exercises. His knowledge of language and basic semantic information about the world was fully functional. However, nothing about his experiences after the surgery was stored, resulting in the peculiar experience that if you left the room and returned, he was unaware that

Henry Molaison

After his passing, the famous patient H.M.'s real name was revealed to the broader scientific community. He was referred to by his initials in the scientific literature for the majority of his life as a measure to protect his privacy. This was his family's preference instead of the potential notoriety associated with one of the most famous individual case studies in the history of neuropsychology.

he had ever seen you before.

This case was revolutionary in the understanding of the organization of memory and the brain overall. His selective impairment indicated that complex cognitive processes such as memory were localized to or at least dependent on specific structures. Once the specific pattern of cognitive impairment exhibited by H.M. had been described, many other patients with similar patterns of memory impairment were identified. Most of these patients have memory deficits much less severe, but exhibit the same distinction between acquisition of new memories for facts and events with largely intact retrieval of older memories. These patients also exhibit intact learning of procedural skills acquired through repetitive practice even though they do not remember the event of practicing. The study of patient HM also inspired a wide range of further research looking for other cognitive processes that could be strongly connected to specific brain regions, an area of research known as cognitive neuropsychology.

In this chapter we will review the methodology associated with analysis of famous single cases such as H.M., **case study** research. In addition, we will review some techniques associated specifically with **developmental research**, which aims to characterize changes in psychological processes across the lifespan. Each of these methodologies will have elements that are fundamentally non-experimental. Particularly interesting single cases are not assigned to that condition and age-related change is not something possible to control experimentally. However, techniques to study these areas often involve using experimental methods together with the non-experimental aspects of research to better understand the underlying phenomena.

Case Study Research

A case study is an in-depth examination of an individual. Sometimes case studies are also completed on social units (e.g., a cult) and events (e.g., a natural disaster). Most commonly in psychology, however, case studies provide a detailed description and analysis of an individual. Often the individual has a rare or unusual condition or disorder or has damage to a specific region of the brain. These studies can bear some similarity to non-experimental research approaches that are qualitative or observational.

Like many observational research methods, case studies tend to be more qualitative in nature. Case study methods involve an in-depth, and often a longitudinal examination of an individual. Depending on the focus of the case study, individuals may or may not be observed in their natural setting. If the natural setting is not what is of interest, then the individual may be brought into a therapist's office or a researcher's lab for study. Also, the bulk of the case study report will focus on in-depth descriptions of the person rather than on statistical analyses. With that said some quantitative data may also be included in the write-up of a case study. For instance, an individual's depression score may be compared to normative scores or their score before and after treatment may be compared. As with other qualitative methods, a variety of different methods and tools can be used to collect information on the case. For instance, interviews, naturalistic observation, structured observation, psychological testing (e.g., IQ test), and/or physiological measurements (e.g., brain scans) may be used to collect information on the individual.

Case studies as illustrating descriptions

The history of psychology is filled with influential cases studies, such as Sigmund Freud's description of "Anna O." Sigmund Freud used the case of a young woman to illustrate many principles of his theory of psychoanalysis (Freud, 1961). (Her real name was Bertha Pappenheim, and she was an early

feminist who went on to make important contributions to the field of social work.) Anna had come to Freud's colleague Josef Breuer around 1880 with a variety of odd physical and psychological symptoms. One of them was that for several weeks she was unable to drink any fluids. According to Freud,

She would take up the glass of water that she longed for, but as soon as it touched her lips she would push it away like someone suffering from hydrophobia....She lived only on fruit, such as melons, etc., so as to lessen her tormenting thirst. (p. 9)

But according to Freud, a breakthrough came one day while Anna was under hypnosis.

[S]he grumbled about her English "lady-companion," whom she did not care for, and went on to describe, with every sign of disgust, how she had once gone into this lady's room and how her little dog—horrid creature!—had drunk out of a glass there. The patient had said nothing, as she had wanted to be polite. After giving further energetic expression to the anger she had held back, she asked for something to drink, drank a large quantity of water without any difficulty, and awoke from her hypnosis with the glass at her lips; and thereupon the disturbance vanished, never to return. (p.9)

Freud's interpretation was that Anna had repressed the memory of this incident along with the emotion that it triggered and that this was what had caused her inability to drink. Furthermore, he believed that her recollection of the incident, along with her expression of the emotion she had repressed, caused the symptom to go away.

As an illustration of Freud's theory, the case study of Anna O. is quite effective. As evidence for the theory, however, it is essentially worthless. The description provides no way of knowing whether Anna had really repressed the memory of the dog drinking from the glass, whether this repression had caused her inability to drink, or whether recalling this "trauma" relieved the symptom. It is also unclear from this case study how typical or atypical Anna's experience was.

Case studies are useful because they provide a level of detailed analysis not found in many other research methods and greater insights may be gained from this more detailed analysis. As a result of the case study, the researcher may gain a sharpened understanding of what might become important to look at more extensively in future more controlled research. Case studies are also often the only way to study rare conditions because it may be impossible to find a large enough sample of individuals with the condition to use quantitative methods. Although at first glance a case study of a rare individual might seem to tell us little about ourselves, they often do provide insights into normal behavior. The case of H.M. provided important insights into the role of the hippocampus in memory consolidation.

However, it is important to note that while case studies can provide insights into certain areas and variables to study, and can be useful in helping develop theories, they should never be used as evidence for theories. In other words, case studies can be used as inspiration to formulate theories and hypotheses, but those hypotheses and theories then need to be formally tested using more rigorous quantitative methods. The reason case studies shouldn't be used to provide support for theories is that they suffer from problems with both internal and external validity. Case studies lack the proper controls that true experiments contain. As such, they suffer from problems with internal validity, so they cannot be used to determine causation.

Neuropsychological Case Studies

When case studies are based on known specific damage to brain regions provide some ability to draw a causal inference – damage to that area leads to the observed impairment – the difficult question to answer is about generalizability. It can never be fully determined if the same damage to any brain region will end up having the same impact for everybody else. Essentially, we cannot tell with current methods how similar brain organization is across people. The power of case studies is the ability to document even for one person that the observed pattern of damage and

intact function is possible to occur.

Another particularly impactful case study for this idea is the case of **Phineas Gage**. In the mid-19th century, Phineas Gage suffered an injury in a factor that led to an iron bar being blasted through part of his skill causing the loss of an eye and damage to the prefrontal cortex just above the eye. He survived the injury and additionally surprisingly exhibited very little impairment in general cognitive function. He did not exhibit difficulty in speaking (aphasia), recognizing objects (agnosia) or any pattern of motor impairment that is often associated with brain damage due to stroke (that impact motor control regions). However, reports at the time documented a very robust change in his personality and mannerisms. His behavior became extremely rude, marked by gross profanity and he appeared to unable to follow plans.

Unlike H.M., other patients with damage to the general region impacted by Gage's injury do not exhibit precisely the same pattern of impairment. Deficits like Gage are termed problems with *comportment* that do occur in some syndromes associated with prefrontal cortex damage. Modern studies of cognitive neuroscience have found a wide range of cognitive operations that depend on this brain region and also documented that there is great variety across individuals in how this system is organized. This part of the human brain is also notable as being the most distinct from all other great apes, likely reflecting the parts of the brain most greatly changed in the evolution of modern humans. As a case study, Gage clearly indicates a relationship of comportment and planning to these regions and then served to motivate studies since to identify the neural basis and operation of these processes.

Cognitive Neuroscience

The field of Cognitive Neuroscience reflects research on the understanding of the relationship of the neural operation of the brain to human cognitive processes. Many modern studies of this relationship have been inspired

by case studies of specific damage and constrained patterns of cognitive impairment. Some of this research is done with neuropsychological methods of looking for new cases with similar patterns or differing patterns associated with damage to similar brain regions. A great deal of this research aims to build from the neuropsychological studies to studies of cognitively healthy operation using methods of neuroimaging.

Neuroimaging methods are techniques for collecting data about brain function noninvasively from humans. Examples of these techniques include functional magnetic resonance imaging (fMRI), electro-encephalography (EEG) and transcranial magnetic stimulation (TMS). Detailed discussion of these methods is beyond the scope of this text but depends on the same underlying research method processes used here. Constructs must be implemented the same way we draw inferences from simpler behavioral measures in the experimental psychology methods described here.

Developmental Research

In this chapter, we will review a few specialized research methodologies that apply to research across human age range. Research with young children or infants depends on many of the basic experimental design techniques discussed previously but requires some additional consideration of the kinds of dependent measures that are available and the ethics of research with *vulnerable populations* (Chapter 8). In addition, research comparing changes across age introduces methodological terminology contrasting cross-sectional and longitudinal designs.

Developmental research often refers to experimental methodology applied to young children. Obviously, the operational definitions for this kind of research need to account for the participant population. The research protocol and especially any task instructions need to be prepared to be age appropriate so that the participants understand the experiment details. The dependent variable needs to be a task that the children can perform.

In research with very young infants, measuring behavior can be challenging. Infants can indicate preference or in some cases familiarity by measures of preferential looking. In these paradigms, the infant is presented with a display that contains two different kinds of information, e.g., on two sides of a display. The infant is carefully observed to identify which direction they are looking or for how long they look. This can be used to infer babies' understanding of the physics of the world around them by demonstrating that they look longer at displays constructed to contain apparent physical violations of collisions or violations of numeracy.

Among the challenges of implementing this technique is the need to make subjective ratings of the direction of looking. This is often done with a blind-rater technique where a camera records the direction of looking from an angle where the rater cannot see what is visible to the infant. The rating process done this way avoids any potential bias in rating based on knowing what the infant was supposed to do based on the experimental hypothesis. This research is still challenging to carry out because of the potential for many extraneous variables unrelated to the experiment affecting looking. These studies also frequently have a challenge in accumulating enough participants to support robust statistical inference. Not every child is comfortable in the experimental situation and research is often highly restricted to very specific age ranges (e.g., 3-6 months), making the available population very limited.

An additional practical challenge is that in many of these procedures, it is necessary to have the implicit participation of an adult who is a parent or guardian of the child. It is common in these paradigms to have the child seated on the lap of a parent in order to maintain comfort so that behavior can be observed. However, this can introduce concerns about the child's behavior being biased by parental expectations. The parent has effectively become part of the research staff and needs to try to act in a consistent, unbiased manner so that behavioral differences solely reflect the child.

Cross-Sectional and Longitudinal Research

When psychologists wish to study change over time (for example, when developmental psychologists wish to study aging) they usually take one of three non-experimental approaches: cross-sectional, longitudinal, or cross-sequential. **Cross-sectional** studies involve comparing two or more pre-existing groups of people (e.g., children at different stages of development). What makes this approach non-experimental is that there is no manipulation of an independent variable and no random assignment of participants to groups. Using this design, developmental psychologists compare groups of people of different ages (e.g., young adults spanning from 18-25 years of age versus older adults spanning 60-75 years of age) on various dependent variables (e.g., memory, depression, life satisfaction). Of course, the primary limitation of using this design to study the effects of aging is that differences between the groups other than age may account for differences in the dependent variable. For instance, differences between the groups may reflect the generation that people come from (a cohort effect) rather than a direct effect of age. For this reason, longitudinal studies, in which one group of people is followed over time as they age, offer a superior means of studying the effects of aging. However, **longitudinal studies** are by definition more time consuming and so require a much greater investment on the part of the researcher and the participants.

A third approach, known as cross-sequential studies, combines elements of both cross-sectional and longitudinal studies. Rather than measuring differences between people in different age groups or following the same people over a long period of time, researchers adopting this approach choose a smaller period of time during which they follow people in different age groups. For example, they might measure changes over a ten year period among participants who at the start of the study fall into the following age groups: 20 years old, 30 years old, 40 years old, 50 years old, and 60 years old. This design is advantageous because the researcher reaps the immediate benefits of being able to compare the age groups after the first assessment. Further, by following the different age groups over time they can subsequently

determine whether the original differences they found across the age groups are due to true age effects or cohort effects.

Research across the lifespan also needs to be sensitive to the implementation of best ethical practices across age ranges. Participants under the age of 18 cannot provide written consent to participate in research but are generally consulted in addition to a parent or guardian consenting to their participation. The research protocol still needs to be arranged to acknowledge that the child still understands that participation is voluntary. Interactions between parents and children can be complex to anticipate, especially for older children such as adolescents who have different expectations about their preferences for research or privacy of data. All participants under 18 are considered *vulnerable populations* and oversight of research practices is elevated compared with research on cognitively healthy adults.

Research that aims to understand changes in behavior or cognition as a result of clinical syndromes is often termed neuropsychological research. Much of this work is done at the other end of the lifespan development with older adults. This work can be done as systematic research across groups of patients with clinical syndromes such as Alzheimer's disease or Parkinson's disease (and also with younger patients for syndrome such as schizophrenia). The cognitive impairments associated with some neuropsychological or neurological research means that these patients are sometimes also treated as vulnerable populations, requiring consent of guardian, etc. In general, these studies just require attention to the basic element of respect for persons for these older adults and acknowledgment of their challenges. This area of research is also one where the method of single case studies has been used to characterize particularly interesting single patients with unexpected patterns of cognitive function.

Developmental stages

A common use of cross-sectional studies in young children is to identify and characterize developmental stages as children grow from infancy. In these designs, participants of specific age ranges will be compared. For example, measures related to object permanence might have been seen to be different for infants at 4 months of age compared with 8 month olds. This research will usually follow the familiar form with manipulated independent variables and a measured dependent variable. Age, of course, is a participant variable not under experimental control in cross-sectional studies. Developmental studies will often use fairly creative dependent measures because very young participants may not provide accurate verbal answers.

For pre-verbal infants, measures of behavior often focus on subtle shifts based on eye-gaze or attention. These are examples of measures that need rigorous scoring procedures to ensure accuracy of data and to avoid bias. These techniques often rely on video recordings of the participants that are scored by blind raters unaware of the experimental condition of the measure. Further complicating these methods is that very young infants must often be tested while on a parent's lap in order to observe attentive behavior. In these cases, bias can creep into a design through expectations of the parents. In some cases, it can even be necessary to occlude the vision of the parents, e.g., through a blindfold, to keep them unaware of the experimental condition in which the infant is in.

Developmental studies also are notably difficult and time-consuming to carry out due to the complexity of designs, unpredictability in behavior of children and the need to have well-controlled face to face interactions. Recruiting can also be a significant challenge when very specific age ranges are targeted. Challenges in scheduling the research protocol can lead to a participant no longer meeting the intended age requirements for the planned study.

Research on children also notably requires informed consent from a parent or guardian. However, the willingness of the parent to have the child participate is not necessarily the same as willingness of the child, which makes confident

compliance with general ethical procedures complex. For older children, dual consent processes may allow children to consent verbally in conjunction with written consent from the parent. Any discomfort or disinterest in the research protocol has to be considered as a possibly interest in withdrawing from the protocol. Particular care must be paid to participant privacy, especially with video recordings of behavior which have to be handled so that access is restricted to study personnel and these are stored securely.

In addition, recent concerns about creating a robust culture of child protection raised the question of whether researchers are required to report any evidence they might observe of mistreatment of children by the parent or guardian. A substantial debate arose around this question. Research team members are trained in research and are not necessarily experts in detecting issues related to child welfare and safety. Mistakes leading to accusations of parents inappropriately would raise a new form of harm emerging from research participation. However, not acting on obvious evidence of abuse seems entirely unethical. Common current practice is to consider researchers in many developmental labs as *mandatory reporters* who are required to report obvious evidence of mistreatment. However, detailed reporting mechanisms aim to provide expert review of concerns to best maintain fair treatment of all participants.

Life-span Development

While much developmental research is aimed at understanding growth and maturation from infancy to childhood and then adulthood, research on late-life aging has some similar methodological challenges. Ethical concerns about voluntary consent, patient privacy (especially for health records) and even mandatory reporting (elder abuse) are present although to a much lesser extent. The primary methodological common element is the intention to characterize changes in cognition or behavior across aging, which is a variable that cannot be experimentally controlled. Comparing across large age ranges also compares across cohort effects caused by broad environmental or social

changes over the lifespan age range. For example, comparing some aspects of health in participants who are currently in their 60s to participants in their 30s varies a substantial period of exposure to lead in the atmosphere (in the 1970s) in addition to age. Social, historical and cultural changes over time are extraneous variables all potentially confounded with age that provide challenges to drawing robust conclusions.

As an example of the difficulty in drawing causal inferences, it has long been observed that many more women than men are diagnosed with Alzheimer's disease. One hypothesis is that this emerges from some key genetic difference related to the X chromosome, such as hormonal differences across the lifespan. However, it is also the case that women currently tend to live longer than men and age is a very strong predictor of Alzheimer's disease in that the older you are the more likely you are to express symptoms. Current aged populations grew up and lived in a social context where women live longer than men, putting them potentially at greater risk for later life onset. This creates the usual problem with confounded variables in that we have two perfectly reasonable explanations for an observed relationship and no easy way to tell which is true. Additional scientific work aimed at a better understanding of the underlying neurobiology of Alzheimer's disease may eventually clarify why this difference emerges.

A more recent similar example of a more positive orientation is recent work in aging identifying a sub-population of older adults called **super-agers**. These are adults who at the age of 80 or more exhibit a high level cognitive functioning that would match participants under the age of 50. These individuals were originally identified as a series of case studies. More systematic research then identified larger numbers of these individuals who may have been previously overlooked in health research because their functioning is entirely intact. Ongoing research aiming to identify "their secret" for aging so successfully has not yet found any robust lifestyle, activity or genetic differences that account for their retained high level of cognitive ability. This work will necessarily be correlational at first but could potentially lead to some future intervention based research to improve aging

outcomes.

Key Takeaways

- Case studies are an observational technique with particularly interesting individuals that often illuminate new areas of science
- Neuropsychological studies of unusual patients have been very influential in developing theories of brain-behavior relationships
- Cognitive Neuroscience is a research area within psychology and neuroscience that uses research methodology to understand cognitive functioning in the brain based on patient studies and neuroimaging methods
- Developmental psychology relies on using traditional research methods together with the participant variable of age
- Comparisons across ages can be done as a cross-sectional design between-participants or as a longitudinal design within-participants over time
- Developmental methods with young children pose specific challenges for developing operational definitions that can be implemented with rigor and without bias
- Ethical research with children, who are a vulnerable population, requires additional oversight and care
- Life-span development examines age-related change in older adults, often through correlational research methods which have the usual challenges for drawing inferences from non-experimental designs

Exercises

TBA