

# 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.

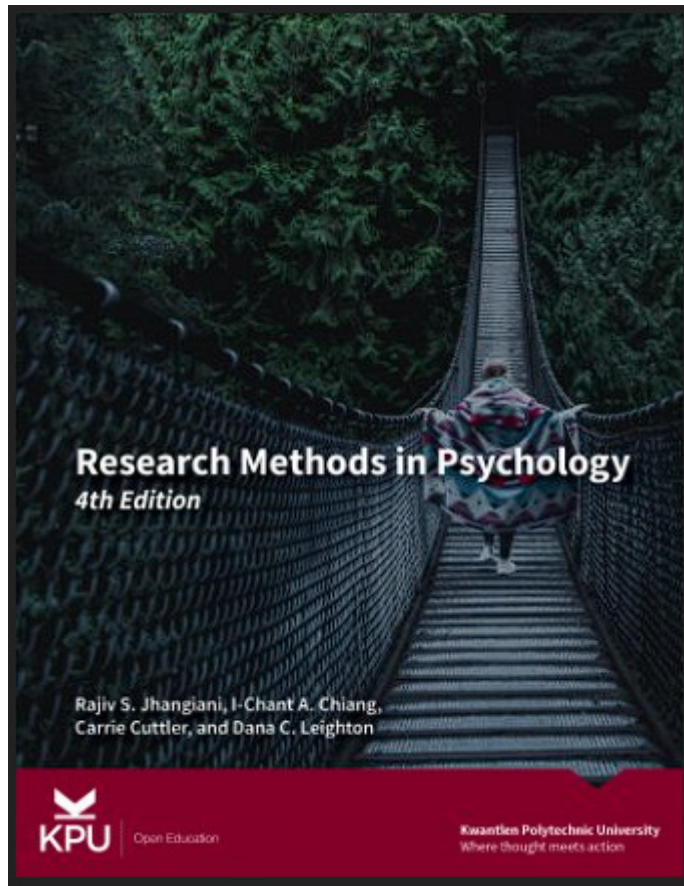
We will start by reviewing methods of experimental psychology research. Typical textbooks for Research Methods start with a review of the scientific method, some history on psychological science, discussion of non-experimental methods and then the process of the design, implementation, and analysis of 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.

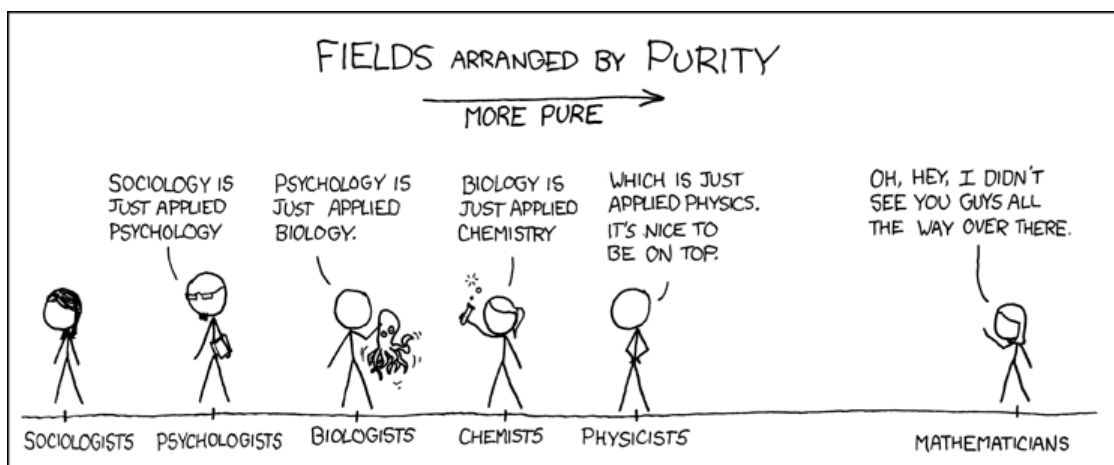
In recent years, psychology's conversation has shifted to an introspective one, looking inward and re-examining the knowledge that we considered foundational. We began to find that some of that unshakable foundation was not as strong as we thought; some of the bedrock findings in psychology were being 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 Chi<sup>2</sup>

Additional statistical tools for simple designs

17. Practicalities

Field 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.

# 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 operation 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 ("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?***

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.

Statistics are the bridge from your numeric, quantitative data to statements

about the conclusions and meaning of your study. Our review of experimental methodology assumes prior familiarity with basic statistical methods from a prerequisite class. However, statistics will be used here in a potentially different manner than in prior classes.

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 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.

We will consider these ideas of reliability and validity of research from two different perspectives as well. When carrying out research, scientists spend a great deal of time worrying about whether their

## ***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**

study will produce a reliable result where reliability is defined as meeting specific statistical criteria. This is because for a study to have scientific impact, it needs to be reported and published in a peer-reviewed journal. The review process always involves an element of evaluating the reliability of the data, so if the results do not meet the field standard criteria, the result will not be published and available to other scientists to read and learn from. For experimental research, obtaining a reliable result requires a well-designed study much more than any statistical sophistication so techniques for effective design will be our main concern.

An additional important goal of the study of research methods is to be able to read and understand results that have been published. As a reader/consumer of science, reliability of the published results is generally less of a concern because the authors and reviewers have already judged that the data and analysis meet the standards for reliability. However, we will observe that when we ask “how do they know that?” about published research that we will identify validity challenges that may weaken or limit the conclusions drawn even from published research.

***Psychology is the science of human behavior.***

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

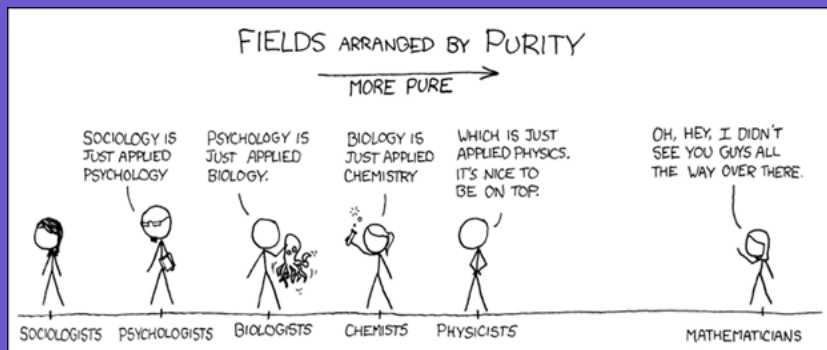
Much of your ability to identify weaknesses in scientific methodology will come from your understanding of human behavior as a human. In this class, we will augment this with some practice applying critical thinking skills systematically to these questions.

## Experiment 1

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

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

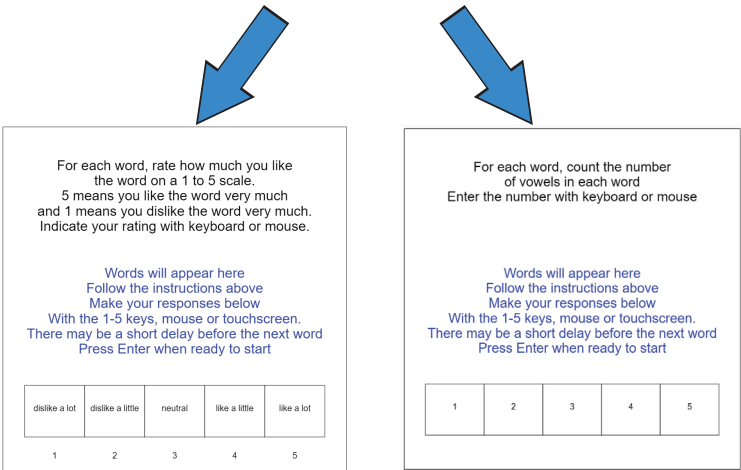
**XKCD**



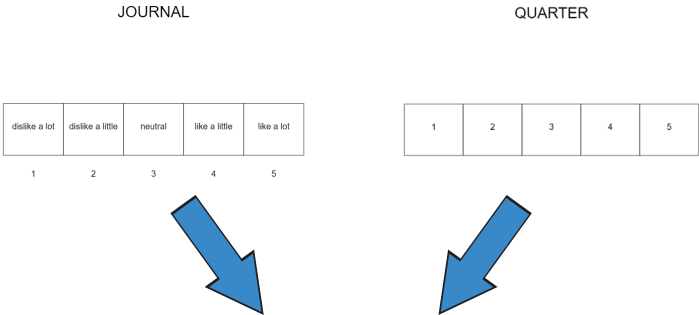
drawn by Randall Monroe, is an exceptionally sharp source of perspectives on science, <https://xkcd.com/435/>

Experiment 1 Design

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



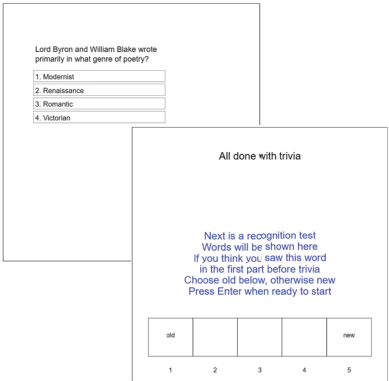
Although the instructions differ, every word is shown for 4s to maintain matched viewing time



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 IV, 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 ( $9.8 \text{ m/s}^2$ ). 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.

Stevens's 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

## **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.**

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.

## ***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?