

Research Methods in Psychology

Understanding Science:
How Do They Know That?

Paul J. Reber, Ph.D.
Northwestern University



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.

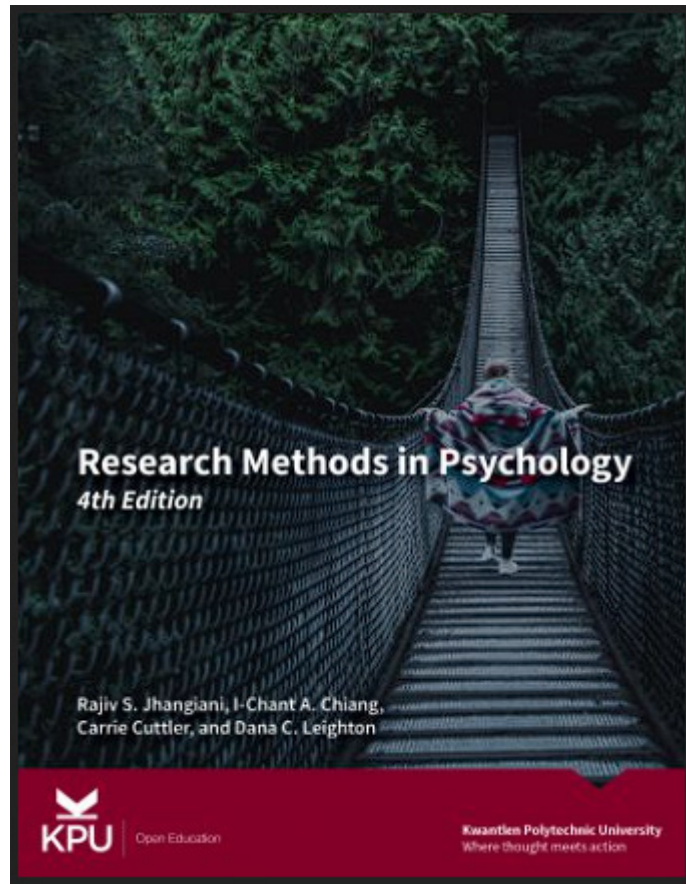
The content is aimed not to completely overlap with my classroom lecture slides content. I do not want the students to feel that classroom time is spent completely rehashing the text. I prefer to have different, novel examples illustrating the concepts and to use a very question-and-answer style in the 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.

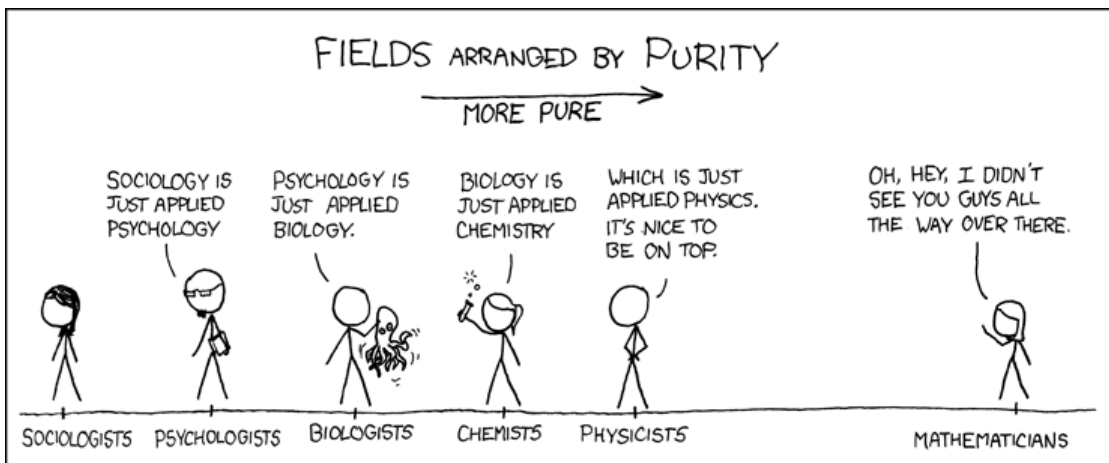


Table of Contents



1. Introduction to Experimental Research: Experiment 1

A simple memory experiment to illustrate methodology

2. Psychological Measurement: the Dependent Variable

Operational definition, constructs to quantities

3. Experimental Control: Extraneous Variables

Avoiding confounds, constance and counterbalancing

4. Experimental Control: Participants and Procedure

Random assignment, demand characteristics and avoiding bias

5. Statistics 1: Carrying out a Students t-test

Running a two-sample t-test in R

6. Report Writing in the Scientific (APA) Style

How-to manual on formal scientific writing

7. Experimental Design: Within-participants

Participants as their own controls, managine history effects

8. Research Ethics 1: Protection of Research Participants

The Institutional Review Board and informed consent

9. Experimental Design: 2x2 Factorial designs

Extending methods to more complex designs and hypotheses

10. Interpreting Factorial Design, Main Effects and Interactions

Inferences from complex design

11. Participant Sampling, Generalizability of Results

Recruiting and external validity

12. Statistics 2: Analysis of Variance

Running an ANOVA in R

13. Developing a Research Proposal

Developing hypotheses for independent research

14. Non-experimental design: Correlational and Qualitative Research

Methods when the IV cannot be manipulated

15. Surveys as Research Instruments

A brief overview of using and constructing surveys

16. Statistics 3: Correlation and Chi²

Additional statistical tools for simple designs

17. Practicalities

Field research, intervention research, user experience measures

18. Research Ethics 2: Responsible Conduct of Research

Working as a fair and ethical scientist

19. Quasi-experimental Design

20. Development and Neuropsychology Methods, Case Studies

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, generalizeability 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.

Acknowledgements



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. I believe the current version is substantially improved from the prior version but I do think there is a lot of room for additional improvement that I hope to continue to add over the next few years.

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 word 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 words, 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.

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.

Psychology is the science of human behavior.

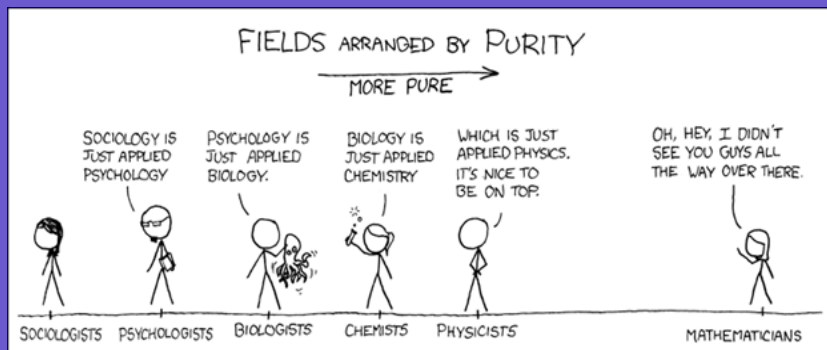
As a human, you have pretty good intuitions about how humans behave. Personal knowledge and experience can be a good starting point for the tricky problem of coming up with the operational definitions for psychological constructs.

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).

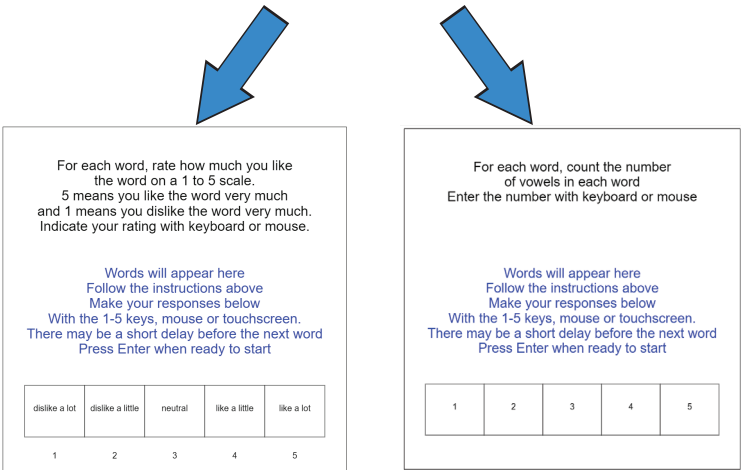
XKCD



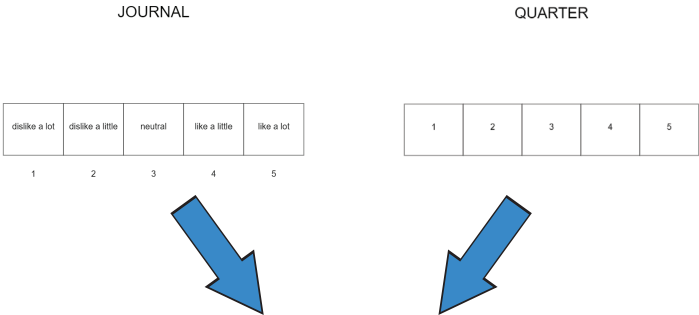
drawn by Randall Monroe, is an exceptionally sharp source of perspectives on science, <https://xkcd.com/435/>

Experiment 1 Design

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



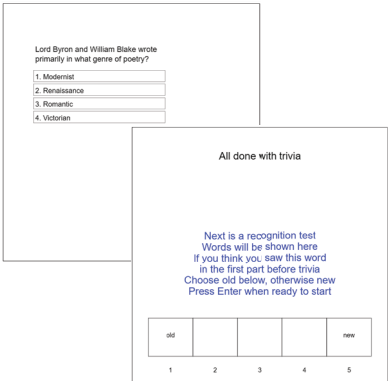
Although the instructions differ, every word is shown for 4s to maintain matched viewing time



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

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



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.

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.

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.

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 leads 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 use 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 chapters, 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 an 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 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.

Confounds

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 participate 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 of 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.

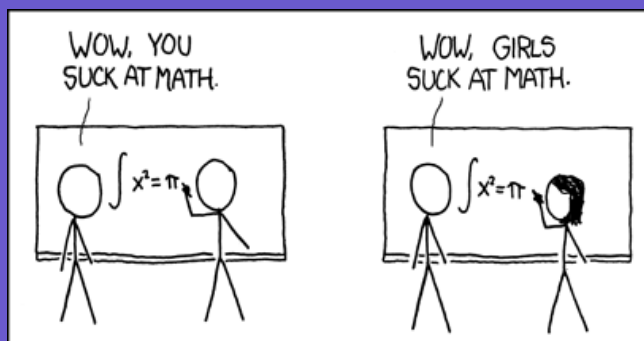
Stereotypes and bias are also well-known to be challenging to study in laboratory conditions because of effects related to **demand characteristics**, which we will also discuss below. These have to do with participants

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 logistically 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 logistically 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 generalizeability 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 I

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:

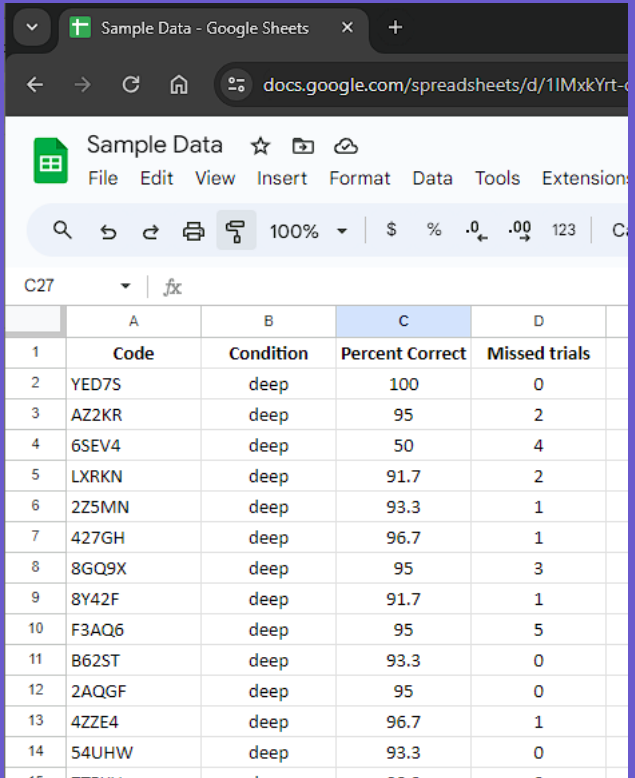
$$SE = \frac{SD}{\sqrt{n}}$$

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
	Code	Condition	Percent Correct	Missed trials
1				
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	7T8VU	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 Sheet with the following data:

	A	B	C	D
	Code	Condition	Percent Correct	Missed trials
1				
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	7TBYU	deep	98.3	0
16	AWC15	deep	83.3	6
17	BDYVR	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	XPTYX	deep	97	0
27				
28			<code>=AVERAGE(C2:C26)</code>	
29				

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 in 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 help 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.

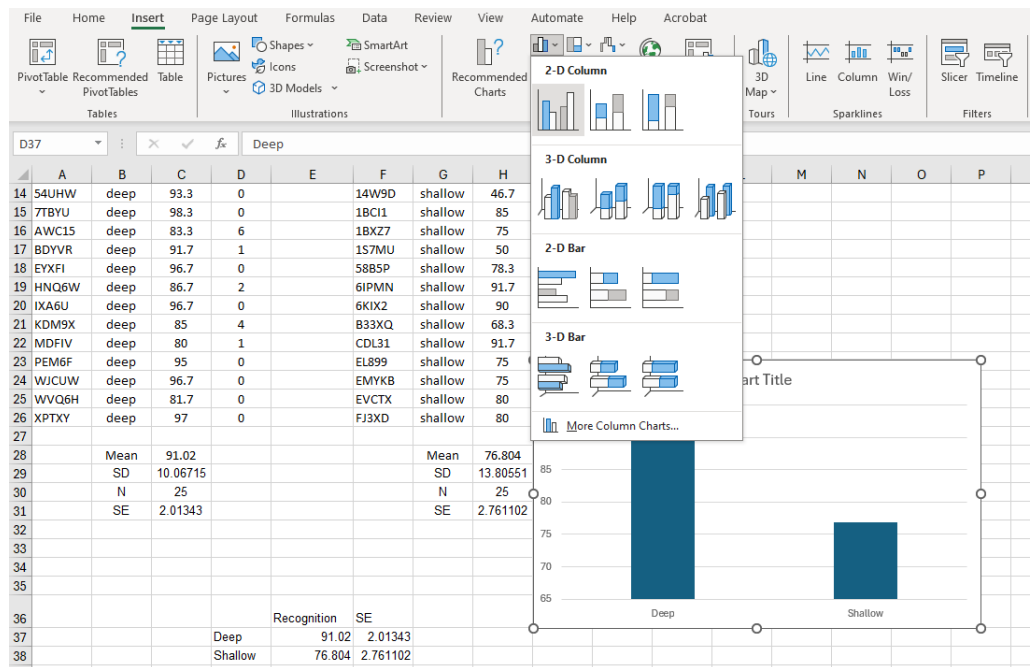
Data Visualization

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.

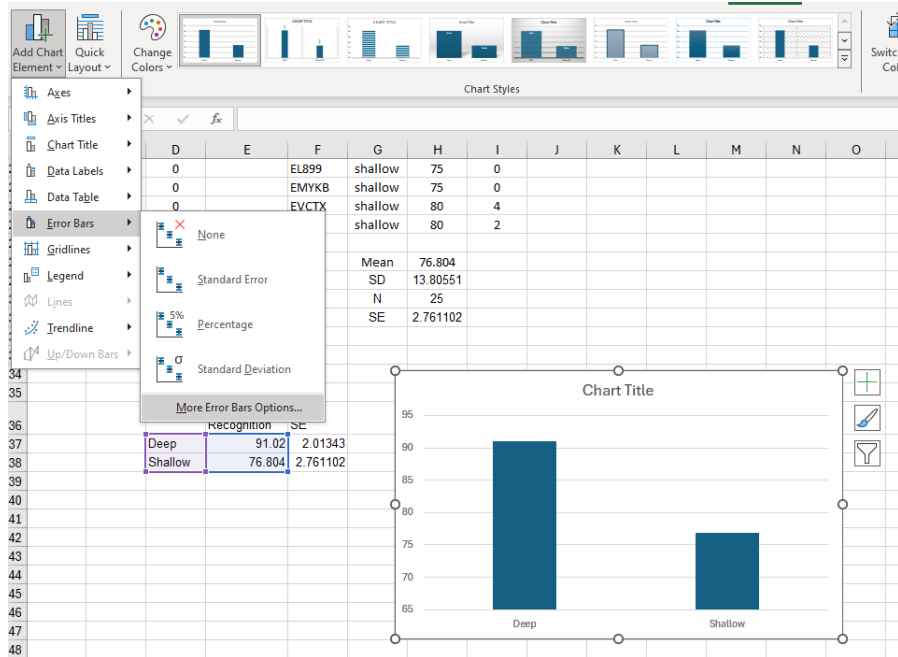
The first step towards creating a figure is generally to create a separate, labeled table of the key numbers that will contribute to the graph. The numbers we need for the graph will be the mean and the SE for each of the two conditions in our study: deep and shallow. For simplicity, we want these organized 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.

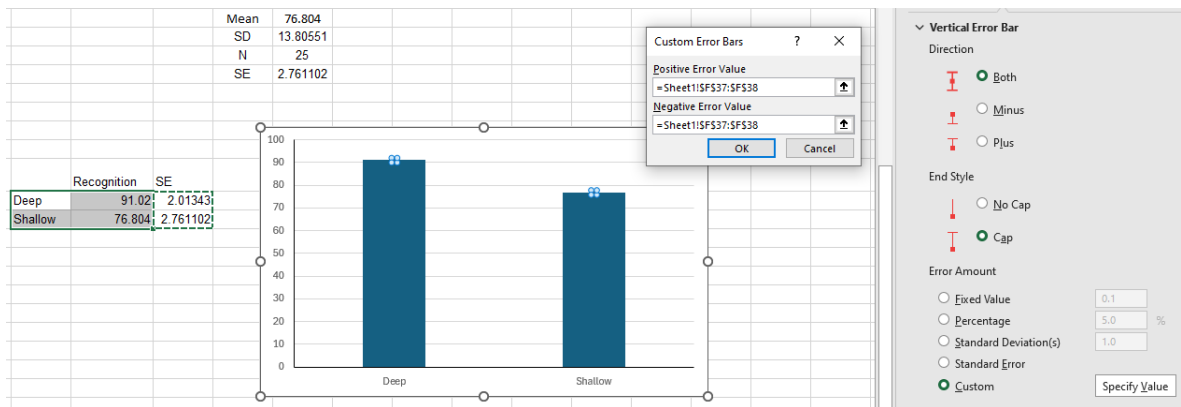


From this initial draft, we need to do some editing to the layout of this chart to make it effective and in approximately standard format. First, the Chart Title can be cut as we generally do not include titles on figures in manuscripts. Titles are used to help describe data in presentation formats, 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.

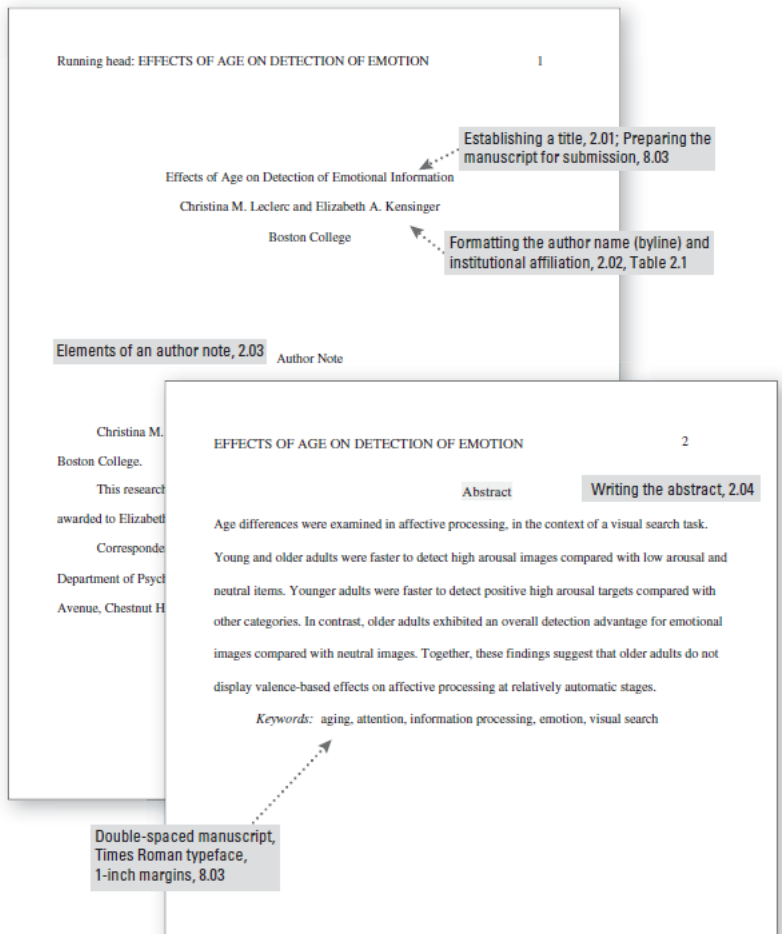
- Sex Differences in Coping Styles and Implications for Depressed Mood
- Effects of Aging and Divided Attention on Memory for Items and Their Contexts
- Computer-Assisted Cognitive Behavioral Therapy for Child Anxiety: Results of a Randomized Clinical Trial
- Virtual Driving and Risk Taking: Do Racing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

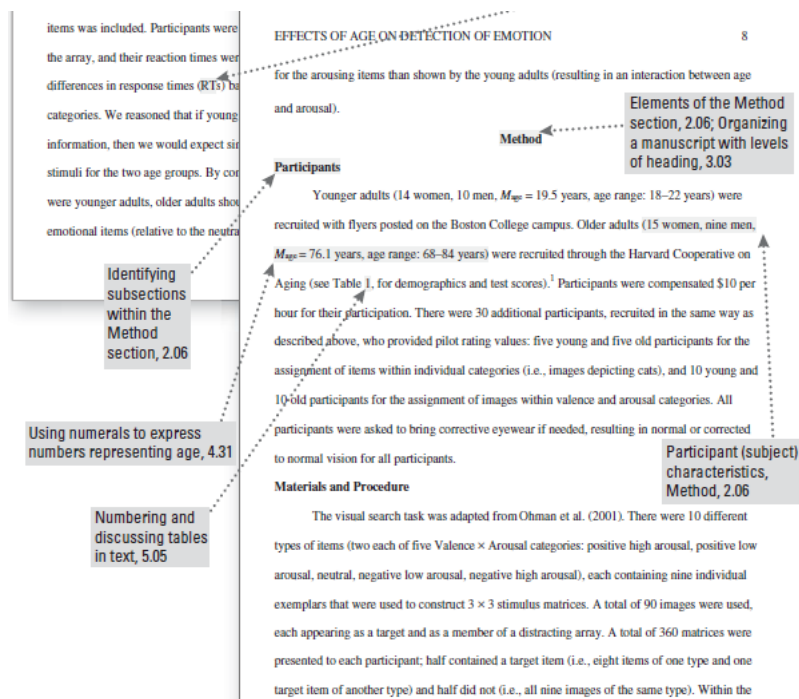
Method

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 assess 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 statistical 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 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

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

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?

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 HURTLES 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 vary 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 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?