Research Methods in Psychology: Thinking Like a Psychological Scientist

Nicolette P. Rickert Virginia Wickline Cassandra Baldwin
Daniel Gambacorta Amy Hackney Audrey Molter

2026-08-11

Table of contents

Preface

This textbook for *Research Methods in Psychology* consists of 26 chapters divided into six larger modules: Foundations of Psychological Research, Measurement and Sampling, Descriptive Statistics, Calculating Inferential Statistics, Research Designs, and Communicating your Outcomes & Growth.

The text was created under a Round 28 Affordable Materials Grant.

This is Open Access Textbook typeset with Quarto books.

To learn more about Quarto books visit https://quarto.org/docs/books.

Textbook Information

Research Methods in Psychology: Thinking Like a Psychological Scientist

For questions or feedback about this textbook, please contact the first author:

Nicolette P. Rickert, Associate Professor, Department of Psychology, Georgia Southern University, Statesboro, GA nrickert@georgiasouthern.edu

Acknowledgements

Thank you to Dr. Nikki Cannon-Rech for her assistance throughout this process. This work was made possible through the Round 28 Affordable Learning Georgia Affordable Materials Grant (Transformation Grant #796).

Part I

Module 1: Foundations of Psychological Research

1 Chapter 1: Thinking like a Psychological Scientist

In this chapter, you will examine how people acquire knowledge and why some sources of knowledge are more reliable than others. You will also explore the limits of human judgment, including the biases that can shape how we interpret information. Finally, you will learn how scientific thinking and scientific theories help psychologists develop more dependable explanations of behavior and why these skills matter even if you never plan to become a researcher.

1.1 How do we know what we know?

Take a minute to ponder some of what you know and how you acquired that knowledge. Perhaps you know that you should make your bed in the morning because your mother or father told you this is what you should do, perhaps you know that swans are white because all of the swans you have seen are white, perhaps you know that your friend is lying to you because she is acting strange and won't look you in the eye, or perhaps you know the skincare products that will get you glass skin because your favorite influencer showed you on Instagram. But should we trust knowledge from these sources? The methods we use to acquire knowledge can be broken down into four categories each with its own strengths and weaknesses.

1.2 Intuition

The first method of knowing is intuition. When we use our intuition, we are relying on our guts, our emotions, and/or our instincts to guide us. Rather than examining facts or using rational thought, intuition involves believing what *feels* true. Imagine you are walking alone to your car at night and notice someone approaching in the distance. Before consciously analyzing the situation, you may experience an immediate feeling that the person is either safe or potentially threatening. This rapid judgment is an example of intuition. The problem with relying on intuition is that our intuitions can be wrong because they are driven by cognitive and motivational biases rather than logical reasoning or scientific evidence. For example, while any strange behavior you see from your friend may lead you to think she is lying to you, it

may just be that she is trying to hold in a bit of gas or is preoccupied with some other issue that is irrelevant to you.

1.3 Authority

Perhaps one of the most common methods of acquiring knowledge is through authority. This method involves accepting new ideas because some authority figure states that they are true. These authorities include parents, the media, doctors, priests and other religious authorities, the government, professors, and even social media influencers. While in an ideal world we should be able to trust authority figures, history has taught us otherwise and many instances of atrocities against humanity are a consequence of people unquestioningly following unethical authority (e.g., Salem Witch Trials, Nazi War Crimes). On a more benign level, while your parents may have told you that you should make your bed in the morning, making your bed provides the warm damp environment in which mites thrive. Keeping the sheets open provides a less hospitable environment for mites. These examples illustrate that the problem with using authority to obtain knowledge is that they may be wrong, they may just be using their intuition to arrive at their conclusions, and they may have their own reasons to mislead you. Nevertheless, much of the information we acquire is through authority because we don't have time to question and independently research every piece of knowledge we learn through authority. But we can learn to evaluate the credentials of authority figures, to evaluate the methods they used to arrive at their conclusions, and evaluate whether they have any reasons to mislead us.

1.4 Rationalism

Rationalism involves using logic and reasoning to acquire new knowledge. Using this method, premises are stated and logical rules are followed to arrive at sound conclusions. For instance, if I am given the premise that all swans are white and the premise that this is a swan then I can come to the rational conclusion that this swan is white without actually seeing the swan. The problem with this method is that if the premises are wrong or there is an error in logic then the conclusion will not be valid. For instance, the premise that all swans are white is incorrect; there are black swans in Australia. Also, unless formally trained in the rules of logic it is easy to make an error. Nevertheless, if the premises are correct and logical rules are followed appropriately then this is a sound means of acquiring knowledge.

1.5 Empiricism

Empiricism involves acquiring knowledge through observation and experience. Once again many of you may have believed that all swans are white because you have only ever seen white

swans. For centuries people believed the world is flat because it appears to be flat. These examples and the many visual illusions that trick our senses illustrate the problems with relying on empiricism alone to derive knowledge. We are limited in what we can experience and observe and our senses can deceive us. Moreover, our prior experiences can alter the way we perceive events. Nevertheless, empiricism is at the heart of the scientific method. Science relies on observations. But not just any observations, science relies on structured observations which is known as ***systematic empiricism***, a topic we'll turn to later on.

1.6 The Limits of Human Judgment

Although intuition, authority, rationalism, and empiricism can provide valuable information, none of these ways of knowing is perfect. Humans are not objective observers of the world. Instead, our judgments are often influenced by cognitive biases, systematic patterns of thinking that can lead us to wrongful conclusions. Cognitive biases affect what we notice, how we interpret information, and how confident we are in our beliefs. Because these biases operate automatically, even intelligent and well-intentioned people are susceptible to them.

Four particularly important biases are the confirmation bias, the availability heuristic, the representativeness heuristic, and the bias blind spot. Understanding these biases helps explain why people sometimes reach incorrect conclusions and highlights the importance of scientific methods for evaluating information and claims.

The ***confirmation bias*** is the tendency to seek out, notice, and remember information that supports our existing beliefs while overlooking or discounting information that contradicts them. For example, if you believe that your professor is unfriendly, you may pay close attention to a few negative interactions while ignoring numerous friendly and supportive ones. As a result, your belief may become stronger even if it is inaccurate.

The ***availability heuristic*** occurs when people judge the likelihood of an event based on how easily examples come to mind. Because vivid, recent, or memorable events are easier to recall, they can exert an outsized influence on our judgments. For example, after hearing several news stories about shark sightings, a person may overestimate the likelihood of encountering a shark during their next trip to the beach despite the rarity of such events.

The ***representativeness heuristic*** occurs when people make judgments based on how closely a person, object, or event matches their existing expectations or stereotypes. For example, you may assume that your professors spend their free time reading books and engaging in intellectual conversations because those activities fit your mental image of a professor. Likewise, a healthcare provider may be slower to recognize symptoms of a heart attack in a woman than in a man if the provider's mental prototype of a heart attack patient is male.

Finally, people are susceptible to the ***bias blind spot***, the tendency to recognize cognitive biases in other people while failing to see those same biases in themselves. Most individuals readily acknowledge that others can be influenced by stereotypes, selective attention, or faulty

reasoning, yet believe that their own judgments are objective and unbiased. Ironically, the bias blind spot can make it more difficult for people to recognize and correct their own errors in thinking.

Pause and Reflect: *Can you think of a time when you searched for information that confirmed something you already believed? How might confirmation bias have influenced your thinking?*

Taken together, these biases illustrate an important limitation of human judgment: people are not always aware of the factors influencing their beliefs and decisions. Because intuition and everyday reasoning are vulnerable to systematic errors, psychologists rely on scientific methods designed to minimize the influence of bias and evaluate claims more objectively.

1.7 Why Science Matters

The scientific method is a process of systematically collecting and evaluating evidence to test ideas and answer questions. While scientists may use intuition, authority, rationalism, and empiricism to generate new ideas, they do not stop there. Scientists go a step further by using systematic empiricism to make careful observations under controlled conditions and rationalism to draw logical conclusions from those observations. While the scientific method is the most likely of all methods of knowing to produce valid knowledge, it is not without limitations. Scientific research often requires substantial time, effort, and resources, and not all questions can be answered scientifically. The scientific method is best suited for answering empirical questions—questions that can be addressed through observation and evidence. This textbook examines how psychologists use the scientific method to advance our understanding of human behavior and mental processes.

1.8 Scientific Versus Everyday Reasoning

Each day, people offer statements as if they are facts, such as, "It looks like rain today," or, "Dogs are very loyal." These conclusions represent ***hypotheses*** about the world: best guesses as to how the world works. Scientists also draw conclusions, claiming things like, "There is an 80% chance of rain today," or, "Dogs tend to protect their human companions." You'll notice that the two examples of scientific claims use less certain language and are more likely to be associated with probabilities. Understanding the similarities and differences between scientific and everyday (non-scientific) statements is essential to our ability to accurately evaluate the trustworthiness of various claims.

Scientific and everyday reasoning both employ ***induction*****,** drawing general conclusions from specific observations. For example, a person's opinion that cramming for a test increases performance may be based on her memory of passing an exam after pulling an all-night study session. Similarly, a researcher's conclusion *against* cramming might be based on studies

comparing the test performances of people who studied the material in different ways (e.g., cramming versus study sessions spaced out over time). In these scenarios, both scientific and everyday conclusions are drawn from a limited ***sample*** of potential observations.

Science also uses ***deduction***, the process of applying general principles or theories to generate specific predictions. For example, if a theory suggests that learning improves when study sessions are spaced out over time, a researcher might predict that students who spread out their studying will perform better on an exam than students who cram the night before. Deductive reasoning allows scientists to move from broad theories to specific hypotheses that can be tested empirically.

Although both everyday thinking and science use induction, science does not rely on induction alone. By combining inductive observations with deductive testing of predictions, science provides a more reliable way of evaluating competing explanations. Consider a student trying to decide how to prepare for an exam. One source of information encourages her to cram, while another suggests that spacing out her studying time is the best strategy. How should she decide which advice to follow? To make the best decision with the information at hand, we need to appreciate the differences between personal opinions and scientific statements, which requires an understanding of science and the nature of scientific reasoning.

1.9 Characteristics of Scientific Thinking

Psychological scientists differ from everyday thinkers not because they are smarter or less susceptible to bias, but because they use methods designed to minimize the influence of bias. Scientific thinking is characterized by skepticism, systematic observation, testable explanations, and a willingness to revise conclusions when new evidence becomes available. Rather than accepting claims at face value, scientists ask whether the available evidence supports the claim and whether alternative explanations have been ruled out.

One hallmark of scientific thinking is ***skepticism***. Scientific skepticism does not mean believing that all news is fake news, rejecting every claim made by others, or assuming that new ideas are false. Instead, it means *withholding judgment* until sufficient evidence is available. Skeptical scientists remain open to new ideas while also requiring that those ideas be supported by reliable evidence. In this sense, skepticism is a healthy habit of questioning rather than doubting everything.

Scientific thinking also relies on ***systematic empiricism***. Unlike everyday observations, scientific observations are made in a carefully organized and standardized manner. Researchers develop procedures for collecting data, measuring variables, and documenting their methods so that others can evaluate and replicate their work. Systematic observation helps reduce many of the errors that arise from relying solely on personal experience or anecdotal evidence.

Another defining characteristic of scientific thinking is that scientific claims must be ***testable***. Researchers should be able to collect evidence that either supports or challenges their ideas.

If no possible evidence could demonstrate that a claim is incorrect, then the claim cannot be evaluated scientifically. This principle, known as ***falsifiability***, has become one of the defining characteristics of modern science and is discussed in more detail below.

Finally, scientific thinking is ***self-correcting***. Scientific conclusions are not considered final truths. Instead, researchers continually evaluate new evidence, refine existing theories, and revise conclusions when warranted. As additional evidence accumulates over time, scientific understanding becomes increasingly accurate, even though individual studies may sometimes produce conflicting findings.

Together, these characteristics help explain why science is one of the most reliable methods humans have developed for acquiring knowledge. Although scientists are not immune to cognitive biases, the scientific process provides tools for recognizing and reducing their influence.

1.9.1 Falsifiability

Although each of these characteristics contributes to the strength of scientific thinking, one has had an especially profound influence on modern science: ***falsifiability***. Proposed by philosopher Karl Popper, falsifiability provides a useful way of distinguishing scientific claims from claims that cannot be evaluated scientifically. Understanding falsifiability helps explain why scientists emphasize testable hypotheses and why the scientific method is designed to rule out incorrect explanations rather than simply accumulate evidence in favor of preferred beliefs.

In the early 20th century, Karl Popper (1902-1994) suggested that science can be distinguished from ***pseudoscience*** (or just everyday reasoning) because scientific claims are capable of being ***falsified***. That is, a claim can be conceivably demonstrated to be untrue. For example, a person might claim that "all people are right handed." This claim can be tested and—ultimately—thrown out because it can be shown to be false: There are people who are left-handed. An easy rule of thumb is to not get confused by the term "falsifiable" but to understand that—more or less—it means testable.

On the other hand, some claims cannot be tested and falsified. Imagine, for instance, that a magician claims that he can teach people to move objects with their minds. The trick, he explains, is to *truly believe* in one's ability for it to work. When his students fail to budge chairs with their minds, the magician scolds, "Obviously, you don't truly believe." The magician's claim does not qualify as falsifiable because there is no way to disprove it. It is unscientific.

Popper was particularly concerned about theories that could explain every possible outcome. He argued that if a theory can account for any result, it becomes difficult or impossible to test scientifically. As an example, Popper criticized some of Freud's explanations of personality and mental illness. Imagine a person who grows up to be an obsessive perfectionist. If she were raised by messy, relaxed parents, Freud might argue that her adult perfectionism is a reaction to her early family experiences—an effort to maintain order and routine instead of

chaos. Alternatively, imagine the same person being raised by harsh, orderly parents. In this case, Freud might argue that her adult tidiness is simply her internalizing her parents' way of being.

As you can see, according to Freud's rationale, both opposing scenarios are possible; no matter what the disorder, Freud's theory could explain its childhood origin—thus failing to meet the principle of falsifiability. Popper argued that when a theory can explain multiple contradictory scenarios equally well, it becomes difficult to identify evidence that would disprove it, making the theory less scientifically useful.

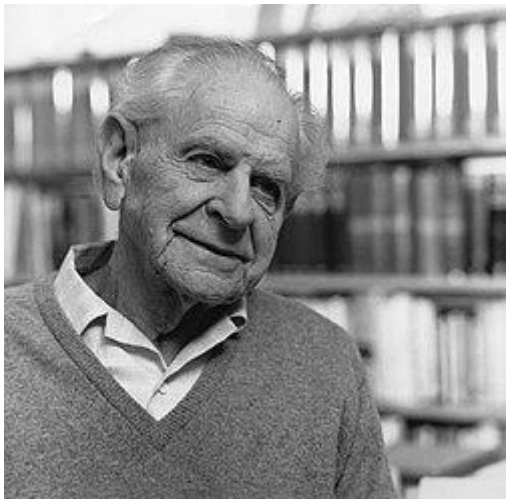


Figure 1.1: Portrait of philosopher Karl Popper, whose concept of falsifiability influenced modern scientific thinking.

Figure 1.1 *Karl Popper*[1]

Popper argued that scientific progress comes not from proving ideas correct, but from eliminating explanations that do not fit the evidence. The explanations that survive repeated testing become the ones scientists have the greatest confidence in. In other words, scientists should be able to identify, *before collecting data*, what evidence would cause them to question or revise their hypothesis.

Test Yourself: Can It Be Falsified?

For each statement below, determine whether it is falsifiable. If it is, describe what evidence could demonstrate that the statement is false.

A. Chocolate tastes better than pasta.

B. We live in the most violent time in history.

C. Time can run backward as well as forward.

D. All swans are white.

[See answer at end of this module]

Although the idea of falsification remains central to scientific data and theory development, these days it's not used strictly the way Popper originally envisioned it. To begin with, scientists aren't solely interested in demonstrating what *isn't.* Scientists are also interested in providing descriptions and explanations for the way things *are.* Whether researchers are investigating when children begin speaking in complete sentences, whether exercise reduces depression, or why people are happier on weekends, they must draw conclusions from limited samples of data. As a result, scientific conclusions are based on probability rather than absolute certainty. Evidence may support or challenge a hypothesis, but scientific knowledge always remains open to revision as new evidence accumulates.

1.10 Scientific Theories

Because scientific conclusions are based on evidence rather than absolute proof, scientific knowledge is organized into theories that explain observations and guide future research. In everyday conversation, people often use the word *theory* to mean a guess or speculation. For example, someone might say, "I have a theory about why my roommate is always late," or "My theory is that our team will win the championship." In these cases, the word theory refers to a personal opinion or educated guess.

In science, however, a theory is much more than a guess. A ***scientific theory*** is a well-supported explanation for a set of observations or findings. Scientific theories are based on extensive empirical evidence and are used to organize existing knowledge, explain why phenomena occur, and generate predictions about future observations.

Scientific theories differ from facts, but they are not opposed to facts. Facts are observations about the world, whereas theories explain those observations. For example, it is a fact that objects released near the Earth's surface fall toward the ground. The theory of gravity helps explain why this occurs. Similarly, psychologists develop theories to explain patterns of human thought, emotion, and behavior.

One of the strengths of science is that theories can change when new evidence emerges. For example, people once believed that the Sun revolved around the Earth. As astronomers collected more systematic observations, the evidence supported a different explanation: the Earth and other planets revolve around the Sun. Scientists view the willingness to revise theories in light of new evidence as a strength rather than a weakness.

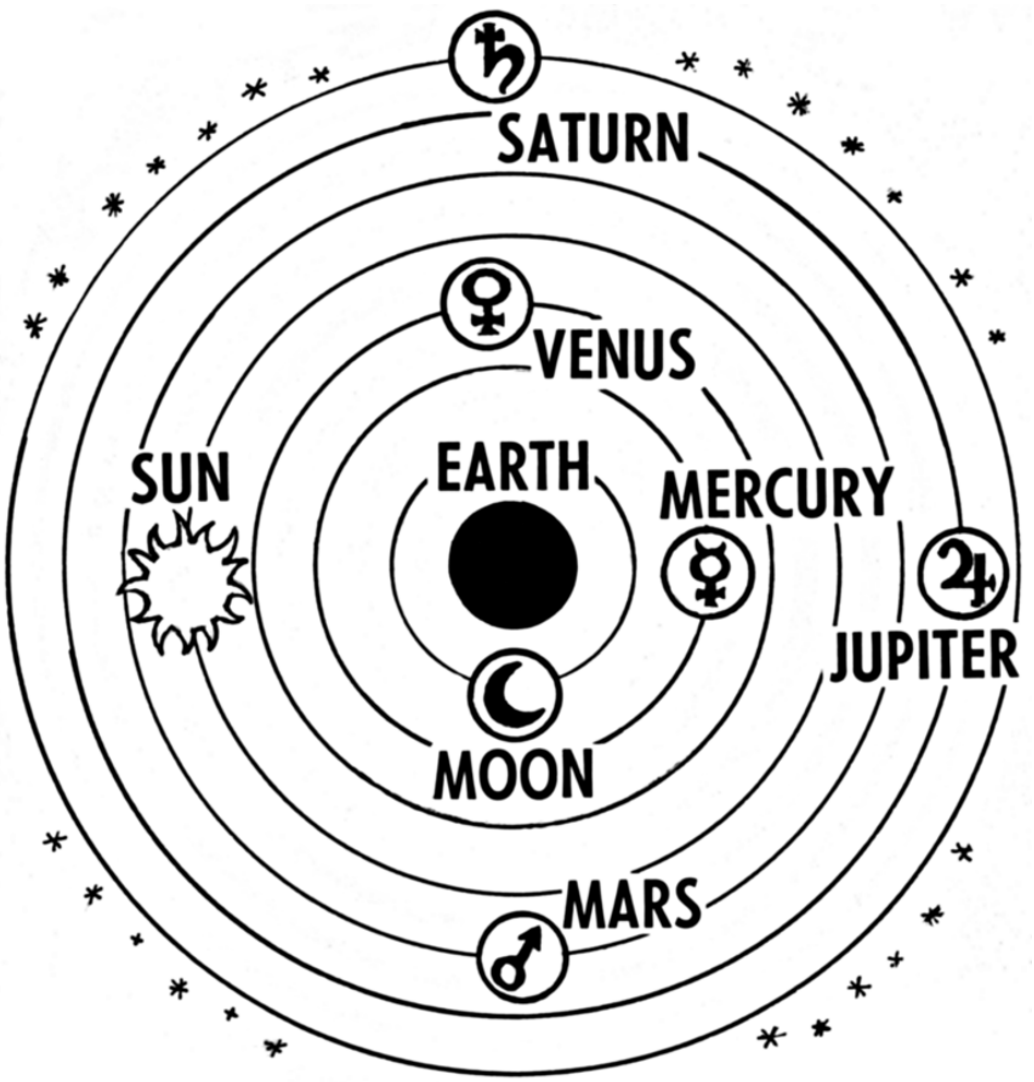


Figure 1.2: Diagram of a geocentric model showing Earth at center surrounded by Moon, Mercury, Venus, Sun, Mars, Jupiter, and Saturn in concentric circular orbits. Each celestial body is labeled with its name and astrological symbol, with small stars scattered outside Saturn's orbit representing outer space. This diagram represents early theories that placed the Earth at the center of the solar system. We now know that the Earth revolves around the sun.

Figure 1.2 *Early theories placed the Earth at the center of the solar system.*[2]

Scientists often develop multiple theories to explain the same phenomenon. Not all scientific theories are equally useful. Scientists evaluate theories using several criteria, including their ability to accurately explain observations, remain internally consistent, account for a broad

range of phenomena, provide simple explanations when appropriate, and stimulate new research. Table 1 summarizes these characteristics of strong scientific theories.

Table 1.1: Five features of scientific theories (Kuhn, 2011)[3]^. A table that describes the five features of good scientific theories - accuracy, consistency, scope, simplicity, and fruitfulness. The table contains a definition and example for each feature.

Feature	**Definition**	**Example**
Accuracy	Explanations and theories match real-world observations	E.g., Although people say, "opposites attract," theories that focus on the role of partner similarity do a better job of explaining the observed data.
Consistency	A theory has few exceptions and shows agreement with other theories within and across disciplines.	E.g., The theory of evolution explains many findings across biology and psychology predicting, for example, that humans are better able to solve problems presented in concrete rather than abstract terms.
Scope	Extent to which a theory extends beyond currently available data, explaining a wide array of phenomena.	E.g., There is a theory that people use mental "short cuts" when making decisions rather than weighing every single piece of evidence. This can be seen in consumer purchasing behavior, in romantic relationships, in charitable donations, and in health choices.
Simplicity	When multiple explanations are equally good at explaining the data, the simplest should be selected.	E.g., The simplest explanation for why "good" people sometimes do "bad" things is because the succumb to some outside influence.

Feature	Definition	Example
Fruitfulness	The usefulness of the theory in guiding new research by predicting new, testable relationships.	E.g., The explanation that competition leads to improved performance can be tested by researching different types of competition.

Together, these characteristics help scientists evaluate competing theories. Strong theories not only explain existing evidence but also make accurate predictions, organize knowledge, and inspire new questions for researchers to investigate. Over time, theories that consistently perform well across these criteria become increasingly influential in guiding scientific research.

Pause and Reflect: *Have you ever heard someone say, "It's just a theory"? How has your understanding of the word theory changed after reading this chapter?*

Psychological scientists rarely begin a study with no direction. Instead, they use existing theories, previous research, and observations of the world to identify unanswered questions and develop testable hypotheses. In the next chapter, we will examine how psychologists transform broad theories into specific research questions and hypotheses that can be investigated using the scientific method.

1.11 Why Understanding Psychological Science Matters

Although relatively few psychology students will become researchers, every student will become a consumer of research. Throughout your life, you will encounter claims about what causes depression, whether certain parenting styles are more effective than others, how to improve memory, which study strategies work best, or how social media affects mental health. News articles, podcasts, social media posts, advertisements, politicians, and influencers often present scientific findings in ways that are incomplete, misleading, or even inaccurate.

Understanding psychological science provides the tools to evaluate these claims critically. Rather than accepting information because it "sounds right" or because it comes from an authority figure, you will learn to ask important questions. What evidence supports the claim? Could there be an alternative explanation? Was the conclusion based on a single study or many studies? Are the findings consistent with existing scientific theories? By applying the principles of scientific reasoning, you can make more informed decisions in your personal life, your career, and as an engaged member of society.

The goal of this textbook is not simply to teach you how psychologists conduct research. It is to help you develop scientific thinking skills that you can use long after this course ends.

1.12 Answer Key: Test Yourself 1

A. Chocolate tastes better than pasta.
Not falsifiable (as written). "Tastes better" is a subjective preference that can differ from person to person. Because there is no objective criterion for determining which food "tastes better," the statement cannot be conclusively shown to be false.

B. We live in the most violent time in history.
Potentially falsifiable. Researchers could collect historical data on rates of violence across different time periods and compare them. Although obtaining complete historical data is challenging, evidence could support or contradict the claim.

C. Time can run backward as well as forward.
Potentially falsifiable. Scientists could search for evidence that time runs backward or make predictions that would be expected if it did. If repeated observations consistently fail to support the claim, confidence in the hypothesis decreases.

D. All swans are white.
Falsifiable. Finding a single swan that is not white would contradict the claim. In fact, black swans exist in Australia, demonstrating that the statement is false.

1.13 Media Attributions

[1] *Karl Popper*. Photograph by Lucinda Douglas-Menzies. Source: Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Karl_Popper2.jpg. Rights status: No known copyright restrictions (Flickr Commons; permission verified by Wikimedia VRT).

[2] *Ptolemaic (geocentric) model*. Pearson Scott Foresman. Source: Wikimedia Commons, File:Ptolemaic system (PSF).png - Wikimedia Commons. License: Public domain.

[3] *Features of good scientific theories.* Smith, E. I. (2026). Thinking like a psychological scientist. In R. Biswas-Diener & E. Diener (Eds), *Noba textbook series: Psychology.* Champaign, IL: DEF publishers. Retrieved from http://noba.to/nt3ysqcm

1.14 Text Attributions

Anguiano, R., & O'Neil, M. (n.d.). Methods of Knowing. In S.D'Costa, M.Ukeye, M. O'Neil, & R. Anguiano, Critical research methods in psychology. Saint Mary's College of California. Licensed under CC BY-NC-SA 4.0. Modified by current authors.

Smith, E. I. (2026). Thinking like a psychological scientist. In R. Biswas-Diener & E. Diener (Eds), *Noba textbook series: Psychology.* DEF publishers. Retrieved from http://noba.to/nt3ysqcm. Modified by current authors.

Spielman, R. M., Jenkins, W. J., & Lovett, M. D. (2020). Problem Solving. In Psychology 2e. OpenStax, Rice University. Access for free at https://openstax.org/books/psychology-2e/pages/1-introduction Licensed under CC BY-NC-SA 4.0. Modified by current authors.

1.15 References

Kuhn, T. S. (2011). Objectivity, value judgment, and theory choice, in T. S. Kuhn (Ed.), *The essential tension: Selected studies in scientific tradition and change* (pp. 320-339). Chicago: University of Chicago Press. Retrieved from http://ebookcentral.proquest.com

2 Chapter 2: Generating Research Questions and Hypotheses

2.1 From Theories to Research Questions

In Chapter 1, we learned that scientific theories organize existing knowledge, explain observations, and generate predictions about future events. But theories alone do not produce scientific discoveries. To evaluate a theory, psychologists must design research that tests its predictions. Every research study begins with a question: What do we want to understand about behavior or mental processes? Research questions and hypotheses are closely related, but they are not the same thing**.** A research question asks what a researcher wants to understand, whereas a hypothesis is a specific prediction about what the researcher expects to find. Throughout this chapter, you will learn how psychologists move from broad theories to focused research questions and finally to hypotheses that can be tested using scientific methods.

2.1.1 Research Questions and Variables

Before we address where research questions in psychology come from—and what makes them more or less interesting—it is important to understand the kinds of questions psychologists ask and the variables they study. This requires a brief introduction to several basic concepts, many of which we will return to in more detail later in the book.

Research questions in psychology are about ***variables***. A variable is a quantity or quality that varies across people or situations. In other words, a variable is any characteristic that *differs* across people or situations. For example, the height of the students in a psychology class is a variable because it varies from student to student. The sex of the students is also a variable as long as there are both male and female students in the class.

Variables are the building blocks of psychological research. Some research questions are descriptive and focus on a single variable, such as how common depression is in college students. Others are associational, asking whether two or more variables are related, such as whether sleep quality is associated with academic performance. Still others are causal, asking whether changes in one variable produce changes in another, such as whether a new type of jury instruction improves jurors' understanding of the law. Understanding how psychologists ask these questions is the first step toward designing meaningful research. Different types of research questions

require different research designs. Throughout this textbook, you'll learn how psychologists choose the most appropriate design to answer each type of question.

In the next sections, we'll examine where research questions come from, what makes some questions stronger than others, and how psychologists transform research questions into specific, testable hypotheses.

2.2 Where do Research Questions Come From?

Research questions rarely appear out of thin air. Instead, they often begin with curiosity about human (or other animal) behavior or mental processes. A researcher may notice an interesting pattern, encounter a practical problem, read about a previous study, or recognize that an existing scientific theory makes a prediction that has not yet been tested. These initial ideas are then refined into research questions that can be investigated scientifically.

2.3 Generating Good Research Questions

2.3.1 Finding Inspiration

Research questions often begin as more general research ideas—usually focusing on some behavior or psychological characteristic: talkativeness, memory, depression, jury verdicts, and so on. Before looking at how to turn such ideas into empirically testable research questions, it is worth looking at where such ideas come from in the first place. Four common sources of inspiration are informal observations, practical problems, previous research, and scientific theories.

2.3.2 Informal Observations

Informal observations include direct observations of our own and others' behavior as well as secondhand observations from nonscientific sources such as newspapers, books, and so on. For example, you might notice that you always seem to be in the slowest moving line at the grocery store. Could it be that most people think the same thing? Or you might read in the local newspaper about people donating money and food to a local family whose house has burned down and begin to wonder about who makes such donations and why. Some of the most famous research in psychology has been inspired by informal observations. Stanley Milgram's famous research on obedience, for example, was inspired in part by journalistic reports of the trials of accused Nazi war criminals—many of whom claimed that they were only obeying orders. This led him to wonder about the extent to which ordinary people will commit immoral acts simply because they are ordered to do so by an authority figure (Milgram, 1963).

2.3.3 Practical Problems

Practical problems can also inspire research ideas, leading directly to applied research in such domains as law, health, education, and sports. For example:

- Can human figure drawings help children remember details about being physically or sexually abused?
- How effective is psychotherapy for depression compared to drug therapy?
- Does using a cell phone impair driving performance?
- What study strategies produce the greatest long-term learning?
- What is the best mental preparation for running a marathon?

Questions like these often lead directly to applied psychological research.

2.3.4 Previous Research

Probably the most common inspiration for new research ideas, however, is previous research. Science is cumulative. Researchers read one another's work, identify unanswered questions, and design new studies that extend existing knowledge. Of course, experienced researchers are familiar with previous research in their area of expertise and probably have a long list of ideas. This suggests that novice researchers can find inspiration by consulting with a more experienced researcher (e.g., students can consult a faculty member). But they can also find inspiration by picking up a copy of almost any professional journal and reading the titles and abstracts. In one issue of *Psychological Science*, for example, you can find articles on the perception of shapes, anti-Semitism, police lineups, the meaning of death, second-language learning, people who seek negative emotional experiences, and many other topics. If you can narrow your interests down to a particular topic (e.g., memory) or domain (e.g., health care), you can also look through more specific journals, such as *Memory & Cognition* or *Health Psychology*. Reading the discussion section of a published research article is often an excellent way to identify ideas for future research because authors typically describe limitations of their studies and suggest important directions for future investigation.

2.3.5 Scientific Theories

Scientific theories are another important source of research questions. As you learned in Chapter 1, theories organize existing knowledge and generate predictions about future observations. Researchers often begin by asking whether a prediction derived from a theory is supported by empirical evidence. If the prediction has not yet been tested, or if previous findings are inconsistent, it becomes an opportunity for new research.

For example, a theory suggesting that sleep improves memory might lead a researcher to ask whether students who obtain eight hours of sleep before an exam perform better than students who stay awake studying. In this way, theories provide a roadmap for future research by identifying questions that remain unanswered.

Regardless of where a research idea originates, it is only the starting point. Researchers must still transform broad ideas into focused, empirical questions that can be answered through scientific observation. In the next section, we'll examine how psychologists develop research questions that are both scientifically meaningful and practically feasible.

2.4 Generating Empirically Testable Research Questions

Once you have a research idea, you need to use it to generate one or more empirically testable research questions, that is, questions expressed in terms of a single variable or relationship between variables. Once you have an idea, ask yourself:

- "What causes this behavior?"
- "What are the consequences of this behavior?"
- "Who is most likely to exhibit this behavior?"
- "Under what conditions does it occur?"

Table 2.1: **Table 2.1** Summary of Research Ideas and Research Questions

Research Idea	**Possible Research Question**
Text anxiety	Does test anxiety predict exam performance?
Alcohol consumption	Does alcohol consumption disrupt long-term memory?
Social media	Who is most likely to feel lonely after looking at social media posts?
Exercise	Does exercise reduce symptoms of depression?

Pause and Reflect: *Think of a topic that interests you, such as personality, social media, romantic relationships, sports, or music. What is one research question a psychologist could investigate about that topic?*

2.5 Evaluating Research Questions

Researchers usually generate many more research questions than they can realistically study. This means they must have some way of evaluating the research questions they generate so that they can choose which ones to pursue. Two important criteria are whether the question is **interesting** and whether it is **feasible** to answer.

2.5.1 Is the Question Interesting?

How often do people tie their shoes? Do people feel pain when you punch them in the jaw? Are women more likely to wear makeup than men? Do people prefer vanilla or chocolate ice cream? Although it would be a fairly simple matter to design a study and collect data to answer these questions, you probably would not want to because they are not interesting. We are not talking here about whether a research question is interesting to us personally but whether it is interesting to people more generally and, especially, to the scientific community. But what makes a research question interesting in this sense? Here we look at three factors that affect the interestingness of a research question: the answer is in doubt, the answer fills a gap in the research literature, and the answer has important practical implications.

First, a research question is interesting to the extent that its answer is in doubt. Obviously, questions that have been answered by scientific research are no longer interesting as the subject of new empirical research. But the fact that a question has not been answered by scientific research does not necessarily make it interesting. There has to be some reasonable chance that the answer to the question will be something that we did not already know. But how can you assess this before actually collecting data? One approach is to try to think of reasons to expect different answers to the question—especially ones that seem to conflict with common sense. If you can think of reasons to expect at least two different answers, then the question might be interesting. If you can think of reasons to expect only one answer, then it probably is not. The question of whether women are more talkative than men is interesting because there are reasons to expect both answers. The existence of the stereotype itself suggests the answer could be yes, but the fact that women's and men's verbal abilities are fairly similar suggests the answer could be no. The question of whether people feel pain when you punch them in the jaw is not interesting because there is little reason to think that the answer could be anything other than a resounding yes.

A second important factor to consider when deciding if a research question is interesting is whether answering it will fill a gap in the research literature. Researchers often identify these gaps by reading published studies and noting the limitations or future directions suggested by the authors.

A final factor to consider when deciding whether a research question is interesting is whether its answer has important practical implications. Questions about effective treatments for

depression, reducing distracted driving, improving eyewitness memory, or enhancing student learning have the potential to improve people's lives.

2.5.2 Is the Question Feasible?

A second important criterion for evaluating research questions is the feasibility of successfully answering them. There are many factors that affect feasibility, including time, money, equipment and materials, technical knowledge and skill, ethical considerations, and access to research participants. Clearly, researchers need to take these factors into account so that they do not waste time and effort pursuing research that they cannot complete successfully.

Looking through a sample of professional journals in psychology will reveal many studies that are complicated and difficult to carry out. These include longitudinal designs in which participants are tracked over many years, neuroimaging studies in which participants' brain activity is measured while they carry out various mental tasks, and complex non-experimental studies involving several variables and complicated statistical analyses. Keep in mind, though, that such research tends to be carried out by teams of highly trained researchers whose work is often supported in part by government and private grants. Keep in mind also that research does not have to be complicated or difficult to produce interesting and important results. Looking through a sample of professional journals will also reveal studies that are relatively simple and easy to carry out—perhaps involving a convenience sample of college students and a paper-and-pencil task.

A final point here is that it is generally good practice to use methods that have already been used successfully by other researchers. For example, if you want to manipulate people's moods to make some of them happy, it would be a good idea to use one of the many approaches that have been used successfully by other researchers (e.g., paying them a compliment). This is good not only for the sake of feasibility—the approach is "tried and true"—but also because it provides greater continuity with previous research. This makes it easier to compare your results with those of other researchers and to understand the implications of their research for yours, and vice versa.

2.5.3 Thinking Like a Researcher

When evaluating a possible research question, ask yourself:

- Is the answer genuinely unknown?
- Does the question build on previous research or theory?
- Does it have scientific or practical importance?
- Can it be answered ethically?

- Do I have the time, resources, and participants needed to answer it?

If the answer to these questions is "yes," then the research question is probably worth pursuing.

2.6 From Research Questions to Hypotheses

A good research question identifies what a psychologist wants to understand, but it does not specify what the researcher expects to find. To answer a research question scientifically, psychologists develop one or more hypotheses. A hypothesis is a specific, testable prediction about the relationship between variables or about the outcome of a study. Unlike a research question, which asks a question, a hypothesis proposes a possible answer that can be evaluated with empirical evidence.

For example, consider the following research question:

> Does sleep affect academic performance?

This question identifies an important topic, but it does not predict what the researcher expects to observe. One possible hypothesis is:

> College students who sleep between seven and nine hours the night before an exam will earn higher exam scores than students who sleep fewer than six hours.

Notice that this hypothesis is much more specific than the research question. It identifies the variables being studied, predicts the expected relationship between them, and can be tested through observation. Throughout the remainder of this chapter, you will learn how psychologists develop strong hypotheses that can be investigated using scientific research methods.

2.7 Characteristics of Good Hypotheses

Not all hypotheses are equally useful. A strong hypothesis provides a clear prediction that can be tested through scientific research. Well-written hypotheses help researchers design studies, select appropriate variables, and interpret their findings. Although there is no single formula to writing a hypothesis, effective hypotheses share several important characteristics.

2.7.1 A Good Hypothesis is Testable

A hypothesis must be capable of being evaluated using empirical evidence. Researchers should be able to collect observations or measurements that either support or do not support the prediction. If no possible evidence could demonstrate that a hypothesis is incorrect, it cannot be tested scientifically.

2.7.2 A Good Hypothesis is Specific

A useful hypothesis clearly identifies the variables being studied and predicts the expected relationship between them. Vague hypotheses make it difficult to determine what should be measured or how the results should be interpreted.

2.7.3 A Good Hypothesis is Falsifiable

As you learned in Chapter 1, scientific hypotheses must be capable of being shown to be incorrect if contradictory evidence exists. Research should be able to describe what findings would cause them to question or reject the hypothesis.

2.7.4 A Good Hypothesis is Grounded in Existing Knowledge

Researchers do not develop hypotheses at random. Most hypotheses are based on previous research, established scientific theories, or systematic observations. Building on existing knowledge helps ensure that research contributes meaningfully to scientific understanding.

2.7.5 A Good Hypothesis Makes a Prediction

Unlike a research question, which asks what the researcher wants to understand, a hypothesis predicts what the researcher expects to find. Although predictions are not always correct, they provide a clear expectation that can be evaluated through research.

When researchers write strong hypotheses, they create a clear roadmap for the rest of the research process. A well-written hypothesis helps determine what variables should be measured, what research design is appropriate, and how the results should be interpreted.

Although research hypotheses often propose a relationship, difference, or effect of one variable on another, a well-designed study can also test the prediction that there will be no difference or no relationship. For example, in clinical psychology, demonstrating that two different types of therapy provide equivalent mental health outcomes is an important scientific finding.

Table 2.2: **Table 2.2** Examples of Poor and Better Hypotheses

Poor Hypothesis	Why it Needs Improvement	Better Hypothesis
Exercise is good for people.	Too vague. What kind of exercise? Good in what way? For whom?	Adults who engage in at least 150 minutes of moderate exercise per week will report lower depression scores than adults who engage in less than 60 minutes per week.
Social media affects teenagers.	Doesn't specify which social media or what outcome is affected.	Teenagers who spend more than 4 hours per day on social media will report higher levels of loneliness than teenagers who spend less than 1 hour per day.
Sleep helps students.	Doesn't specify how much sleep or what outcome.	College students who sleep at least 8 hours the night before an exam will score higher on the exam than students who sleep fewer than 6 hours.

Pause and Reflect: *Think back to your research question. Can you rewrite it as a specific, testable hypothesis?*

2.8 Identifying Variables

Earlier in this chapter, you learned that research questions are built around variables - characteristics, behaviors, or conditions that can vary across people, situations, or time. Once a researcher has developed a hypothesis, the next step is to identify the variables it contains.

Consider the following hypothesis:

> College students who sleep at least eight hours the night before an exam will earn higher exam scores than students who sleep fewer than six hours.

This hypothesis contains two variables: sleep duration and exam performance.

Identifying the variables is only the beginning. Researchers must also determine exactly how each variable will be measured or manipulated. For example, should sleep duration be measured

using a sleep diary, a smartwatch, or participants' self-reports? Should exam performance be measured using a classroom exam, a standardized test, or a laboratory memory task?

These questions illustrate an important feature of psychological research. Many concepts studied by psychologists, such as stress, intelligence, prejudice, motivation, and happiness, cannot be observed directly. Before these concepts can be studied scientifically, researchers must decide exactly how they will be measured or manipulated. These precise descriptions are called ***operational definitions***, which are discussed in Module 2.

2.9 Types of Hypotheses

Not all hypotheses make the same kind of prediction. Some hypotheses predict the direction of a relationship, whereas others simply predict that a relationship or difference exists.

A ***directional hypothesis*** predicts both that a relationship exists and the direction of that relationship. For example:

> College students who sleep at least eight hours the night before an exam will earn higher exam scores than students who sleep fewer than six hours.

This hypothesis predicts that students who sleep more will perform *better* on the exam.

A ***nondirectional hypothesis*** predicts that a relationship or difference exists but does not predict which direction the results will go. For example:

Sleep duration will be related to exam performance.

Researchers often use nondirectional hypotheses when previous research is limited or when there is insufficient evidence to predict the direction of the relationship.

Every research hypothesis also has a corresponding ***null hypothesis***, which predicts that no relationship or difference exists between the variables. Researchers use null hypotheses when conducting statistical analyses, a topic that will be covered in a later module.

2.10 Independent and Dependent Variables

Once researchers have developed a hypothesis, they must identify the role of each variable. In experimental research, the variable that the researcher manipulates is called the independent variable (IV). The variable that is measured to determine whether it changes is called the dependent variable (DV).

Recall our hypothesis that:

> College students who sleep at least eight hours the night before an exam will earn higher exam scores than students who sleep fewer than six hours.

In this example, the independent variable is sleep duration, and the dependent variable is exam score. A helpful way to remember the difference is to ask yourself, "Which variable is expected to influence the other?" Which variables is the "independent leader?" The variable leading in this example is the independent variables. The variable expected to follow, or change, is the dependent variable.

Remember that an independent variable is manipulated in experimental research. Thus in our example, a researcher could randomly assign some college students to sleep at least eight hours the night before an exam and randomly assign other college students to sleep fewer than six hours before an exam. Then the exam performances between the two groups of students would be compared.

Pause and Reflect: If you wanted to test whether listening to music while studying affects exam performance, what would be the independent variable? What would be the dependent variable?

2.11 Predictor and Outcome Variables

If you are thinking that it doesn't sound very easy or even ethical to ask some students to get eight hours of sleep and others to get less before an exam, you're right. Many times it doesn't make sense to manipulate a variable of interest. In many psychological studies, researchers simply measure variables as they naturally occur. For example, researchers cannot ethically assign people to experience different levels of traumatic experiences or to different levels of good parenting. Instead, they measure these variables and examine how they relate to one another.

In these studies, researchers often use the terms predictor variables and outcome variables rather than independent and dependent variables.

A ***predictor variable*** is a measured variable, and used to predict or explain another variable. An ***outcome variable*** is a measured variable being predicted. Notice that both variables are measured variables - nothing is manipulated by the researcher.

For example, imagine that researchers want to know whether stress predicts college students' grade point average (GPA). The predictor variable is stress, and the outcome variable is GPA.

As you learn more about research methods, you'll discover that independent and dependent variables correspond with experiments, and predictor and outcome variables correspond with nonexperimental or correlational research.

2.12 Putting it All Together

So far, you've learned how psychologists move from a broad idea to a testable study. The process begins with a scientific theory or an observation that inspires a research question. The researcher then develops a hypothesis, identifies the variables involved, and determines the role of each variable in the study. The remaining chapters of this textbook will build on these ideas as you learn how psychologists design studies, collect data, and draw conclusions about behavior and mental processes. See Figure 2.1.

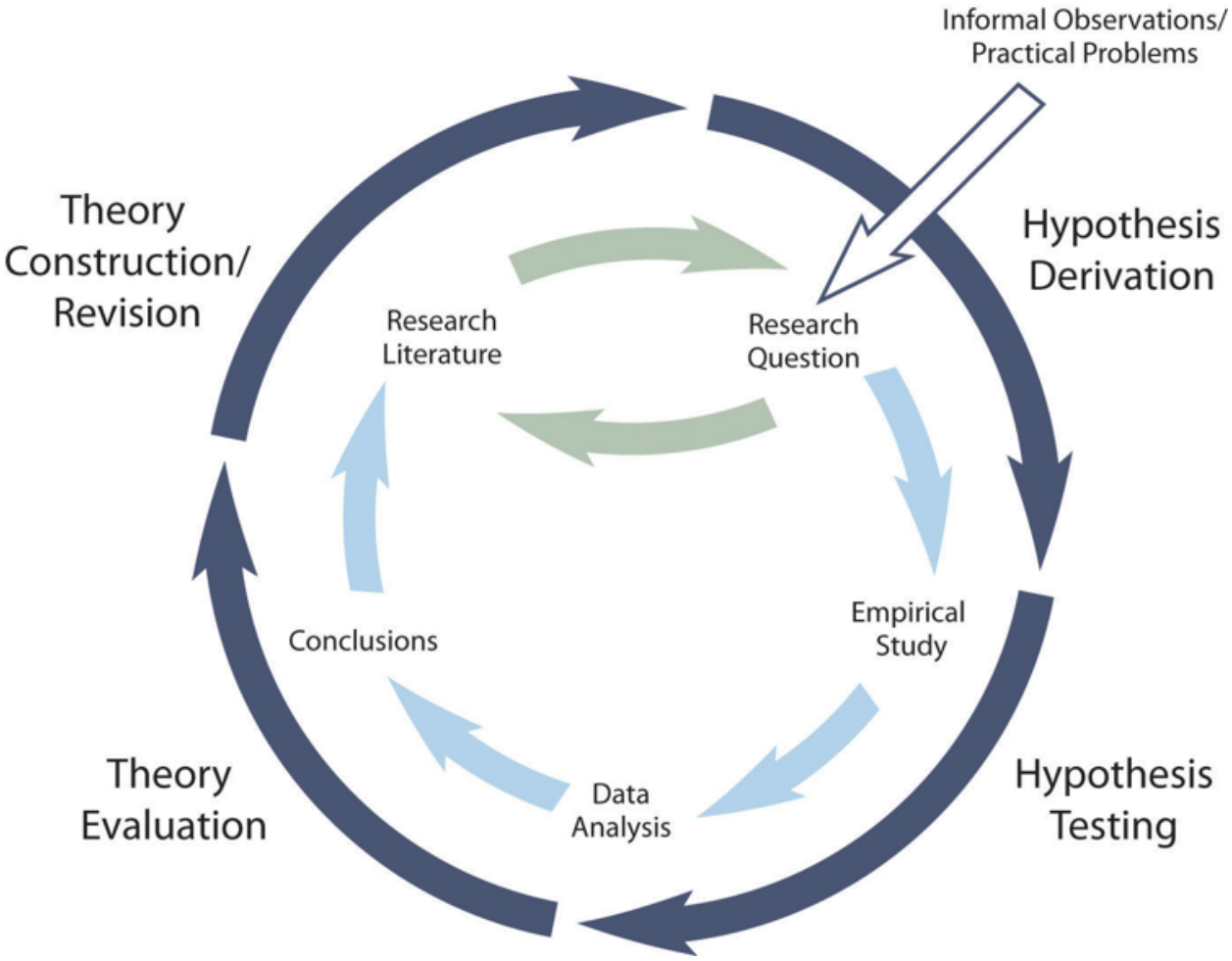


Figure 2.1: The theory data cycle

Figure 2.1 *The Theory Data Cycle*[1]

Pause and Reflect Answer: *If you wanted to test whether listening to music while studying affects exam performance, what would be the independent variable? What would be the dependent*

variable?

In this example, the independent variable is whether students listen to music while studying. This is the variable the researcher changes. The dependent variable is exam performance, usually measured by the students' scores on an exam. This is the outcome variable the researcher measures to determine whether the independent variable had an effect.

2.13 Media Attributions

[1] *The theory data cycle*. Developing a hypothesis. In Research Methods in Psychology (4th ed.), by R.S. Jhangiani, I.-C.A. Chiang, C. Cuttler, & D.C. Leighton (2019). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

2.14 Text Attributions

Dudley, M. (2019, January 24). Research methods in psychology". OER Commons. https://oercommons.org/authoring/51456-research-methods-in-psychology Licensed under CC BY-NC-SA 4.0. Modified by current authors.

2.15 References

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*(4), 371-378. https://doi.org/10.1037/h0040525

3 Chapter 3: Finding and Evaluating Psychological Research

3.1 Why Conduct a Literature Review?

Recall that one of the most common sources of inspiration for a research question is previous research. Therefore, it is important to review the literature early in the research process. The ***research literature*** in any field is all the published research in that field. Reviewing the research literature means finding, reading, and summarizing the published research relevant to your topic of interest. In addition to helping you discover new research questions, reviewing the literature early in the research process can help you in several other ways.

- It can tell you if a research question has already been answered.
- It can help you evaluate the interestingness of a research question.
- It can give you ideas for how to conduct your own study.
- It can tell you how your study fits into the research literature.

Reviewing the literature may seem overwhelming at first, especially when you discover hundreds or even thousands of articles on your topic. The good news is that experienced researchers rarely read every article they find. Instead, they use effective search strategies to quickly locate the most relevant and highest-quality research. This chapter will show you how to search efficiently so you can spend less time searching and more time learning.

Although the research literature includes millions of publications, not all sources are equally valuable when conducting a literature review. Psychologists rely primarily on peer-reviewed journal articles and scholarly books because these sources provide the strongest scientific evidence.

Pause and Reflect: Imagine you're studying the effects of caffeine on anxiety. Why might it be a mistake to conduct a study before reviewing what other researchers have already discovered?

3.2 Professional Journals

Professional journals are periodicals that publish original research articles. There are thousands of professional journals that publish research in psychology and related fields. Most articles in professional journals are one of two basic types: empirical research reports and review articles. ***Empirical research* reports** describe one or more new empirical studies conducted by the authors. They introduce a research question, explain why it is interesting, review previous research, describe their method and results, and draw their conclusions. ***Review articles*** summarize previously published research on a topic and usually present new ways to organize or explain the results. When a review article is devoted primarily to presenting a new theory, it is often referred to as a theoretical article. When a review article provides a statistical summary of all the previous results it is referred to as a ***meta-analysis.*** Table 3.1 summarizes the characteristics of empirical research reports, review articles, and meta-analyses.

Table 3.1: **Table 3.1** Types of Sources

Type of Source	**Purpose**	**Example**
Empirical Article	Reports a new study	Researchers examine whether sleep improves memory.
Review article	Summarizes many studies	Overview of research on stress and college students
Meta-analysis	Combines results statistically	Calculates the average effect of mindfulness interventions across hundreds of experiments

3.3 How do researchers get their work published?

Most professional journals in psychology undergo a process of ***double-blind peer review***. Researchers who want to publish their work in the journal submit a manuscript to the editor—who is generally an established researcher too—who in turn sends it to two or three experts on the topic. Each reviewer reads the manuscript, writes a critical but constructive review, and sends the review back to the editor along with recommendations about whether the manuscript should be published or not. The editor then decides whether to accept the article for publication, ask the authors to make changes and resubmit it for further consideration, or reject it outright. In any case, the editor forwards the reviewers' written comments to the researchers so that they can revise their manuscript accordingly. This entire process is double-blind, as the reviewers do not know the identity of the researcher(s) and vice versa. Double-blind peer review is helpful because it ensures that the work meets basic standards of the field before it can enter

the research literature. However, in order to increase transparency and accountability, some journals now utilize an open peer review process wherein the identities of the reviewers (which remain concealed during the peer review process) are published alongside the journal article.

Because of peer review, journal articles are generally more trustworthy than websites, blogs, or social media posts. Although peer review isn't perfect, it helps ensure that published research meets the standards of the scientific community before becoming part of the scientific literature.

3.4 Scholarly Books

Scholarly books are books written by researchers and practitioners mainly for use by other researchers and practitioners. A ***monograph*** is written by a single author or a small group of authors and usually gives a coherent presentation of a topic much like an extended review article. ***Edited volumes*** have an editor or a small group of editors who recruit many authors to write separate chapters on different aspects of the same topic. In general, scholarly books undergo a peer review process similar to that used by professional journals.

3.5 Literature Search Strategies

3.5.1 A Step-by-Step Literature Search

Finding relevant research articles usually takes more time than just typing a topic into a database and finding exactly what you want. Using effective searching strategies can help you locate higher quality sources, save time, and ensure you are finding relevant literature for you.

3.5.1.1 Step 1: Start with your research question.

Example: Does social media affect loneliness among college students?

3.5.1.2 Step 2: Identify keywords.

Table 3.2: **Table 3.2** Examples of main ideas and keywords

Main Idea	Keywords
Social Media	Instagram, TikTok, Snapchat, Facebook
Loneliness	Isolation, belonging, social connection

Main Idea	Keywords
College Students	Undergraduate, university students

3.5.1.3 Step 3: Use Boolean Operators.

Boolean operators help you broaden or narrow your search by combining keywords.

- **AND** narrows your search by returning articles that contain both search terms
  - Example: social media AND loneliness
  - This search will return articles that discuss both social media and loneliness.
- **OR** broadens your search by returning articles that contain any of the search terms. This is especially useful for searching synonyms or related items.
  - Example: social media OR TikTok OR Instagram OR Facebook
  - This search will return articles that mention any of these terms.

Putting it all together, our search terms might look like this:

> (Social media OR TikTok OR Instagram OR Facebook) AND (loneliness OR social isolation)

3.5.1.4 Step 4: Search PsycINFO.

PsycINFO is a database produced by the American Psychological Association (APA). It consists of individual records for each article, book chapter, or book in the database. Each record includes basic publication information, an abstract or summary of the work (like the one presented at the start of this chapter), and a list of other works cited by that work. A computer interface allows entering one or more search terms like modeled above, and returns any records that contain those search terms. Each record also contains lists of keywords that describe the content of the work and also a list of index terms. The index terms are especially helpful because they are standardized. Research on differences between females and males, for example, is always indexed under "Human Sex Differences." Research on note-taking is always indexed under the term "Learning Strategies." If you do not know the appropriate index terms, PsycINFO includes a thesaurus that can help you find them.

Given that there are nearly four million records in PsycINFO, you may have to try a variety of search terms in different combinations and at different levels of specificity before you find what you are looking for. Imagine, for example, that you are interested in the question of whether males and females differ in terms of their ability to recall experiences from when they were very

young. If you were to enter the search term "memory," it would return far too many records to look through individually. This is where the thesaurus helps. Entering "memory" into the thesaurus provides several more specific index terms—one of which is "early memories." While searching for "early memories" among the index terms still returns too many to look through individually—combining it with "human sex differences" as a second search term returns fewer articles, many of which are highly relevant to the topic.

Another helpful tip is to filter your results.

- To find empirical journal articles only, select Scholarly/Peer reviewed
- To find full text articles, select Online (Full Text) Only

If your search returns too many articles, try adding another keyword or using more specific search terms. If it returns too few articles, broaden your search by removing a keyword or adding synonyms with OR.

3.6 Other Sources of Research

3.6.1 Open Science Framework (OSF)

3.6.1.1 What is OSF?

It is a free online platform for researchers and students. OSF provides access to research materials, datasets, preprints, and study registrations. It allows researchers to share and organize their work, can be used to find research resources, and learn from existing studies. It promotes openness and collaboration in research.

3.6.2 Google Scholar

Google Scholar is a free search engine to access scholarly literature. Google Scholar's "Cited by" feature is a powerful way to discover newer research on a topic. When you click "Cited by" beneath an article, Google Scholar displays more recent publications that have references to that article in their own research. This allows you to see how the scientific conversation has evolved over time and quickly identify newer studies related to your topic.

3.6.2.1 How to access Google Scholar

1. Search **Google Scholar** or go directly to scholar.google.com
2. Search to find sources
3. Use the **Cited by** feature in Google Scholar to find newer studies that have referenced an article.

Here is a helpful tutorial: How to Use Google Scholar

3.7 AI: Benefits

AI can be a valuable tool during the research process when used appropriately and ethically. AI can help generate keywords and search terms for literature searches. Research-focused AI databases (e.g., **Consensus**) can help identify relevant scholarly articles, generate keywords for literature searches, and summarize research findings. AI can also assist in brainstorming research questions, refining topics, and discovering related concepts.

Helpful Videos:

Using AI for literature review - ethically!

Figure 3.1: Using AI for literature review - ethically!

Consensus vs ChatGPT - using AI for Literature Review and Research

Figure 3.2: Consensus vs ChatGPT - using AI for Literature Review and Research

The Best Free AI Tools for Research (2026)

Figure 3.3: The Best Free AI Tools for Research (2026)

3.8 AI: Risks

Although AI can be helpful, it also has important limitations that researchers should understand. General AI chatbots such as ChatGPT can sometimes generate inaccurate information or hallucinate sources that do not actually exist. Due to this be sure to check your sources through reliable databases such as PsycINFO or Google Scholar, and read the original source before citing it or drawing conclusions from it. AI generated summaries may also omit important details, misinterpret findings, or oversimplify complex research. Always verify any AI-generated summaries by consulting the original source. Additionally, relying too much on AI can hinder the development of critical thinking and research skills. To also maintain academic integrity, AI should be used as a tool, not a replacement for your own thinking, analysis, and writing.

Helpful Videos:

How to Stop AI from Killing Your Critical Thinking | Advait Sarkar

Figure 3.4: How to Stop AI from Killing Your Critical Thinking | Advait Sarkar | TED

Evaluating Sources using the CRAAP test

Figure 3.5: Evaluating Sources using CRAAP

3.9 Finding More Articles

In addition to entering search terms into PsycINFO and other databases, there are several other techniques you can use to search the research literature. First, if you have one good article or book chapter on your topic—a recent review article is best—you can look through the reference list of that article for other relevant articles, books, and book chapters. In fact, you should do this with any relevant article or book chapter you find. You can also start with a classic article or book chapter on your topic, find its record in PsycINFO (by entering the author's name or article's title as a search term), and link from there to a list of other works in PsycINFO that cite that classic article. This works because other researchers working on your topic are likely to be aware of the classic article and cite it in their own work. You can also do a general Internet search using search terms related to your topic or the name of a researcher who conducts research on your topic. This might lead you directly to works that are part of the research literature (e.g., articles in open-access journals or posted on researchers' own websites). The search engine Google Scholar is especially useful for this purpose. A general Internet search might also lead you to websites that are not part of the research literature but might provide references to works that are. Finally, you can talk to people (e.g., your instructor or other faculty members in psychology) who know something about your topic and

can suggest relevant articles and book chapters.

3.10 Read Articles Efficiently

Research articles can seem intimidating because they often contain technical language and detailed statistical analyses. Fortunately, you do not need to read every article from beginning to end before deciding whether it is useful.

A good strategy is to read the sections in this order:

1. Abstract: What was the study about?
2. Introduction: Why was the study conducted?
3. Discussion: What did the researchers conclude?
4. Results: What did they specifically find?
5. Method: How was the study conducted?

Reading articles in this order will help you quickly determine whether an article is relevant to your research question before investing time reading every detail.

Conducting a literature review is much more than collecting articles. It is the process of discovering what researchers already know, identifying unanswered questions, and building a foundation for your own research. As you gain experience, searching the literature will become faster and more efficient. The time you invest in finding high-quality research will be well worth it.

Pause and Reflect: *If you found 500 articles on your topic, would you read every one? Based on this chapter, what strategies would you use to narrow your search?*

3.11 Text Attributions

4 Chapter 4: Conducting Ethical Research

4.1 Why Research Ethics Matter

Imagine you have developed an interesting research question, found a gap in the scientific literature, and written a strong hypothesis. Your study is well designed and has the potential to contribute to our understanding of human behavior. Before collecting a single piece of data, however, the most important question remains: Should this study be conducted?

Not every scientifically interesting study is ethically acceptable. Researchers have a responsibility to protect the rights, safety, and well-being of the people and animals who participate in their research. In psychology, conducting good research means more than using sound scientific methods - it also means treating participants with honesty, fairness, and respect.

Unfortunately, the importance of research ethics was learned through painful mistakes. Throughout history, some researchers have placed scientific discovery above the wellbeing of participants, leading to studies that caused unnecessary harm and violated basic human rights. These cases changed the way research is conducted and led to the ethical principles and oversight procedures that guide psychological research today.

4.2 Historical Cases

4.2.1 Nazi Medical Experiments

One of the darkest chapters in the history of scientific research occurred during World War II, when physicians and scientists in Nazi Germany conducted cruel and often fatal experiments on prisoners held in concentration camps. Many of the victims were Jewish people, but the experiments also targeted Romani people, people with disabilities, political prisoners, Soviet prisoners of war, and other groups persecuted by the Nazi regime. Rather than being treated as human beings with inherent dignity and rights, these individuals were dehumanized and viewed as expendable for the sake of advancing the Nazi regime's political and ideological goals. Prisoners were subjected to extreme cold, high-altitude conditions, infectious diseases, forced sterilization, and experimental surgeries often performed without anesthesia, and other inhumane procedures without their knowledge or consent. Many suffered permanent injuries or died as a result of these experiments.

Pause and Reflect: *What allowed researchers to justify treating certain groups of people so differently from others? Why is recognizing the dignity and worth of every participant essential to ethical research?*

4.2.2 Tuskegee Syphilis Study

Although the Nazi medical experiments are among the most infamous examples of unethical research, serious ethical violations have also occurred in the United States. One of the best-known examples is the ***Tuskegee Syphilis Study***, conducted by the U.S. Public Health Service between 1932 and 1972 - a period of 40 years (Reverby, 2009). The study enrolled hundreds of poor African American men in rural Alabama, many of whom had syphilis. However, the participants didn't know this. Instead, they were told they had "**bad blood**" and that they would receive treatment. This was a lie. The true purpose of the study was to observe the natural progression of the disease without providing effective treatment. Even after penicillin became the standard and highly effective treatment for syphilis in the 1940s, researchers intentionally withheld treatment and prevented many participants from receiving it elsewhere. As a result, many men experienced unnecessary suffering, developed serious health complications, transmitted the disease to their wives, and died. The Tuskegee Syphilis Study demonstrated the importance of honesty, informed consent, justice, and respect for participants, and it led to major reforms in the oversight of research involving human participants.

Pause and Reflect: *If you had been one of the participants and learned years later that an effective treatment had existed but had been intentionally withheld from you, how do you think that would affect your trust in researchers and the medical community?*

4.2.3 Milgram's Obedience Study

Not all ethically important research involved physical harm or the deliberate exploitation of participants. In the early 1960s, psychologist Stanley Milgram conducted a series of studies to examine whether ordinary people would obey an authority figure, even when doing so appeared to harm another person (Milgram, 1963). Participants believed they were administering increasingly painful electric shocks to another individual whenever an incorrect answer was given on a memory task. In reality, no shocks were delivered - the "learner" was a research confederate, playing a role that relied on deception.

Milgram found that many participants continued administering what they believed were dangerous shocks when instructed to do so by the experimenter. The findings provided important insights into the power of authority and have helped psychologists understand events such as the Holocaust and other historical atrocities in which ordinary people obeyed harmful orders.

Unlike the Nazi experiments and the Tuskegee study, Milgram's participants were debriefed and suffered no lasting physical harm. Nevertheless, many psychologists questioned whether

exposing participants to such intense psychological distress was ethically justified. Milgram's work illustrates that ethical decisions are not always clear-cut. Researchers must often weigh the potential scientific benefits of a study against the possible risks to participants.

Although these historical cases differed in important ways, they shared a common lesson: scientific research must be guided by ethical principles that protect participants while allowing important discoveries to be made. In response to these and other events, researchers, professional organizations, and governments developed ethical guidelines that continue to shape psychological research today. We'll review these next.

Pause and Reflect: *If you had served on an ethics review board in the 1960s, would you have approved Milgram's study? Why or why not? What changes, if any, would you require before allowing the study to proceed?*

4.3 A Framework for Thinking about Research Ethics

Table 4.1 presents a framework for thinking through the ethical issues involved in psychological research. The rows of Table 4.1 represent four general moral principles that apply to scientific research: weighing risks against benefits, acting responsibly and with integrity, seeking justice, and respecting people's rights and dignity. (These principles are adapted from those in the American Psychological Association [APA] Ethics Code.) The columns of Table 4.1 represent three groups of people that are affected by scientific research: the research participants, the scientific community, and society more generally. The idea is that a thorough consideration of the ethics of any research project must take into account how each of the four moral principles applies to each of the three groups of people.

Table 4.1: A Framework for Thinking about Ethical Issues in Scientific Research[1]. A table that details several moral principles and who would be affected with the implementation of scientific research.

	Who is affected?		
Moral principle	Research participants	Scientific community	Society
Weighing risks against benefits			
Acting responsibly and with integrity			
Seeking justice			
Respecting people's rights and dignity			

4.4 The Four Ethical Principles of Research

4.4.1 Principle 1 - Scientific research must weigh risks against benefits

Scientific research in psychology is ethical only if its potential benefits outweigh its potential risks. Among the risks to participants are that a treatment might fail to help or even be harmful, a procedure might result in physical or psychological harm, or their right to privacy might be violated. Among the potential benefits are receiving helpful treatment, learning about psychology, experiencing the satisfaction of contributing to scientific knowledge, and receiving money or course credit for participating.

Scientific research can have risks and benefits to the scientific community and to society too. A risk to science is that if a research question is uninteresting or a study is poorly designed, then the time, money, and effort spent on that research could have been spent on more productive research. A risk to society is that research results could be misunderstood or misapplied with harmful consequences. For example, a now-discredited study that falsely linked the measles, mumps, and rubella (MMR) vaccine to autism greatly contributed to vaccine hesitancy and reduced vaccination rates, despite overwhelming scientific evidence showing no such link.

In practice, researchers must carefully consider these risks and benefits before a study begins. Institutional Review Boards (IRBs) review proposed studies to help ensure that participants are protected and that the potential benefits of the research justify any risks.

4.4.2 Principle 2 - Acting Responsibly and With Integrity

Researchers must act responsibly and with integrity. This means carrying out their research in a thorough and competent manner, meeting professional obligations, and being truthful throughout the research process. Acting with integrity is essential because it promotes trust, which is the foundation of ethical research. Participants must be able to trust that researchers are being honest about what a study involves, will keep their promises (such as maintaining confidentiality), and will conduct their research in ways that maximize benefits and minimize risks.

Integrity also requires researchers to collect, analyze, and report their findings honestly. Fabricating data, falsifying results, or plagiarizing the work of others undermines public trust and damages the scientific process.

4.4.3 Principle 3 - Seeking Justice

Researchers must conduct their research in a just manner. They should treat their participants fairly, for example, by giving them adequate compensation for their participation and making sure that benefits and risks are distributed across all participants. For example, in a study of a

new and potentially beneficial psychotherapy, some participants might receive the psychotherapy while others serve as a control group that receives no treatment. If the psychotherapy turns out to be effective, it would be fair to offer it to participants in the control group when the study ends.

Justice also means ensuring that the burdens and benefits of research are shared fairly. Researchers should avoid exploiting vulnerable populations simply because they are more accessible or less able to refuse participation.

4.4.4 Principle 4 - Respecting People's Rights and Dignity

Researchers must respect people's rights and dignity as human beings. One element of this is respecting participants' ***autonomy***—their right to make their own decisions and choose whether to participate in research free from coercion or undue pressure. This principle is reflected in the practice of ***informed consent***, which requires researchers to explain a study and obtain participants' voluntary agreement before they take part.

Another element of respecting people's rights and dignity is respecting their ***privacy***—their right to decide what information about them is shared with others. This means that researchers must maintain ***confidentiality***, which is essentially an agreement not to disclose participants' personal information without their consent or some appropriate legal authorization.

These four principles provide the ethical foundation for psychological research. In practice, however, ethical decisions are not always simple. Researchers must often balance competing priorities, such as advancing scientific knowledge while protecting participants from harm. To help researchers make these decisions consistently, professional organizations and governments have developed formal ethics codes and oversight procedures.

4.5 From Principles to Practice: Ethics Codes

The four ethical principles of weighing risks against benefits, acting with integrity, seeking justice, and respecting people's rights and dignity provide a useful framework for thinking about ethical research. However, principles alone do not tell researchers exactly what to do in every situation. To help researchers make consistent ethical decisions, governments and professional organizations have developed formal ethics codes and oversight procedures. These guidelines help protect participants while allowing important scientific discoveries to continue.

4.5.1 Historical Overview

4.5.1.1 The Nuremberg Code

The first major code of research ethics was the ***Nuremberg Code***, developed in 1947 following the trials of Nazi physicians who conducted horrific experiments on concentration camp prisoners during World War II. The Nuremberg Code established that **participation in research must be voluntary** and based on ***informed consent***. It also emphasized that researchers should minimize unnecessary physical and psychological harm, ensure that the potential benefits of research outweigh its risks, and allow participants to withdraw from a study at any time. Although originally developed in response to the Nazi medical experiments, the Nuremberg Code became the foundation for modern research ethics and influenced many of the ethical guidelines that followed.

4.5.2 The Declaration of Helsinki

In 1964, the World Medical Association expanded upon the Nuremberg Code by adopting the Declaration of Helsinki. This document emphasized that research involving human participants should follow a carefully written research protocol and be reviewed by an independent committee before data collection begins. These ideas helped establish the practice of independent ethical review that is now common in research institutions around the world.

4.5.3 The Belmont Report

In the United States, public concern following the Tuskegee Syphilis Study and other unethical research led to the publication of the Belmont Report in 1979. The Belmont Report became the foundation of modern research ethics in the United States and identified three core principles that should guide research involving human participants:

1. **Respect for Persons**: Individuals should be treated as autonomous decision makers and should voluntarily choose whether to participate in research. This principle is the basis for informed consent.
2. **Beneficence**: Researchers should maximize the potential benefits of research while minimizing possible risks and harm to participants.
3. **Justice**: The benefits and burdens of research should be distributed fairly, and vulnerable populations should not be exploited simply because they are more accessible or less able to refuse participation.

Although the wording differs slightly, these principles closely parallel the ethical principles discussed earlier in this module. Together, they continue to guide the ethical conduct of research in psychology and many other scientific disciplines.

4.6 Institutional Review Boards

Today, most colleges, universities, hospitals, and research institutions have an ***Institutional Review Board (IRB)***, a committee responsible for reviewing research involving human participants before data collection begins. Importantly, the IRB committee consists of a diverse group of individuals with different backgrounds, areas of expertise, and perspectives. Typically, an IRB includes scientists and nonscientists, as well as at least one member who is not affiliated with the institution. This diversity in membership helps ensure that research proposals are evaluated from multiple viewpoints, making it more likely that potential ethical concerns will be recognized before a study begins. Furthermore, because everyone is susceptible to cognitive biases, a diverse committee is also less likely than a single reviewer or a homogeneous group to overlook important ethical concerns because of shared assumptions or blind spots.

Researchers submit a detailed research proposal describing the purpose of the study, the procedures, potential risks and benefits, and the steps they will take to protect participants. The IRB evaluates whether the study meets ethical standards and may approve the study, require modifications, or determine that it cannot be conducted as proposed.

4.7 APA Ethics Code

Although the Nuremberg Code, Declaration of Helsinki, and Belmont Report provide broad ethical guidance, psychologists in the United States also follow the **[American Psychological Association (APA) Ethical Principles of Psychologists and Code of Conduct]**. This document provides detailed standards for psychologists in research, teaching, clinical practice, and other professional activities. For researchers, the most relevant section is Standard 8: Research and Publication, which outlines ethical responsibilities before, during, and after a study.

Several parts of the APA Ethics Code are especially important for beginning researchers. In the following sections, we will examine informed consent, deception, debriefing, research with nonhuman animals, and scholarly integrity. Together, these standards help ensure that psychological research is conducted ethically, responsibly, and with respect for participants.

4.7.1 Informed Consent

Standards 8.02 to 8.05 are about informed consent. Again, ***informed consent*** means obtaining and documenting people's agreement to participate in a study, having informed them of everything that might reasonably be expected to affect their decision. This includes details of the procedure, the risks and benefits of the research, the fact that they have the right to decline to participate or to withdraw from the study, the consequences of doing so, and any

legal limits to confidentiality. For example, some states require researchers who learn of child abuse or other crimes to report this information to authorities.

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

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

4.7.2 Deception

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

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

Note, however, that the APA Ethics Code takes a more moderate approach—allowing deception when the benefits of the study outweigh the risks, participants cannot reasonably be expected

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

4.7.3 Debriefing

Standard 8.08 is about ***debriefing***. This is the process of informing participants as soon as possible of the purpose of the study, revealing any deception, and correcting any other misconceptions they might have as a result of participating. Debriefing also involves minimizing harm that might have occurred. For example, an experiment on the effects of being in a sad mood on memory might involve inducing a sad mood in participants by having them think sad thoughts, watch a sad video, or listen to sad music. Debriefing would be the time to return participants' moods to normal by having them think happy thoughts, watch a happy video, or listen to happy music.

4.7.4 Nonhuman Animal Subjects

Standard 8.09 is about the humane treatment and care of nonhuman animal subjects. Although most contemporary research in psychology does not involve nonhuman animal subjects, a significant minority of it does—especially in the study of learning and conditioning, behavioral neuroscience, and the development of drug and surgical therapies for psychological disorders.

The use of nonhuman animal subjects in psychological research is like the use of deception in that there are those who argue that it is rarely, if ever, ethically acceptable (Bowd & Shapiro, 1993). Clearly, nonhuman animals are incapable of giving informed consent. Yet they can be subjected to numerous procedures that are likely to cause them suffering. They can be confined, deprived of food and water, subjected to pain, operated on, and ultimately euthanized. (Of course, they can also be observed benignly in natural or zoolike settings.) Others point out that psychological research on nonhuman animals has resulted in many important benefits to humans, including the development of behavioral therapies for many disorders, more effective pain control methods, and antipsychotic drugs (Miller, 1985). It has also resulted in benefits to nonhuman animals, including alternatives to shooting and poisoning as means of controlling them.

As with deception, the APA acknowledges that the benefits of research on nonhuman animals can outweigh the costs, in which case it is ethically acceptable. However, researchers must use alternative methods when they can. When they cannot, they must acquire and care for their subjects humanely and minimize the harm to them. For more information on the APA's position on nonhuman animal subjects, see the website of the APA's Committee on Animal Research and Ethics.

4.7.5 Scholarly Integrity

Standards 8.10 to 8.15 are about scholarly integrity. These include the obvious points that researchers must not fabricate data or plagiarize. Plagiarism means using others' words or ideas without proper acknowledgment. Proper acknowledgment generally means indicating direct quotations with quotation marks and providing a citation to the source of any quotation or idea used.

According to the APA Ethics Code, faculty advisers should discuss publication credit—who will be an author and the order of authors—with their student collaborators as early as possible in the research process.

The remaining standards make some less obvious but equally important points. Researchers should not publish the same data a second time as though it were new, they should share their data with other researchers, and as peer reviewers they should keep the unpublished research they review confidential. Note that the authors' names on published research—and the order in which those names appear—should reflect the importance of each person's contribution to the research. It would be unethical, for example, to include as an author someone who had made only minor contributions to the research (e.g., analyzing some of the data) or for a faculty member to make himself or herself the first author on research that was largely conducted by a student.

Pause and Reflect: *A researcher finds that the results do not support the original hypothesis. The researcher is tempted to leave out a few participants whose responses seem unusual because doing so would make the results statistically significant. Is this ethical? Why or why not? What should the researcher do instead?*

4.8 Putting Ethics Into Practice

In this section, we look at some practical advice for conducting ethical research in psychology. Again, it is important to remember that ethical issues arise well before you begin to collect data and continue to arise through publication and beyond.

4.8.1 Know and Accept Your Ethical Responsibilities

As the American Psychological Association (APA) Ethics Code notes in its introduction, "Lack of awareness or misunderstanding of an ethical standard is not itself a defense to a charge of unethical conduct." This is why the very first thing that you must do as a new researcher is know and accept your ethical responsibilities. At a minimum, this means reading and understanding the relevant standards of the APA Ethics Code, distinguishing minimal risk from at-risk research, and knowing the specific policies and procedures of your institution—including how to prepare and submit a research protocol for institutional review board (IRB) review. If you are conducting research as a course requirement, there may be specific course standards, policies, and procedures. If any standard, policy, or procedure is unclear—or you are unsure what to do about an ethical issue that arises—you must seek clarification. You can do this by reviewing the relevant ethics codes, reading about how similar issues have been resolved by others, or consulting with more experienced researchers, your IRB, or your course instructor. Ultimately, you as the researcher must take responsibility for the ethics of the research you conduct.

4.8.2 Identify and Minimize Risks

As you design your study, you must identify and minimize risks to participants. Start by listing all the risks, including risks of physical and psychological harm and violations of confidentiality. Remember that it is easy for researchers to see risks as less serious than participants do or even to overlook them completely. For example, one student researcher wanted to test people's sensitivity to violent images by showing them gruesome photographs of crime and accident scenes. Because she was an emergency medical technician, however, she greatly underestimated how disturbing these images were to most people. Remember too that some risks might apply only to some participants. For example, while most people would have no problem completing a survey about their fear of various crimes, those who have been a victim of one of those crimes might become upset. This is why you should seek input from a variety of people, including your research collaborators, more experienced researchers, and even from nonresearchers who might be better able to take the perspective of a participant.

Once you have identified the risks, you can often reduce or eliminate many of them. One way is to modify the research design. For example, you might be able to shorten or simplify the procedure to prevent boredom and frustration. You might be able to replace upsetting or offensive stimulus materials (e.g., graphic accident scene photos) with less upsetting or offensive ones (e.g., milder photos of the sort people are likely to see in the newspaper). A good example of modifying a research design is a 2009 replication of Milgram's study conducted by Jerry Burger. Instead of allowing his participants to continue administering shocks up to the 450-V maximum, the researcher always stopped the procedure when they were about to administer the 150-V shock (Burger, 2009). This made sense because in Milgram's study (a) participants' severe negative reactions occurred after this point and (b) most participants who administered

the 150-V shock continued all the way to the 450-V maximum. Thus the researcher was able to compare his results directly with Milgram's at every point up to the 150-V shock and also was able to estimate how many of his participants would have continued to the maximum—but without subjecting them to the severe stress that Milgram did. (The results, by the way, were that these contemporary participants were just as obedient as Milgram's were.)

A second way to minimize risks is to use a pre-screening procedure to identify and eliminate participants who are at high risk. You can do this in part through the informed consent process. For example, you can warn participants that a survey includes questions about their fear of crime and remind them that they are free to withdraw if they think this might upset them. Pre-screening can also involve collecting data to identify and eliminate participants. For example, Burger used an extensive pre-screening procedure involving multiple questionnaires and an interview with a clinical psychologist to identify and eliminate participants with physical or psychological problems that put them at high risk.

A third way to minimize risks is to take active steps to maintain confidentiality. You should keep signed consent forms separately from any data that you collect and in such a way that no individual's name can be linked to his or her data. In addition, beyond people's sex, age, and ethnicity, you should only collect personal information that you actually need to answer your research question. If people's sexual orientation or ethnicity is not clearly relevant to your research question, for example, then do not ask them about it. Be aware also that certain data collection procedures can lead to unintentional violations of confidentiality. When participants respond to an oral survey in a shopping mall or complete a questionnaire in a classroom setting, it is possible that their responses will be overheard or seen by others. If the responses are personal, it is better to administer the survey or questionnaire individually in private or to use other techniques to prevent the unintentional sharing of personal information.

4.8.3 Identify and Minimize Deception

Remember that deception can take a variety of forms, not all of which involve actively misleading participants. It is also deceptive to allow participants to make incorrect assumptions (e.g., about what will be on a "memory test") or simply withhold information about the full design or purpose of the study. This is called passive deception. Overall, it is best to identify and minimize all forms of deception.

Remember that according to the APA Ethics Code, deception is ethically acceptable only if there is no way to answer your research question without it. Therefore, if your research design includes any form of active deception, you should consider whether it is truly necessary. Imagine, for example, that you want to know whether the age of college professors affects students' expectations about their teaching ability. You could do this by telling participants that you will show them photos of college professors and ask them to rate each one's teaching ability. But if the photos are not really of college professors but of your own family members and friends, then this would be deception. This deception could easily be eliminated, however,

by telling participants instead to imagine that the photos are of college professors and to rate them as if they were.

In general, it is considered acceptable to wait until debriefing before you reveal your research question as long as you describe the procedure, risks, and benefits during the informed consent process. For example, you would not have to tell participants that you wanted to know whether the age of college professors affects people's expectations about them until the study was over. Not only is this information unlikely to affect people's decision about whether or not to participate in the study, but it has the potential to invalidate the results. Participants who know that age is the independent variable might rate the older and younger "professors" differently because they think you want them to. Alternatively, they might be careful to rate them the same so that they do not appear prejudiced. But even this extremely mild form of deception can be minimized by informing participants—orally, in writing, or both—that although you have accurately described the procedure, risks, and benefits, you will wait to reveal the research question until afterward. In essence, participants give their consent to be deceived or to have information withheld from them until later.

4.8.4 Weigh the Risks Against the Benefits

Once the risks of the research have been identified and minimized, you need to weigh them against the benefits. This requires identifying all the benefits. Remember to consider benefits to the participants, to science, and to society. If you are a student researcher, remember that one of the benefits is the knowledge you will gain about how to conduct scientific research in psychology—knowledge you can then use to complete your studies and succeed in graduate school or in your career.

If the research poses minimal risk—no more than in people's daily lives or routine physical or psychological examinations—then even a small benefit to participants, science, or society is generally considered enough to justify it. If it poses more than minimal risk, then there should be more benefits. If the research has the potential to upset some participants, for example, then it becomes more important that the study be well designed and answer a scientifically interesting research question or have clear practical implications. It would be unethical to subject people to pain, fear, or embarrassment for no better reason than to satisfy one's personal curiosity. In general, psychological research that has the potential to cause harm that is more than minor or lasts for more than a short time is rarely considered justified by its benefits. Consider, for example, that Milgram's study—as interesting and important as the results were—would be considered unethical by today's standards.

4.8.5 Create Informed Consent and Debriefing Procedures

Once you have settled on a research design, you need to create your informed consent and debriefing procedures. Start by deciding whether informed consent is necessary according

to APA Standard 8.05. If informed consent is necessary, there are several things you should do. First, when you recruit participants—whether it is through word of mouth, posted advertisements, or a participant pool—provide them with as much information about the study as you can. This will allow those who might find the study objectionable to avoid it. Second, prepare a script or set of "talking points" to help you explain the study to your participants in simple everyday language. This should include a description of the procedure, the risks and benefits, and their right to withdraw at any time. Third, create an informed consent form that covers all the points in Standard 8.02a that participants can read and sign after you have described the study to them. Your university, department, or course instructor may have a sample consent form that you can adapt for your own study. If not, an Internet search will turn up several samples. Remember that if appropriate, both the oral and written parts of the informed consent process should include the fact that you are keeping some information about the design or purpose of the study from them but that you will reveal it during debriefing.

Debriefing is similar to informed consent in that you cannot necessarily expect participants to read and understand written debriefing forms. So again it is best to write a script or set of talking points with the goal of being able to explain the study in simple everyday language. During debriefing, you should reveal the research question and full design of the study. For example, if participants are tested under only one condition, then you should explain what happened in the other conditions. If you deceived your participants, you should reveal this as soon as possible, apologize for the deception, explain why it was necessary, and correct any misconceptions that participants might have as a result. Debriefing is also a good time to provide additional benefits to research participants by giving them relevant practical information or referrals to other sources of help. For example, in a study of attitudes toward domestic abuse, you could provide pamphlets about domestic abuse and referral information to the university counseling center for those who might want it.

Remember to schedule plenty of time for the informed consent and debriefing processes. They cannot be effective if you have to rush through them.

4.8.6 Get Approval

The next step is to get institutional approval for your research based on the specific policies and procedures at your institution or for your course. This will generally require writing a protocol that describes the purpose of the study, the research design and procedure, the risks and benefits, the steps taken to minimize risks, and the informed consent and debriefing procedures. Do not think of the institutional approval process as merely an obstacle to overcome but as an opportunity to think through the ethics of your research and to consult with others who are likely to have more experience or different perspectives than you. If the IRB has questions or concerns about your research, address them promptly and in good faith. This might even mean making further modifications to your research design and procedure before resubmitting your protocol.

4.8.7 Follow Through

Your concern with ethics should not end when your study receives institutional approval. It now becomes important to stick to the protocol you submitted or to seek additional approval for anything other than a minor change. During the research, you should monitor your participants for unanticipated reactions and seek feedback from them during debriefing. One criticism of Milgram's study is that although he did not know ahead of time that his participants would have such severe negative reactions, he certainly knew after he had tested the first several participants and should have made adjustments at that point (Baumrind, 1985). Be alert also for potential violations of confidentiality. Keep the consent forms and the data safe and separate from each other and make sure that no one, intentionally or unintentionally, has access to any participant's personal information.

Finally, you must maintain your integrity through the publication process and beyond. Address publication credit—who will be authors on the research and the order of authors—with your collaborators early and avoid plagiarism in your writing. Remember that your scientific goal is to learn about the way the world actually is and that your scientific duty is to report on your results honestly and accurately. So do not be tempted to fabricate data or alter your results in any way. Besides, unexpected results are often as interesting, or more so, than expected ones.

Pause and Reflect: *Think back to the three historical cases discussed at the beginning of this chapter: the Nazi medical experiments, the Tuskegee Syphilis Study, and Milgram's obedience study. Which ethical principle—or combination of principles—was most clearly violated in each case?*

4.9 Media Attributions

[[1] *A Framework for Thinking about Ethical Issues in Scientific Research.* Adapted from Research Ethics by M. Ukeye, in Critical Research Methods in Psychology, by] S.D'Costa, M.Ukeye, M. O'Neil, & R. Anguiano, Critical research methods in psychology. Saint Mary's College of California. Licensed under CC BY-NC-SA 4.0.

4.10 Text Attributions

[Dudley, M. (2019). Research methods in psychology. OER Commons.] [https://oercommons.org/authoring/51456-research-methods-in-psychology]. Licensed under CC BY-NC-SA 4.0. [Modified by current authors.]

4.11 References

Baumrind, D. (1985). Research using intentional deception: Ethical issues revisited. *American Psychologist*, *40*(2), 165-174. https://doi.org/10.1037/0003-066X.40.2.165

Bowd, A.D. & Shapiro, K.J. (1993). The case against laboratory animal research in psychology. *Journal of Social Issues*, *49*(1), 133-142.

Burger, J.M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist*, *64*(1), 1 – 11. https://doi.org/10.1037/a0010932

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*, 371-378. https://psycnet.apa.org/doi/10.1037/h0040525

Miller, N.E. (1985). The value of behavioral research on animals. *American Psychologist*, *40*(4), 423-440. https://psycnet.apa.org/doi/10.1037/0003-066X.40.4.423

Reverby, S.M. (2009). Examining Tuskegee: The infamous syphilis study and its legacy. Chapel Hill, NC: University of North Carolina Press.

Part II

Module 2: Getting Good Data - Measurement and Sampling

5 Chapter 5: 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. We also look at some practical issues in psychological measurement.

Do You Feel You Are a Person of Worth?

The Rosenberg Self-Esteem Scale (Rosenberg, 1989) is one of the most common measures of self-esteem and the one that MacDonald and Martineau used in their study. Participants respond to each of the 10 items that follow with a rating on a 4-point scale: Strongly Agree, Agree, Disagree, Strongly Disagree. Score Items 1, 2, 4, 6, and 7 by assigning 3 points for each Strongly Agree response, 2 for each Agree, 1 for each Disagree, and 0 for each Strongly Disagree. Reverse the scoring for Items 3, 5, 8, 9, and 10 by assigning 0 points for each Strongly Agree, 1 point for each Agree, and so on. The overall score is the total number of points.

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.

Exercise 5.1
Practice. Complete the Rosenberg Self-Esteem Scale and compute your overall score.

5.1 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 by inserting a meat thermometer. It is also consistent with measurement throughout the 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 ($9.8\ \mathrm{m/s^2}$). The result of this procedure is a score that represents the object's potential energy.

Of course 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—his or her 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, where she reads a list of two digits to the person and asks him or her 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 his or her 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 he or she has 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 his or her current level of depression.

The important point here is that measurement does not require any particular instruments or procedures. It does not require placing individuals or objects on bathroom scales, holding rulers up to them, or inserting thermometers into them. What it *does* require is *some* systematic procedure for assigning scores to individuals or objects so that those scores represent the characteristic of interest.

5.2 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., extroversion), emotional states (e.g., fear), attitudes (e.g., toward taxes), and abilities (e.g., athleticism).

Psychological constructs often 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 college student is highly extroverted (see below) 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 (talking, laughing, etc.) across a variety of situations. Another reason psychological constructs cannot always 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.

5.2.1 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). Table 5.1 describes the Big Five dimensions and facets.

Table 5.1: **Table 5.1** The Big 5 Personality Dimensions[1]

Big Five Dimension	Facets					
Openness to Experience	Fantasy	Aesthetics	Feelings	Actions	Ideas	Values
Conscientiousness	Competence	Order	Dutifulness	Achievement Striving	Self-Discipline	Deliberation
Extraversion	Warmth	Gregariousness	Assertiveness	Activity	Excitement Seeking	Positive Emotions
Agreeableness	Trust	Straightforwardness	Altruism	Compliance	Modesty	Tender-Mindedness
Neuroticism	Worry	Anger	Discouragement	Self-Consciousness	Impulsivity	Vulnerability

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) 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 has a strong genetic component, remains fairly stable over time, and is positively correlated with 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 importantly, researchers develop definitions that are more detailed, more precise, and more accurate than the informal definitions found in a 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 works better. In others, researchers are still in the process of deciding which of various conceptual definitions is the best.

Exercise 5.2
Practice. Write your own conceptual definition of self-confidence, irritability, and athleticism.

5.3 Operational Definitions

An ***operational definition*** is a definition of a variable in terms of precisely how it is to be measured (or manipulated depending on the type of variable you are utilizing and goals of your study; see Module 5 for more information on manipulated variables). 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. ***Behavioral measures*** are those in which some 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 et al., 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 he or she 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 in 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 (an indicator of sweating and physiological arousal), 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 Daily Hassles and Uplifts Scale (Delongis et al., 1982) is similar but focuses on everyday stressors like misplacing things and being concerned about one's weight. The Perceived Stress Scale (Cohen et al., 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" 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 ultimately allows researchers 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."

Exercise 5.3
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.

5.4 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 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.

5.4.1 Nominal Scales

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.

5.4.2 Ordinal Scales

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.)

5.4.3 Interval Scales

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.

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.

5.4.4 Ratio Scales

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 5.2).

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 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 the number of siblings is measured at the ratio level.

Table 5.2: **Table 5.2** Summary of Levels of Measurement[2]

Level of Measurement	Levels/values	Rank order	Equal intervals	True zero
NOMINAL	X			
ORDINAL	X	X		
INTERVAL	X	X	X	
RATIO	X	X	X	X

Exercise 5.4
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.

Please watch: Variable Measurement Scales (Statistics Lectures)

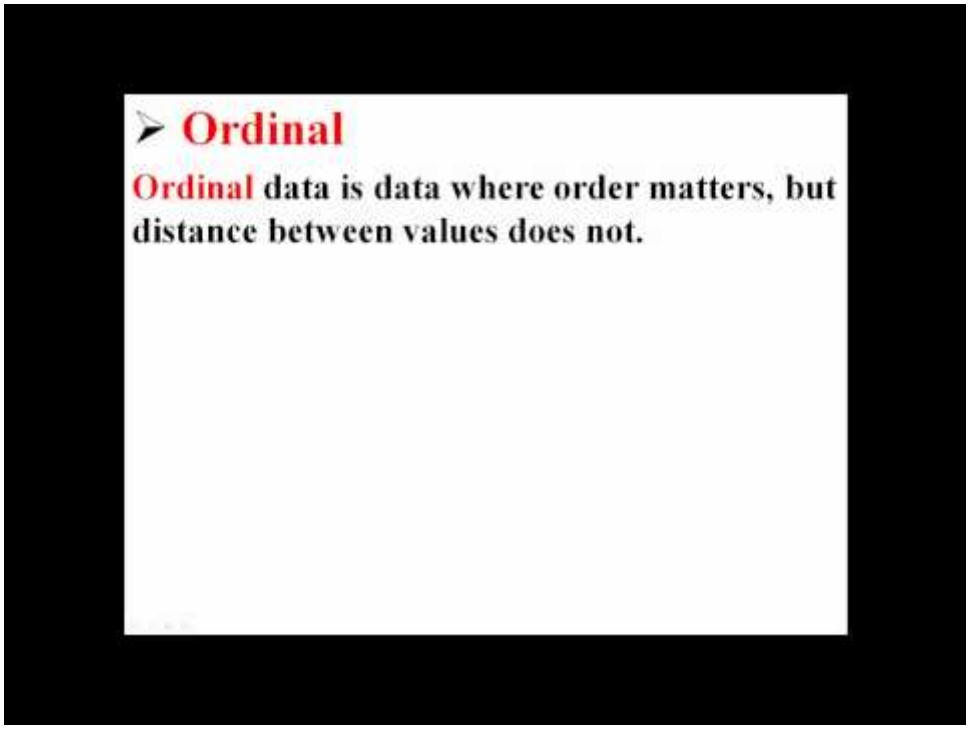


Figure 5.1: Variable Measurement Scales

Please watch: What Are Scales of Measurement? (Nominal, Ordinal, Interval, Ratio) (Daniel Storage)

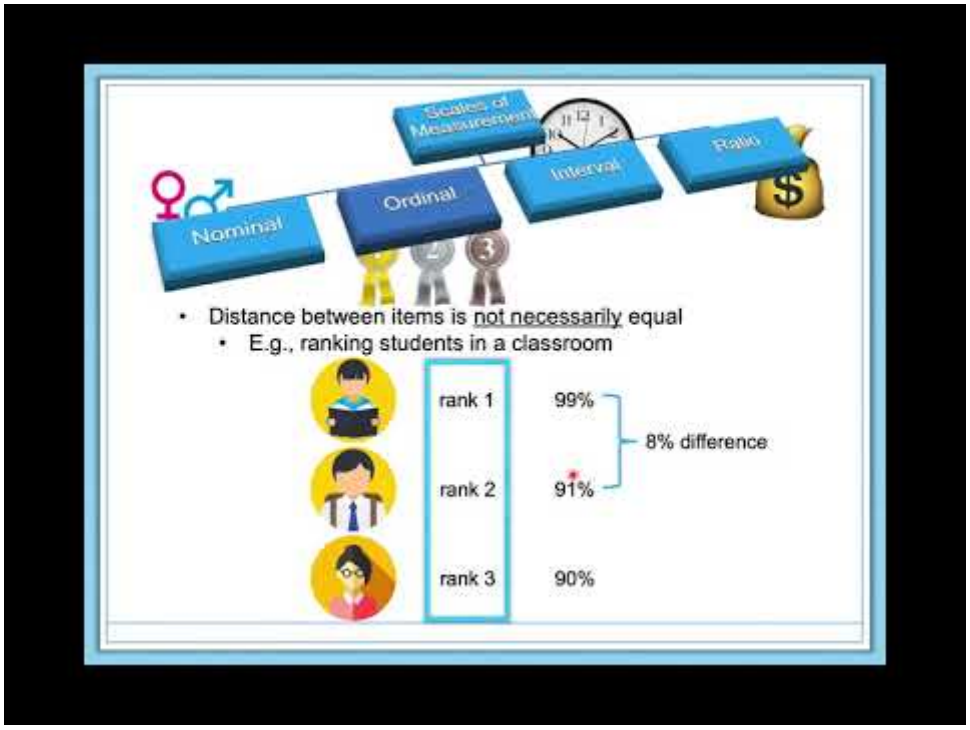


Figure 5.2: What Are Scales of Measurement? (Nominal, Ordinal, Interval, Ratio)

5.5 Reliability

Measurement involves assigning scores to individuals so that they represent some characteristic of the individuals. But how do researchers know that the scores actually represent the characteristic, especially when it is a construct like intelligence, self-esteem, depression, or working memory capacity? The answer is that they conduct research using the measure to confirm that the scores make sense based on their understanding of the construct being measured. This is an extremely important point. Psychologists do not simply assume that their measures work. Instead, they collect data to demonstrate that they work. If their research does not demonstrate that a measure works, they stop using it.

As an informal example, imagine that you have been dieting for a month. Your clothes seem to be fitting more loosely, and several friends have asked if you have lost weight. If at this point your bathroom scale indicated that you had lost 10 pounds, this would make sense and you would continue to use the scale. But if it indicated that you had gained 10 pounds, you would rightly conclude that it was broken and either fix it or get rid of it. In evaluating a measurement method, psychologists consider two general dimensions that fall under ***construct validity***: reliability and validity.

Reliability refers to the consistency of a measure. Psychologists consider three types of consistency: over time (test-retest reliability), across items (internal consistency), and across different researchers (inter-rater reliability).

5.5.1 Test-Retest Reliability

When researchers measure a construct that they assume to be consistent across time, then the scores they obtain should also be consistent across time. ***Test-retest reliability*** is the extent to which this is actually the case. For example, intelligence is generally thought to be consistent across time. A person who is highly intelligent today will be highly intelligent next week. This means that any good measure of intelligence should produce roughly the same scores for this individual next week as it does today. Clearly, a measure that produces highly inconsistent scores over time cannot be a very good measure of a construct that is supposed to be consistent.

Assessing test-retest reliability requires using the measure on a group of people at one time, using it again on the same group of people at a later time, and then looking at the test-retest correlation between the two sets of scores. In general, a test-retest correlation of +.80 or greater is considered to indicate good reliability.

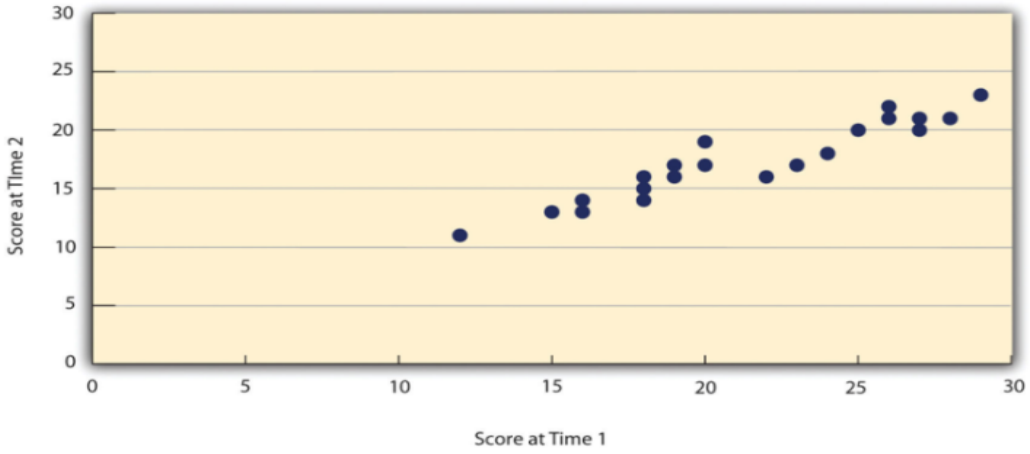


Figure 15.1. Test-Retest Correlation Between Two Sets of Scores of Several College Students on the Rosenberg Self-Esteem Scale, Given Two Times a Week Apart [Image Description]

Figure 5.3: **Figure 5.2** Test-Retest Correlation Between Two Sets of Scores of Several College Students on the Rosenberg Self-Esteem Scale, Given Two Times a Week Apart[3]

Again, high test-retest correlations make sense when the construct being measured is assumed to be consistent over time, which is the case for intelligence, self-esteem, and the Big Five personality dimensions. But other constructs are not assumed to be stable over time. The very

nature of mood, for example, is that it changes. So a measure of mood that produced a low test-retest correlation over a period of a month would not be a cause for concern.

5.5.2 Internal Consistency

A second kind of reliability is ***internal consistency***, which is the consistency of people's responses across the items on a multiple-item measure. In general, all the items on such measures are supposed to reflect the same underlying construct, so people's scores on those items should be correlated with each other. On the Rosenberg Self-Esteem Scale, people who agree that they are a person of worth should tend to agree that they have a number of good qualities. If people's responses to the different items are not correlated with each other, then it would no longer make sense to claim that they are all measuring the same underlying construct. This is as true for behavioral and physiological measures as for self-report measures. For example, people might make a series of bets in a simulated game of roulette as a measure of their level of risk seeking. This measure would be internally consistent to the extent that individual participants' bets were consistently high or low across trials.

Like test-retest reliability, internal consistency can only be assessed by collecting and analyzing data. One approach is to look at a ***split-half correlation***. This involves splitting the items into two sets, such as the first and second halves of the items or the even- and odd-numbered items. Then a score is computed for each set of items, and the relationship between the two sets of scores is examined. A split-half correlation of +.80 or greater is generally considered good internal consistency.

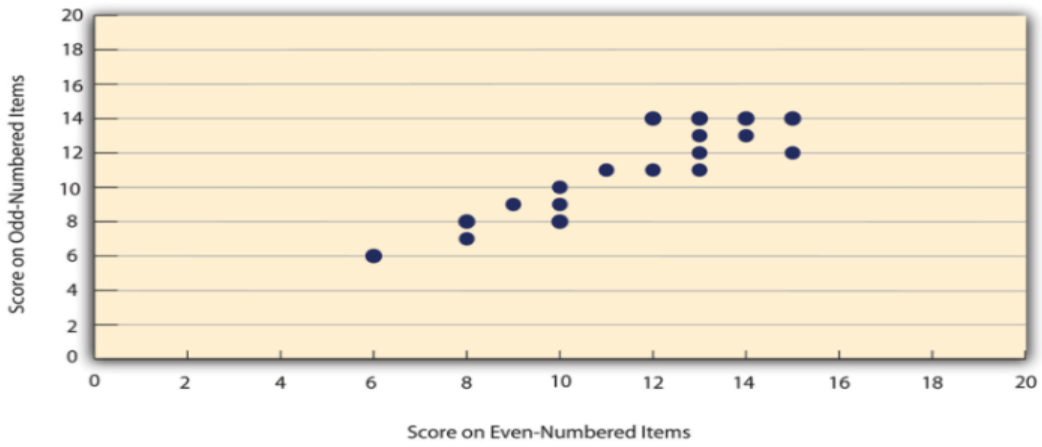


Figure 15.2. Split-Half Correlation Between Several College Students' Scores on the Even-Numbered Items and Their Scores on the Odd-Numbered Items of the Rosenberg Self-Esteem Scale. [Image Description]

Figure 5.4: **Figure 5.2** Split-Half Correlation Between Several College Students' Scores on the Even-Numbered Items and Their Scores on the Odd-Numbered Items of the Rosenberg Self-Esteem Scale[4]

Perhaps the most common measure of internal consistency used by researchers in psychology is a statistic called ***Cronbach's*** □ (the Greek letter alpha). Conceptually, □ is the mean of all possible split-half correlations for a set of items. For example, there are 252 ways to split a set of 10 items into two sets of five. Cronbach's □ would be the mean of the 252 split-half correlations. Note that this is not how □ is actually computed, but it is a correct way of interpreting the meaning of this statistic. Again, a value of +.80 or greater is generally taken to indicate good internal consistency.

Please watch: Cronbach's alpha (Simply Explained)

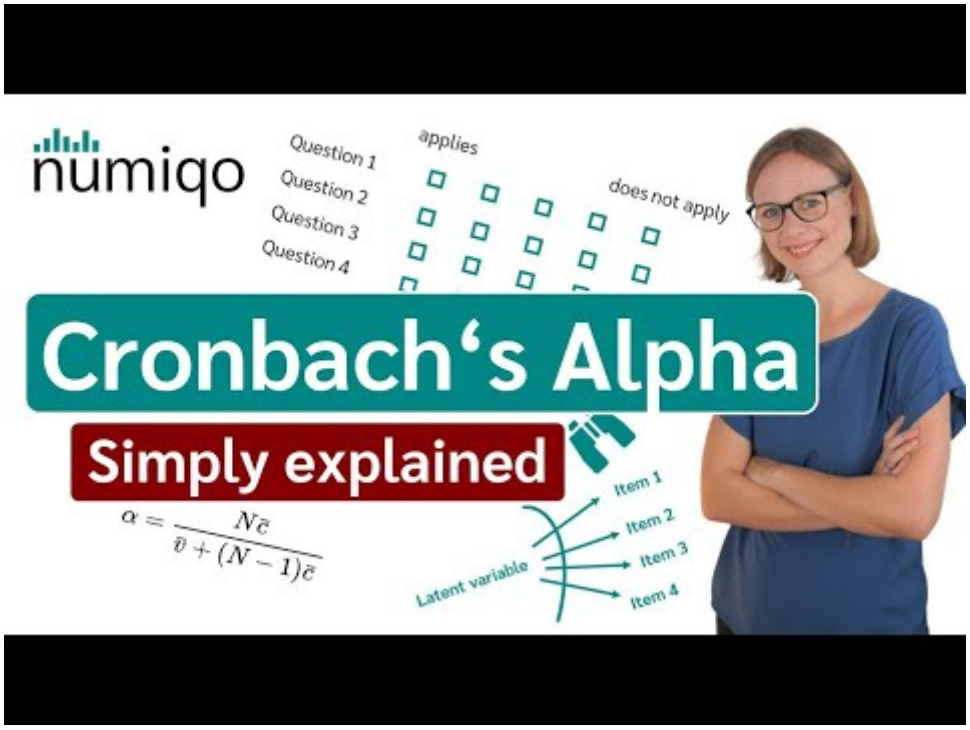


Figure 5.5: Cronbach's Alpha (Simply explained)

5.5.3 Inter-rater Reliability

Many behavioral measures involve significant judgment on the part of an observer or a rater. ***Inter-rater reliability*** is the extent to which different observers are consistent in their judgments. For example, if you were interested in measuring college students' social skills, you could make video recordings of them as they interacted with another student whom they are meeting for the first time. Then you could have two or more observers watch the videos and rate each student's level of social skills. To the extent that each participant does in fact have some level of social skills that can be detected by an attentive observer, different observers' ratings should be highly correlated with each other. If they were not, then those ratings could not be an accurate representation of participants' social skills. Inter-rater reliability is often assessed using Cronbach's 𝛼 when the judgments are quantitative or an analogous statistic called Cohen's 𝜅 (the Greek letter kappa) when they are categorical.

Exercise 5.5
Practice. Ask several friends to complete the Rosenberg Self-Esteem Scale. Then assess its internal consistency by making a scatterplot to show the split-half correlation (even- vs. odd-numbered items). Compute the correlation coefficient too if you know how.

Please watch: Forms of Reliability in Research and Statistics (Daniel Storage)

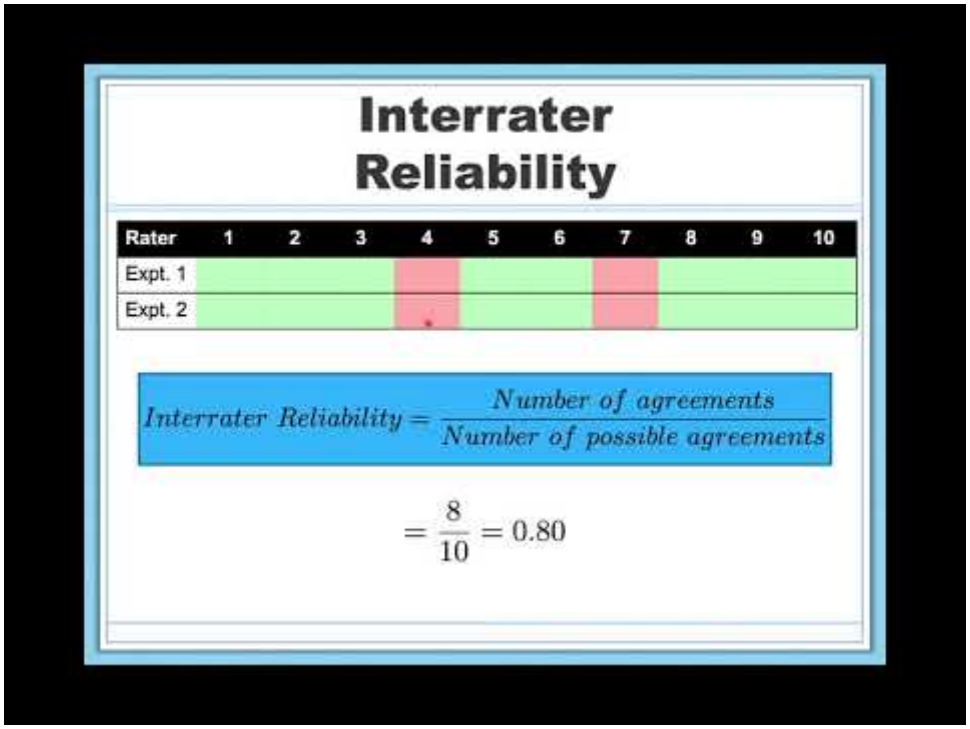


Figure 5.6: Forms of Reliability in Research and Statistics

5.6 Validity

Validity is the extent to which the scores from a measure represent the variable they are intended to. But how do researchers make this judgment? We have already considered one factor that they take into account—reliability. When a measure has good test-retest reliability and internal consistency, researchers should be more confident that the scores represent what they are supposed to. There has to be more to it, however, because a measure can be extremely reliable but have no validity whatsoever. As an absurd example, imagine someone who believes that people's index finger length reflects their self-esteem and therefore tries to measure self-esteem by holding a ruler up to people's index fingers. Although this measure would have extremely good test-retest reliability, it would have absolutely no validity. The fact that one person's index finger is a centimeter longer than another's would indicate nothing about which one had higher self-esteem.

Textbook presentations of validity usually divide it into several distinct "types." But a good way to interpret these types is that they are other kinds of evidence—in addition to reliability—that should be taken into account when judging the validity of a measure.

5.6.1 Face Validity

Face validity is the extent to which a measurement method appears "on its face" to measure the construct of interest. Most people would expect a self-esteem questionnaire to include items about whether they see themselves as a person of worth and whether they think they have good qualities. So a questionnaire that included these kinds of items would have good face validity. The finger-length method of measuring self-esteem, on the other hand, seems to have nothing to do with self-esteem and therefore has poor face validity. Although face validity can be assessed quantitatively—for example, by having a large sample of people rate a measure in terms of whether it appears to measure what it is intended to—it is usually assessed informally.

Face validity is at best a very weak kind of evidence that a measurement method is measuring what it is supposed to. One reason is that it is based on people's intuitions about human behavior, which are frequently wrong. It is also the case that many established measures in psychology work quite well despite lacking face validity. The Minnesota Multiphasic Personality Inventory (MMPI) measures many personality characteristics and disorders by having people decide whether each of over 567 different statements applies to them—where many of the statements do not have any obvious relation to the construct that they measure.

5.6.2 Content Validity

Content validity is the extent to which a measure "covers" the construct of interest. For example, if a researcher conceptually defines test anxiety as involving both sympathetic nervous system activation (leading to nervous feelings) and negative thoughts, then his measure of test anxiety should include items about both nervous feelings and negative thoughts. Or consider that attitudes are usually defined as involving thoughts, feelings, and actions toward something. By this conceptual definition, a person has a positive attitude toward exercise to the extent that he or she thinks positive thoughts about exercising, feels good about exercising, and actually exercises. So to have good content validity, a measure of people's attitudes toward exercise would have to reflect all three of these aspects. Like face validity, content validity is not usually assessed quantitatively. Instead, it is assessed by carefully checking the measurement method against the conceptual definition of the construct.

While face and content validity are good first steps at assessing if a measure is accurately assessing the construct of interest, they are nonetheless subjective in nature and typically considered non-empirical forms for assessing validity. The next three kinds of validity described below are considered empirical or more objective assessments of validity, and therefore tend to be more desirable for researchers.

5.6.3 Criterion Validity – Concurrent, Predictive, and Convergent

Criterion validity is the extent to which people's scores on a measure are correlated with other variables (known as criteria) that one would expect them to be correlated with. For example, people's scores on a new measure of test anxiety should be negatively correlated with their performance on an important school exam. If it were found that people's scores were in fact negatively correlated with their exam performance, then this would be a piece of evidence that these scores really represent people's test anxiety. But if it were found that people scored equally well on the exam regardless of their test anxiety scores, then this would cast doubt on the validity of the measure.

A ***criterion*** can be any variable that one has reason to think should be correlated with the construct being measured, and there will usually be many of them. For example, one would expect test anxiety scores to be negatively correlated with exam performance and course grades and positively correlated with general anxiety and with blood pressure during an exam. Or imagine that a researcher develops a new measure of physical risk taking. People's scores on this measure should be correlated with their participation in "extreme" activities such as snowboarding and rock climbing, the number of speeding tickets they have received, and even the number of broken bones they have had over the years. When the criterion is measured at the same time as the construct, criterion validity is referred to as ***concurrent validity***; however, when the criterion is measured at some point in the future (after the construct has been measured), it is referred to as ***predictive validity*** (because scores on the measure have "predicted" a future outcome).

Criteria can also include other measures of the same construct. For example, one would expect new measures of test anxiety or physical risk taking to be positively correlated with existing measures of the same constructs. This is known as ***convergent validity***.

Assessing criterion validity requires collecting data using the measure. Researchers John Cacioppo and Richard Petty did this when they created their self-report Need for Cognition Scale to measure how much people value and engage in thinking (Cacioppo & Petty, 1982). In a series of studies, they showed that college faculty scored higher than assembly-line workers, that people's scores were positively correlated with their scores on a standardized academic achievement test, and that their scores were negatively correlated with their scores on a measure of dogmatism (which represents a tendency toward obedience). In the years since it was created, the Need for Cognition Scale has been used in literally hundreds of studies and has been shown to be correlated with a wide variety of other variables, including the effectiveness of an advertisement, interest in politics, and juror decisions (Petty et al., 2009).

Exercise 5.6
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?

5.6.4 Discriminant Validity

Discriminant validity (also referred to as ***divergent validity***) is the extent to which scores on a measure are *not* correlated with measures of variables that are conceptually distinct. For example, self-esteem is a general attitude toward the self that is fairly stable over time. It is not the same as mood, which is how good or bad one happens to be feeling right now. So people's scores on a new measure of self-esteem should not be very highly correlated with their moods. If the new measure of self-esteem were highly correlated with a measure of mood, it could be argued that the new measure is not really measuring self-esteem; it is measuring mood instead.

When they created the Need for Cognition Scale, Cacioppo and Petty also provided evidence of discriminant validity by showing that people's scores were not correlated with certain other variables. For example, they found only a weak correlation between people's need for cognition and a measure of their cognitive style—the extent to which they tend to think analytically by breaking ideas into smaller parts or holistically in terms of "the big picture." They also found no correlation between people's need for cognition and measures of their test anxiety and their tendency to respond in socially desirable ways. All these low correlations provide evidence that the measure is reflecting a conceptually distinct construct.

Please watch: Forms of Validity in Research and Statistics (Daniel Storage)

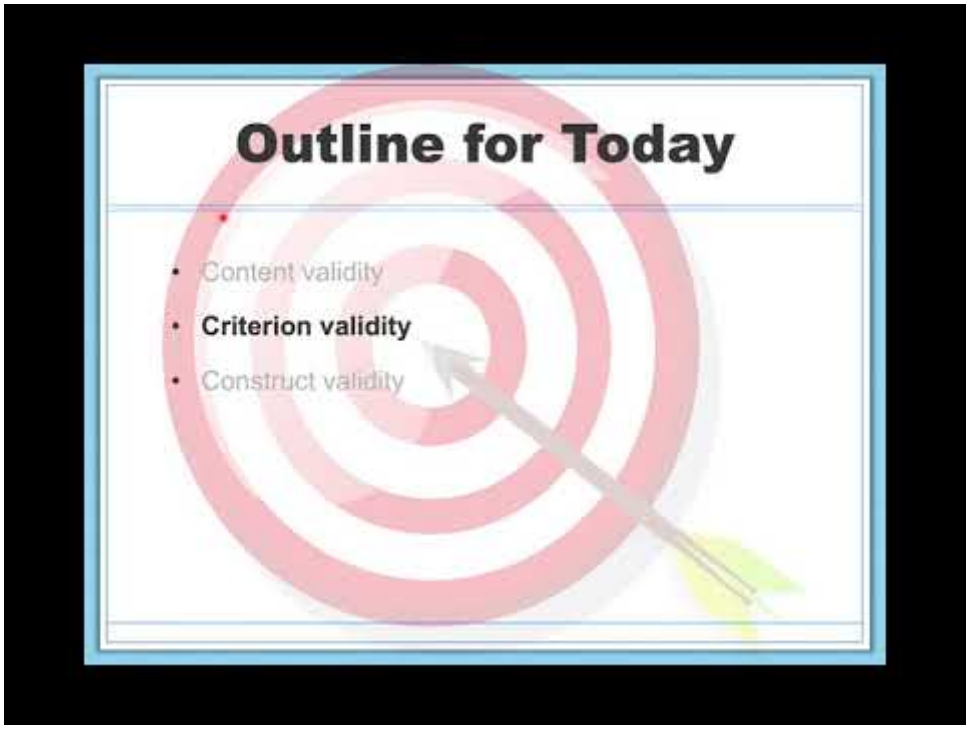


Figure 5.7: Forms of Validity in Research and Statistics

5.7 Illustrating Reliability and Validity of a Measurement

Imagine that a construct we want to measure is the bullseye (the center) of a target, as illustrated by the bullseye in each of the diagrams below. Then imagine that we have a measure, such as a survey, a questionnaire, a laboratory task, a brain imaging device, or any of a myriad types of measurements we use in psychological science.

Now imagine that each time we use our measurement is like shooting an arrow toward the target. Because our goal is to measure each construct reliably and validly, we want our arrow shots to be as consistent (reliable) and accurate (valid) as possible. We want them to consistently (reliably) and accurately (validly) hit the target's bullseye.

In the diagrams below (Figure 5.3), our arrow shots are illustrated with back dots. The larger the dots, the more of our arrows — our measurements — landed on that spot. Each diagram illustrates a relation between reliability and validity.

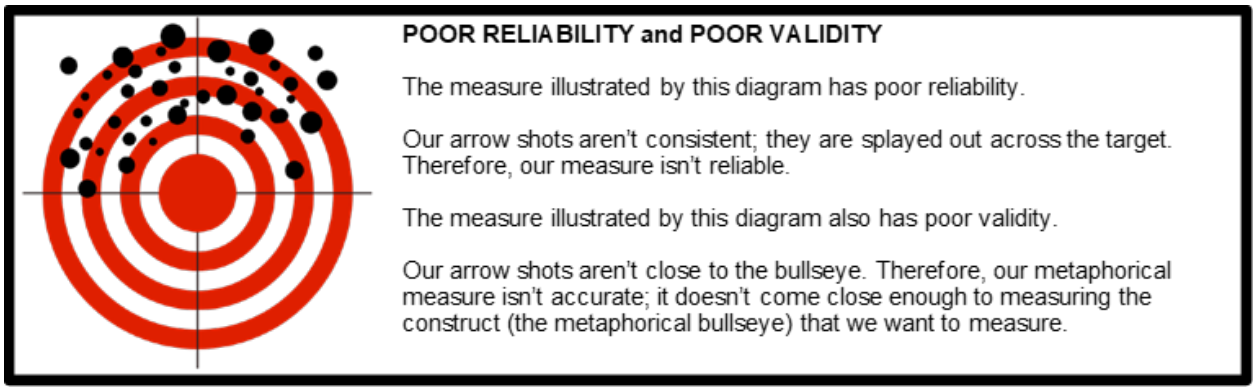


Figure 5.8: Poor reliability and poor validity. The measure illustrated by this diagram has poor reliability. Our arrow shots aren't consistent; they are splayed out across the target. Therefore, the measure isn't reliable. The measure illustrated by this diagram also has poor validity. Our arrow shots aren't close to the bullseye. Therefore, our metaphorical measure isn't accurate; it doesn't come close enough to measuring the construct (the metaphorical bullseye) that we want to measure.

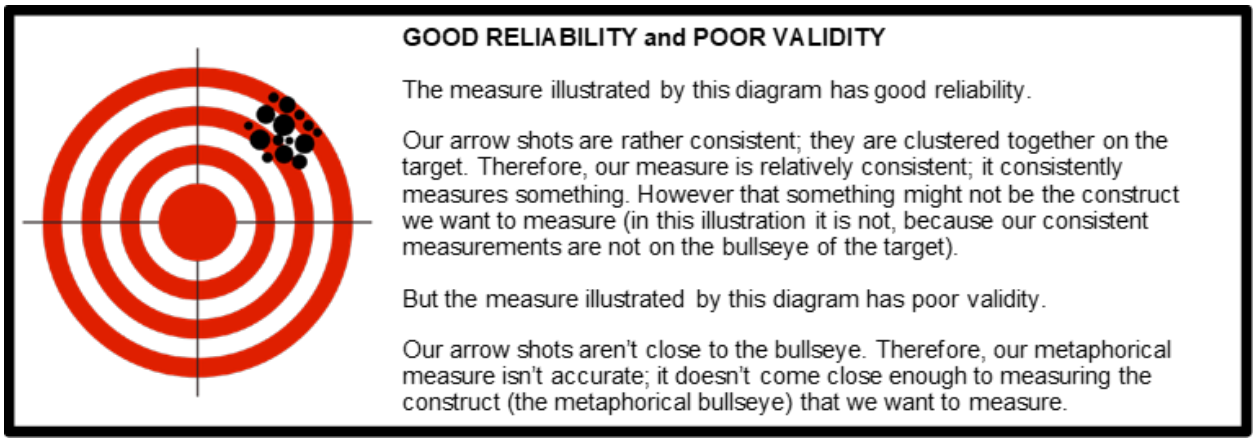


Figure 5.9: Good reliability and poor validity. The measure illustrated by this diagram has good reliability. Our arrow shots are rather consistent; they are clustered together on the target. Therefore, our measure is relatively consistent; it consistently measures something. However, that something might not be the construct we want to measure (in this illustration it is not because our consistent measurements are not on the bullseye of the target). But the measure illustrated by this diagram has poor validity. Our arrow shots aren't close to the bullseye. Therefore, our metaphorical measure isn't accurate; it doesn't come close enough to measuring the construct (the metaphorical bullseye) that we want to measure.

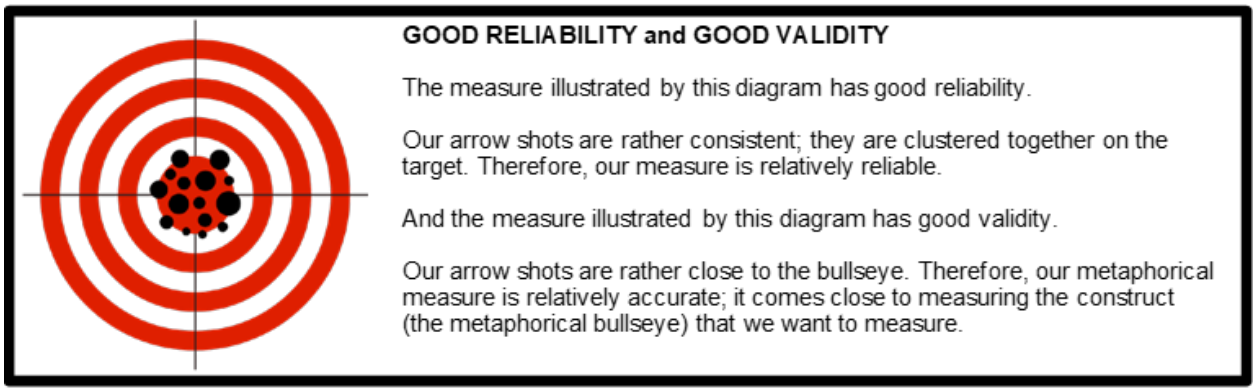


Figure 5.10: **Figure 5.3** Summarizing the Connection Between Reliability and Validity in Psychological Measurement[5]

5.8 Practical Strategies for Psychological Measurement

So far in this chapter, we have considered several basic ideas about the nature of psychological constructs and their measurement. But now imagine that you are in the position of actually having to measure a psychological construct for a research project. How should you proceed? Broadly speaking, there are four steps in the measurement process: (a) conceptually defining the construct, (b) operationally defining the construct, (c) implementing the measure, and (d) evaluating the measure. In this section, we will look at each of these steps in turn. It is important to note that creation of a psychological measure can be a long process and in some cases can even serve as its own research project when validating it on a national level.

5.8.1 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.

5.8.2 Operationally Defining the Construct

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 et al., 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.

5.8.2.1 Using an Existing Measure

It is usually a good idea to use an existing measure that has been used successfully in previous research. Among the advantages are that (a) you save the time and trouble of creating your own, (b) there is already some evidence that the measure is valid (if it has been used successfully), and (c) your results can more easily be compared with and combined with previous results. In fact, if there already exists a reliable and valid measure of a construct, other researchers might expect you to use it unless you have a good and clearly stated reason for not doing so.

If you choose to use an existing measure, you may still have to choose among several alternatives. You might choose the most common one, the one with the best evidence of reliability and validity, the one that best measures a particular aspect of a construct that you are interested in (e.g., a physiological measure of stress if you are most interested in its underlying physiology), or even the one that would be easiest to use. For example, the Ten-Item Personality Inventory (TIPI) is a self-report questionnaire that measures all the Big Five personality dimensions with just 10 items (Gosling et al., 2003). It is not as reliable or valid as longer and more comprehensive measures, but a researcher might choose to use it when testing time is severely limited.

When an existing measure was created primarily for use in scientific research, it is usually described in detail in a published research article and is free to use in your own research—with a proper citation. You might find that later researchers who use the same measure describe it only briefly but provide a reference to the original article, in which case, you would have to get the details from the original article. The American Psychological Association also publishes the Directory of Unpublished Experimental Measures and PsycTESTS, which are extensive catalogs/collections of measures that have been used in previous research. Many existing measures—especially those that have applications in clinical psychology—are proprietary. This means that a publisher owns the rights to them and that you would have to purchase them. These include many standard intelligence tests, the Beck Depression Inventory, and the

Minnesota Multiphasic Personality Inventory (MMPI). Details about many of these measures and how to obtain them can be found in other reference books, including Tests in Print and the Mental Measurements Yearbook. There is a good chance you can find these reference books in your university library.

Exercise 5.7
Practice. Choose a construct (sexual jealousy, self-confidence, etc.) and find two measures of that construct in the research literature. If you were conducting your own study, which one (if either) would you use and why? Be sure to look at any reliability or validity evidence for the measures in the research literature.

5.8.2.2 Creating Your Own Measure

Instead of using an existing measure, you might want to create your own. Perhaps there is no existing measure of the construct you are interested in, or existing ones are too difficult or time-consuming to use. Or perhaps you want to use a new measure specifically to see whether it works in the same way as existing measures—that is, to evaluate convergent validity. In this section, we consider some general issues in creating new measures that apply equally to self-report, behavioral, and physiological measures.

First, be aware that most new measures in psychology are really variations of existing measures, so you should still look to the research literature for ideas. Perhaps you can modify an existing questionnaire, create a paper-and-pencil version of a measure that is normally computerized (or vice versa), or adapt a measure that has traditionally been used for another purpose. For example, the famous Stroop task (Stroop, 1935)—in which people quickly name the colors that various color words are printed in—has been adapted for the study of social anxiety. People high in social anxiety are slower at color naming when the words have negative social connotations such as "stupid" (Amir et al., 2002).

When you create a new measure, you should strive for simplicity. Remember that your participants are not as interested in your research as you are and that they will vary widely in their ability to understand and carry out whatever task you give them. You should create a set of clear instructions using simple language that you can present in writing or read aloud (or both). It is also a good idea to include one or more practice items so that participants can become familiar with the task and to build in an opportunity for them to ask questions before continuing. It is also best to keep the measure brief to avoid boring or frustrating your participants to the point that their responses start to become less reliable and valid.

The need for brevity, however, needs to be weighed against the fact that it is nearly always better for a measure to include multiple items rather than a single item. There are two reasons for this. One is a matter of *content validity*. Multiple items are often required to cover a construct adequately. The other is a matter of *reliability*. People's responses to single items can be influenced by all sorts of irrelevant factors—misunderstanding the particular item, a

momentary distraction, or a simple error such as checking the wrong response option. But when several responses are summed or averaged, the effects of these irrelevant factors tend to cancel each other out to produce more reliable scores. Remember, however, that multiple items must be structured in a way that allows them to be combined into a single overall score by summing or averaging. To measure "financial responsibility," a student might ask people about their annual income, obtain their credit score, and have them rate how "thrifty" they are—but there is no obvious way to combine these responses into an overall score. To create a true multiple-item measure, the student might instead ask people to rate the degree to which 10 statements about financial responsibility describe them on the same five-point scale.

Finally, the very best way to assure yourself that your measure has clear instructions, includes sufficient practice, and is an appropriate length is to test several people. Observe them as they complete the task, time them, and ask them afterward to comment on how easy or difficult it was, whether the instructions were clear, and anything else you might be wondering about. Obviously, it is better to discover problems with a measure before beginning any large-scale data collection.

5.8.3 Implementing the Measure

You will want to implement any measure in a way that maximizes its reliability and validity. In most cases, it is best to test everyone under similar conditions that, ideally, are quiet and free of distractions. Participants are often tested in groups because it is efficient, but be aware that it can create distractions that reduce the reliability and validity of the measure. As always, it is good to use previous research as a guide. If others have successfully tested people in groups using a particular measure, then you should consider doing it too.

Be aware also that people can react in a variety of ways to being measured, which reduces the reliability and validity of the scores. Although some disagreeable participants might intentionally respond in ways meant to disrupt a study, ***participant reactivity*** is more likely to take the opposite form. Agreeable participants might respond in ways they believe they are expected to. Some participants might engage in *socially desirable* responding, doing or saying things because they think it is the socially appropriate thing. For example, people with low self-esteem agree that they feel they are a person of worth, not because they really feel this way, but because they believe this is the socially appropriate response and do not want to look bad in the eyes of the researcher. Additionally, research studies can have built-in ***demand characteristics***: subtle cues that reveal how the researcher expects participants to behave. For example, a participant whose attitude toward exercise is measured immediately after she is asked to read a passage about the dangers of heart disease might reasonably conclude that the passage was meant to improve her attitude. As a result, she might respond more favorably because she believes she is expected to by the researcher. Finally, your own expectations (as a researcher) can bias participants' behaviors in unintended ways.

There are several precautions you can take to minimize these kinds of reactivity. One is to make the procedure as clear and brief as possible so that participants are not tempted to vent their frustrations on your results. Another is to guarantee participants' anonymity and make clear to them that you are doing so. If you are testing them in groups, be sure that they are seated far enough apart that they cannot see each other's responses. Give them all the same type of writing tool so that they cannot be identified by, for example, the pink glitter pen that they used. You can even allow them to seal completed questionnaires into individual envelopes or put them into a drop box where they immediately become mixed with others' questionnaires. Although informed consent requires telling participants what they will be doing, it does not require revealing your hypothesis or other information that might suggest to participants how you expect them to respond. A questionnaire designed to measure financial responsibility need not be titled "Are You Financially Responsible?" It could be titled "Money Questionnaire" or have no title at all. Finally, the effects of your expectations can be minimized by arranging to have the measure administered by a helper who is "blind" or unaware of its intent or of any hypothesis being tested. Regardless of whether this is possible, you should standardize all interactions between researchers and participants—for example, by always reading the same set of instructions word for word.

5.8.4 Evaluating the Measure

Once you have used your measure on a sample of people and have a set of scores, you are in a position to evaluate it more thoroughly in terms of reliability and validity. Even if the measure has been used extensively by other researchers and has already shown evidence of reliability and validity, you should not assume that it worked as expected for your particular sample and under your particular testing conditions. Regardless, you now have additional evidence bearing on the reliability and validity of the measure, and it would make sense to add that evidence to the research literature.

In most research designs, it is not possible to assess test-retest reliability because participants are tested at only one time. For a new measure, you might design a study specifically to assess its test-retest reliability by testing the same set of participants at two separate times. In other cases, a study designed to answer a different question still allows for the assessment of test-retest reliability. For example, a psychology instructor might measure his students' attitude toward critical thinking using the same measure at the beginning and end of the semester to see if there is any change. Even if there is no change, he could still look at the correlation between students' scores at the two times to assess the measure's test-retest reliability. It is also customary to assess internal consistency for any multiple-item measure—usually by looking at a split-half correlation or Cronbach's .

Criterion validity can be assessed in various ways. For example, if your study included more than one measure of the same construct or measures of conceptually distinct constructs, then you should look at the correlations among these measures to be sure that they fit your expectations. Note also that a successful experimental manipulation also provides evidence of

criterion validity. Recall from the beginning of this chapter that MacDonald and Martineau manipulated participants' moods by having them think either positive or negative thoughts, and after the manipulation, their mood measure showed a distinct difference between the two groups. This simultaneously provided evidence that their mood manipulation worked and that their mood measure was valid.

But what if your newly collected data casts doubt on the reliability or validity of your measure? The short answer is that you have to ask why. It could be that there is something wrong with your measure or how you administered it. It could be that there is something wrong with your conceptual definition. It could be that your experimental manipulation failed. For example, if a mood measure showed no difference between people who you instructed to think positive versus negative thoughts, maybe it is because the participants did not actually think the thoughts they were supposed to, or that the thoughts did not actually affect their moods. In short, it is "back to the drawing board" to revise the measure, revise the conceptual definition, or try a new manipulation.

5.9 Summary

- 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 (inter-rater reliability). Validity is the extent to which the scores actually represent the variable they are intended to.

- Validity is a judgment based on various types of evidence. The relevant evidence includes the measure's reliability, whether it covers the construct of interest, and whether the scores it produces are correlated with other variables they are expected to be correlated with and not correlated with variables that are conceptually distinct.
- 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.
- You often have the option of using an existing measure or creating a new measure. You should make this decision based on the availability of existing measures and their adequacy for your purposes.
- Several simple steps can be taken in creating new measures and in implementing both existing and new measures that can help maximize reliability and validity.
- Once you have used a measure, you should reevaluate its reliability and validity based on your new data. Remember that the assessment of reliability and validity is an ongoing process.

5.10 Media Attributions

[1-4] Jhangiani, R.S., Chiang, I-C.A., Cuttler, C., & Leighton, D.C. (2019). Research methods in psychology. 4th edition. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

[5] Gernsbacher, M.A. (n.d.). Open Access Active-Learning Research Methods Course. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

5.11 Text Attributions

D'Costa, S., Ukeye, M., O'Neil, M., & Anguiano, R. (no date). Critical Research Methods in Psychology. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Dudley, M. (2019). Research Methods in Psychology. OER Commons. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Gernsbacher, M.A. (n.d.). Open Access Active-Learning Research Methods Course. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Jhangiani, R.S., Chiang, I-C.A., Cuttler, C., & Leighton, D.C. (2019). Research methods in psychology. 4th edition. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

5.12 References

Amir, N., Freshman, M., & Foa, E. (2002). Enhanced Stroop interference for threat in social phobia. *Journal of Anxiety Disorders*, 16, 1–9.

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology*, 63, 575–582.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42, 116–131.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, 24, 386-396.

Costa, P. T., Jr., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, 4, 5–13.

Delongis, A., Coyne, J. C., Dakof, G., Folkman, S., & Lazarus, R. S. (1982). Relationships of daily hassles, uplifts, and major life events to health status. *Health Psychology*, 1, 119-136.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality*, 37, 504–528.

Holmes, T. H., & Rahe, R. H. (1967). The Social Readjustment Rating Scale. *Journal of Psychosomatic Research*, 11, 213-218.

MacDonald, T. K., & Martineau, A. M. (2002). Self-esteem, mood, and intentions to use condoms: When does low self-esteem lead to risky health behaviors? *Journal of Experimental Social Psychology*, 38, 299–306.

Petty, R. E, Briñol, P., Loersch, C., & McCaslin, M. J. (2009). The need for cognition. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 318–329). New York, NY: Guilford Press.

Rosenberg, M. (1989). *Society and the adolescent self-image* (rev. ed.). Middletown, CT: Wesleyan University Press.

Segerstrom, S. E., & Miller, G. E. (2004). Psychological stress and the human immune system: A meta-analytic study of 30 years of inquiry. *Psychological Bulletin*, 130, 601–630.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662.

6 Chapter 6: Sampling

6.1 Populations and Samples

We are usually interested in understanding a specific group of people. This group is known as the population of interest, or simply the population. The ***population*** is the collection of all people who have some characteristic in common; it can be as broad as "all people" if we have a very general research question about human psychology, or it can be extremely narrow, such as "all freshmen psychology majors at Midwestern public universities" if we have a specific group in mind.

In statistics, we often rely on a ***sample***—that is, a small subset of a larger set of data—to draw inferences about the larger set. The larger set is known as the population from which the sample is drawn. Let's say you have been hired by the National Election Commission to examine how the American people feel about the fairness of the voting procedures in the U.S. Whom will you ask? It is not practical to ask every single American how he or she feels about the fairness of the voting procedures. Instead, we query a relatively small number of Americans and draw inferences about the entire country from their responses. The Americans actually queried constitute our sample of the larger population of all Americans.

A sample is typically a small subset of the population. In the case of voting attitudes, we would sample a few thousand Americans drawn from the hundreds of millions that make up the country. In choosing a sample, it is therefore crucial that it not over-represent one kind of citizen at the expense of others. For example, something would be wrong with our sample if it happened to be made up entirely of Florida residents. If the sample held only Floridians, it could not be used to infer the attitudes of other Americans. The same problem would arise if the sample were composed only of Republicans. Inferences from statistics are based on the assumption that sampling is *representative* of the population. If the sample is not representative, then the possibility of ***sampling bias*** occurs. Sampling bias means that our conclusions apply only to our sample and are not generalizable to the full population.

Example

As another example, imagine we are interested in examining how many math classes have been taken, on average, by current graduating seniors at American colleges and universities during their four years in school. Whereas our population in our election example included all U.S. citizens, now it involves just the graduating seniors throughout the country. This

is still a large set since there are thousands of colleges and universities, each enrolling many students. (New York University, for example, enrolls 48,000 students.) It would be prohibitively costly to examine the transcript of every college senior. We therefore take a sample of college seniors and then make *inferences* to the entire population based on what we find. To make the sample, we might first choose some public and private colleges and universities across the United States. Then we might sample 50 students from each of these institutions. Suppose that the average number of math classes taken by the people in our sample was 3.2. We might speculate that 3.2 approximates the number we would find if we had the resources to examine every senior in the entire population. But we must be careful about the possibility that our sample is non-representative of the population. Perhaps we chose an overabundance of math majors, or chose too many technical institutions that have heavy math requirements. Such bad sampling makes our sample unrepresentative of the population of all seniors.

To solidify your understanding of the importance of representative sampling, consider the following examples. Try to identify the population of interest, and then reflect on whether the sample is likely to yield the information desired.

Example

A substitute teacher wants to know how students in the class did on their last test. The teacher asks the ten students sitting in the front row to state their latest test scores. He concludes from their report that the class did extremely well. What is the sample? What is the population? Can you identify any problems with choosing the sample in the way that the teacher did?
In this example, the population consists of all students in the class. The sample is made up of just the ten students sitting in the front row. The sample is not likely to be representative of the population. Those who sit in the front row tend to be more interested in the class and tend to perform higher on tests. Hence, the sample may perform at a higher level than the population.

Example

A coach is interested in how many cartwheels the average college freshman at his university can do. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping.
In this example, the population is the class of all freshmen at the coach's university. The sample is composed of 8 volunteers. The sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman; people who can't do cartwheels probably did not volunteer! In the example, we are also not told of the gender

of the volunteers. Were they all women, for example? That might affect the outcome, contributing to the non-representative nature of the sample (if the school is co-ed).

6.2 Probability versus Nonprobability Sampling

Essentially all psychological research involves ***sampling***—selecting a sample to study from the population of interest. Sampling falls into two broad categories. ***Probability sampling*** occurs when the researcher can specify the probability that each member of the population will be selected for the sample. It gives the likelihood of the sample being representative of the population.

Nonprobability sampling occurs when the researcher cannot specify these probabilities. Nonprobability sampling cannot assure the representativeness of a sample as well as probability sampling. However, it tends to be simpler and cheaper for researchers to conduct, and thus psychological research more often involves nonprobability sampling techniques. For example, ***convenience sampling***—studying individuals who happen to be nearby and willing to participate—is a very common form of nonprobability sampling used in psychological research. University participant pools (such as through SONA) are examples of convenience sampling. Another form of nonprobability sampling is ***snowball sampling***, in which existing research participants help recruit additional participants for the study. ***Quota sampling***, in which subgroups in the sample are recruited to be proportional to those subgroups in the population, is also common. For example, a researcher might want their sample to have proportional quotas across religious groups and set pre-determined quotas across those groups (e.g., Christian, Jewish, Muslim, etc.). Another nonprobabily sampling technique is ***purposive sampling***, in which only a certain type of people are sampled to fit the researcher's research question based on specific, pre-determined criteria. As an example, if a researcher is studying the psychological impact of transitioning to remote work, they might not be interested in sampling from the general population. Instead, the researcher might set out purposeful criteria of who to sample from, such as IT managers and remote employees who recently transitioned to remote work. Finally, ***self-selection sampling*** is when individuals choose to take part in the research on their own accord, without being approached by the researcher directly. To clarify, self-selection sampling does not mean the researcher participates in their own study (a common misconception).

Survey researchers, however, are much more likely to use some form of probability sampling. This tendency is because the goal of most survey research is to make accurate estimates about what is true in a particular population, and these estimates are most accurate when based on a probability sample. For example, it is important for researchers to base their estimates of election outcomes—which are often decided by only a few percentage points—on probability samples of likely registered voters. However, probability sampling is more difficult to implement

than nonprobability sampling. Therefore, some researchers may choose to use one of the nonprobability sampling techniques described since they tend to be easier to implement.

Researchers may also use nonprobability sampling techniques if generalizability is not a top priority. ***Generalizability*** is the extent to which the results of a study can be accurately applied to the population and is closely related to ***external validity*** (discussed in Module 5). If a researcher is trying to test an association or causal claim (establishing the relation between two variables of interest), then generalizability (and the utilization of probability sampling techniques) is usually a lower priority. In these cases, nonprobability sampling techniques may be preferred. However, if a researcher is trying to test a frequency claim (e.g., election outcomes), then generalizability is essential, making a probability sampling technique appropriate.

6.2.1 Probability Sampling Techniques

Compared with nonprobability sampling, probability sampling requires a very clear specification of the population, which of course depends on the research questions to be answered. The population might be all registered voters in the state of Arkansas, all American consumers who have purchased a car in the past year, women in the United States over 40 years old who have received a mammogram in the past decade, or all the alumni of a particular university. Once the population has been specified, probability sampling requires a ***sampling frame***. This is essentially a list of all the members of the population from which to select the respondents. Sampling frames can come from a variety of sources, including telephone directories, lists of registered voters, and hospital or insurance records. In some cases, a map can serve as a sampling frame, allowing for the selection of cities, streets, or households.

6.2.1.1 Simple Random Sampling

There are a variety of different probability sampling methods. ***Simple random sampling*** is done in such a way that each individual in the population has an equal probability of being selected for the sample. It is a reliable method of obtaining information because every single member of a population is chosen randomly, merely by chance. This could involve putting the names of all individuals in the sampling frame into a hat, mixing them up, and then drawing out the number needed for the sample. Given that most sampling frames take the form of computer files, random sampling is more likely to involve computerized sorting or selection of respondents. A common approach in telephone surveys is random-digit dialing, in which a computer randomly generates phone numbers from among the possible phone numbers within a given geographic area. A specific advantage of simple random sampling is that it is the most straightforward method of probability sampling. A disadvantage is that you may not find enough individuals with your characteristic of interest, especially if that characteristic is uncommon.

Sometimes it is not feasible to build a sample using simple random sampling. To see the problem, consider the fact that both Dallas and Houston competed to be hosts of the 2012 Olympics. Imagine that you had been hired to assess whether most Texans preferred Houston to Dallas as the host, or the reverse. Given the impracticality of obtaining the opinion of every single Texan, you had to construct a sample of the Texas population. But notice how difficult it would have been to proceed by simple random sampling. For example, how would you have contacted those individuals who didn't vote and didn't have a phone? Even among people you found in the telephone book, how could you have identified those who had just relocated to another state (and had no reason to inform you of their move)? What would you have done about the fact that, since the beginning of the study, an additional 4,212 people took up residence in the state of Texas? As you can see, it is sometimes very difficult to develop a truly random procedure. For this reason, other kinds of sampling techniques have been devised.

6.2.1.2 Stratified Sampling

A common alternative to simple random sampling is ***stratified random sampling***, in which the population is divided into different subgroups or "strata" (usually based on demographic characteristics) and then a random sample is taken from each "stratum." ***Proportionate stratified random sampling*** can be used to select a sample in which the proportion of respondents in each of various subgroups matches the proportion in the population. For example, because about 12.5% of the US population is Black, stratified random sampling can be used to ensure that a survey of 1,000 American adults includes about 125 Black respondents.

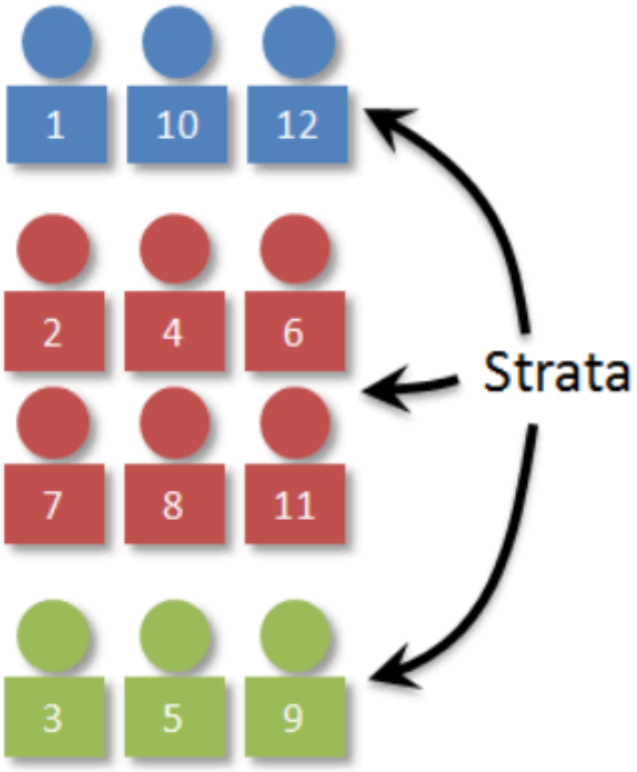


Figure 6.1: **Figure 6.1** Examples of Strat in a Population to be Sampled Via Stratified Random Sampling[1]

Disproportionate stratified random sampling (also referred to as ***oversampling***) can also be used to sample extra respondents from particularly small subgroups—allowing valid conclusions to be drawn about those subgroups. For example, because Asian Americans make up a fairly small percentage of the US population (about 4.5%), a simple random sample of 1,000 American adults might include too few Asian Americans to draw any conclusions about them as distinct from any other subgroup. If representation is important to the research question, however, then disproportionate stratified random sampling could be used to ensure that enough Asian American respondents are included in the sample to draw valid conclusions about Asian Americans as a whole.

6.2.1.3 Cluster Sampling

Yet another type of probability sampling is ***cluster sampling***, in which larger clusters of individuals are randomly sampled. Demographic characteristics, such as race/ethnicity, gender, age, and zip code can be used to identify a cluster. Cluster sampling can be more efficient than

simple random sampling, especially where a study takes place over a wide geographic region. Researchers can go a step further and engage in ***multistage sampling*** by randomly sampling individuals within each cluster. For example, to select a sample of small-town residents in the United States, a researcher might randomly select several small towns and then randomly select several individuals within each town. Cluster and multistage sampling are especially useful for surveys that involve face-to-face interviewing because it minimizes the amount of traveling that the interviewers must do. For example, instead of traveling to 200 small towns to interview 200 residents, a research team could travel to 10 small towns and interview 20 residents of each. The U.S. Census Bureau uses multistage sampling by first taking a simple random sample of counties in each state, then taking another simple random sample of households in each county, and collecting data on those households.

Exercise 6.1

Discussion. Identify an appropriate sampling frame for each of the following populations.

- students at a particular university
- adults living in the state of Washington
- households in Pullman, Washington
- people with low self-esteem

Please watch: Sampling Methods and Bias with Surveys (Crash Course Statistics #10)

Figure 6.2: Sampling Methods and Bias with Surveys: Crash Course Statistics #10

Please watch: Sampling: Simple Random, Convenience, systematic, cluster, stratified - Statistics Help

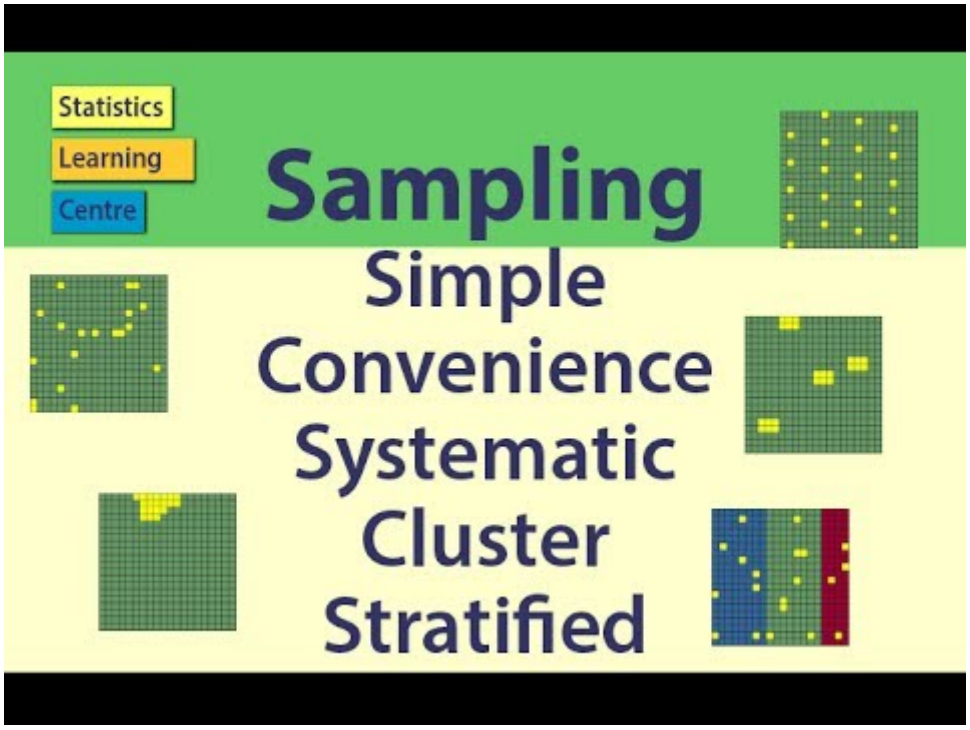


Figure 6.3: Sampling: Simple Random, Convenience, systematic, cluster, stratified - Statistics Help

6.3 Sample Size and Population Size

How large does a survey sample need to be? In general, this depends on two factors. One is the level of confidence in the result that the researcher wants. The larger the sample, the closer any statistic based on that sample will tend to be to the corresponding value in the population. The other factor is the budget of the study. Larger samples provide greater confidence, but they take more time, effort, and money to obtain. Taking these two factors into account, most survey research uses sample sizes that range from about 100 to about 1,000. Conducting a power analysis prior to launching the survey helps to guide the researcher in making this trade-off.

> **Note**
>
> **Why is a sample of 1,000 considered to be adequate for most survey research—even when the population is much larger than that? Consider, for example, that a sample of only 1,000 registered voters is generally considered a good sample of the roughly 120 million**

registered voters in the US population—even though it includes only about 0.0008% of the population! The answer is a bit surprising.
One part of the answer is that a statistic based on a larger sample will tend to be closer to the population value and that this can be characterized mathematically. Imagine, for example, that in a sample of registered voters, exactly 50% say they intend to vote for the incumbent. If there are 100 voters in this sample, then there is a 95% chance that the true percentage in the population is between 40 and 60. But if there are 1,000 voters in the sample, then there is a 95% chance that the true percentage in the population is between 47 and 53. Although this "95% confidence interval" continues to shrink as the sample size increases, it does so at a slower rate. For example, if there are 2,000 voters in the sample, then this only reduces the 95% confidence interval to 48 to 52. In many situations, the small increase in confidence beyond a sample size of 1,000 is not considered to be worth the additional time, effort, and money.
Another part of the answer—and perhaps the more surprising part—is that confidence intervals depend only on the size of the sample and not on the size of the population. So a sample of 1,000 would produce a 95% confidence interval of 47 to 53 regardless of whether the population size was a hundred thousand, a million, or a hundred million.

6.4 Statistical Power and Sample Size

Samples size and statistical power are two closely connected topics. The statistical power of a research design is the probability of rejecting the null hypothesis given the sample size and expected relation strength. For example, the statistical power of a study with 50 participants and an expected Pearson's r of +.30 in the population is .59. That is, there is a 59% chance of rejecting the null hypothesis if indeed the population correlation is +.30. Statistical power is the complement of the probability of committing a Type II error. So in this example, the probability of committing a Type II error would be $1 - .59 = .41$. Clearly, researchers should be interested in the power of their research designs if they want to avoid making Type II errors. In particular, they should make sure their research design has adequate power before collecting data. A common guideline is that a power of .80 is adequate. This guideline means that there is an 80% chance of rejecting the null hypothesis for the expected relationship strength.

The topic of how to compute power for various research designs and null hypothesis tests is beyond the scope of this book. However, there are online tools that allow you to do this by entering your sample size, expected relation strength, and level for various hypothesis tests (see "Computing Power Online"). In addition, Table 6.1 shows the sample size needed to

achieve a power of .80 for weak, medium, and strong relations for a two-tailed independent samples *t*-test and for a two-tailed test of Pearson's *r*. Notice that weak relations require very large samples to provide adequate statistical power.

Table 6.1: **Table 6.1** Sample Size Needed to Achieve Statistical Power of .80 for Different Expected Relation Strengths for an Independent Samples t-Test and a Test of Pearson's r [2]

Null Hypothesis Test **Relation Strength**	**Independent Samples *t*-Test**	**Test of Pearson's *r***
Strong ($d = .80$, $r = .50$)	52	28
Medium ($d = .50$, $r = .30$)	128	84
Weak ($d = .20$, $r = .10$)	788	782

What should you do if you discover that your research design does not have adequate power? Imagine, for example, that you are conducting a between-subjects experiment with 20 participants in each of two conditions and that you expect a medium difference ($d = .50$) in the population. The statistical power of this design is only .34. That is, even if there is a medium difference in the population, there is only about a one in three chance of rejecting the null hypothesis and about a two in three chance of committing a Type II error. Given the time and effort involved in conducting the study, this probably seems like an unacceptably low chance of rejecting the null hypothesis and an unacceptably high chance of committing a Type II error.

Given that statistical power depends primarily on relation strength and sample size, there are essentially two steps you can take to increase statistical power: increase the strength of the relation or increase the sample size. Increasing the strength of the relation can sometimes be accomplished by using a stronger manipulation or by more carefully controlling extraneous variables to reduce the amount of noise in the data (e.g., by using a within-subjects design rather than a between-subjects design). The usual strategy, however, is to increase the sample size. For any expected relation strength, there will always be some sample large enough to achieve adequate power.

i Computing Power Online

The following links are to tools that allow you to compute statistical power for various research designs and null
hypothesis tests by entering information about the expected relation strength, the sample size, and the level. They
also allow you to compute the sample size necessary to achieve your desired level of power (e.g., .80). The first is an
online tool. The second is a free downloadable program called G*Power.

- Russ Lenth's Power and Sample Size Page: http://www.stat.uiowa.edu/~rlenth/Power/index.html
- G*Power: http://www.gpower.hhu.de

6.5 Sampling Bias

Probability sampling was developed in large part to address the issue of sampling bias. ***Sampling bias*** occurs when a sample is selected in such a way that it is not representative of the entire population and therefore produces inaccurate results.

Sampling bias is an important consideration for researchers who seek to engage in social justice aligned research practices. Historically, many research samples consisted of White men but the findings from these studies were applied to individuals across racial groups. One key example was the creation of the body mass index (BMI) and its connection to health access. The original research to develop BMI was done in Europe with White men and women. Further research was done in the US in the 1970s but the sample size was still not racially diverse. Nevertheless, the findings were implemented across many racial and ethnic groups, globally. The implications of BMI have led to difficulties with health and insurance access for many racially marginalized communities. It highlights the importance of considering our sampling procedures to ensure that the generalizability of the findings are accurate.

There is one form of sampling bias that even careful random sampling is subject to. It is almost never the case that everyone selected for the sample actually responds to the survey. Some may have died or moved away, and others may decline to participate because they are too busy, are not interested in the survey topic, or do not participate in surveys on principle. If these survey non-responders differ from survey responders in systematic ways, then this can produce ***non-response bias***. For example, in a mail survey on alcohol consumption, researcher Vivienne Lahaut and colleagues found that only about half the sample responded after the initial contact and two follow-up reminders (Lahaut et al., 2002). The danger here is that the half who responded might have different patterns of alcohol consumption than the half who did not, which could lead to inaccurate conclusions on the part of the researchers. So to test for non-response bias, the researchers later made unannounced visits to the homes of a subset of the non-responders—coming back up to five times if they did not find them at home. They found that the original non-responders included an especially high proportion of abstainers (nondrinkers), which meant that their estimates of alcohol consumption based only on the original responders were too high.

Although there are methods for statistically correcting for non-response bias, they are based on assumptions about the non-responders—for example, that they are more similar to late responders than to early responders—which may not be correct. For this reason, the best approach to minimizing non-response bias is to minimize the number of non-responders—that

is, to maximize the response rate. There is a large research literature on the factors that affect survey response rates (Groves et al., 2004). In general, in-person interviews have the highest response rates, followed by telephone surveys, and then mail and Internet surveys. Among the other factors that increase response rates are sending potential respondents a short prenotification message informing them that they will be asked to participate in a survey in the near future and sending simple follow-up reminders to non-responders after a few weeks. The perceived length and complexity of the survey also makes a difference, which is why it is important to keep survey questionnaires as short, simple, and on topic as possible. Finally, offering an incentive—especially cash—is a reliable way to increase response rates. However, ethically, there are limits to offering incentives that may be so large as to be considered coercive.

6.6 Summary

- Research usually involves probability sampling, in which each member of the population has a known probability of being selected for the sample. Types of probability sampling include simple random sampling, stratified random sampling, and cluster sampling.
- Sampling bias occurs when a sample is selected in such a way that it is not representative of the population and therefore produces inaccurate results. The most pervasive form of sampling bias is non-response bias, which occurs when people who do not respond to the survey differ in important ways from people who do respond.
- The best way to minimize non-response bias is to maximize the response rate by prenotifying respondents, sending them reminders, constructing questionnaires that are short and easy to complete, and offering incentives.

6.7 Media Attributions

[1] D'Costa, S., Ukeye, M., O'Neil, M., & Anguiano, R. (no date). Critical Research Methods in Psychology. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

[2] Jhangiani, R.S., Chiang, I-C.A., Cuttler, C., & Leighton, D.C. (2019). Research methods in psychology. 4th edition. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

6.8 Text Attributions

D'Costa, S., Ukeye, M., O'Neil, M., & Anguiano, R. (no date). Critical Research Methods in Psychology. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Dudley, M. (2019). Research Methods in Psychology. OER Commons. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Jhangiani, R.S., Chiang, I-C.A., Cuttler, C., & Leighton, D.C. (2019). Research methods in psychology. 4th edition. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

6.9 References

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology.* Hoboken, NJ: Wiley.

Lahaut, V. M. H. C. J., Jansen, H. A. M., van de Mheen, D., & Garretsen, H. F. L. (2002). Non-response bias in a sample survey on alcohol consumption. *Alcohol and Alcoholism*, 37, 256–260.

Part III

Module 3: Descriptive Statistics - A Refresher

7 Chapter 7: Descriptive and Inferential Statistics

7.1 What Are Statistics?

Statistics include numerical facts and figures. For instance:

- The largest earthquake measured 9.2 on the Richter scale.
- Men are at least 10 times more likely than women to commit murder.
- One in every eight South Africans is HIV positive.
- By the year 2050, there will be 12 people aged 65 and over for every new baby born.

The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on how the numbers are chosen and how the statistics are interpreted. For example, consider the following three scenarios and the interpretations based on the presented statistics. You will find that the numbers may be right, but the interpretation may be wrong. Try to identify a major flaw with each interpretation before we describe it.

1. A new advertisement for Ben & Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective.

Major flaw: Ice cream consumption generally increases in the months of June, July, and August regardless of advertisements. This effect is called a ***history effect*** and leads people to interpret outcomes as the result of one variable when another variable (in this case, one having to do with the passage of time) is actually responsible.

2. The more churches in a city, the more crime there is. Thus, churches lead to crime.

Major flaw: Both increased churches and increased crime rates can be explained by larger populations. In bigger cities, there are both more churches and more crime. This is an example of the ***third-variable problem***. Namely, a third variable can cause both situations; however, people erroneously believe that there is a causal relationship between the two primary variables rather than recognize that a third variable can cause both.

3. Seventy-five percent more interracial marriages are occurring this year than 25 years ago. Thus, our society now accepts interracial marriages.

Major flaw: We do not have all the information we need. What is the rate at which marriages are occurring? Suppose only 1% of marriages 25 years ago were interracial and so now 1.75% of marriages are interracial (1.75 is 75% higher than 1). This latter number is hardly evidence suggesting the acceptance of interracial marriages. In addition, the statistic provided does not rule out the possibility that the number of interracial marriages has seen dramatic fluctuations over the years and this year is not the highest. Again, there is simply not enough information to understand fully the impact of the statistics. As a whole, these examples show that statistics are not only facts and figures; they are something more than that. In the broadest sense, "***statistics***" refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data. Statistics is the language of science and data. The ability to understand and communicate using statistics enables researchers from different labs, different languages, and different fields to articulate to one another exactly what they have found in their work. It is an objective, precise, and powerful tool in science and in everyday life.

Please watch: What Is Statistics: Crash Course Statistics #1

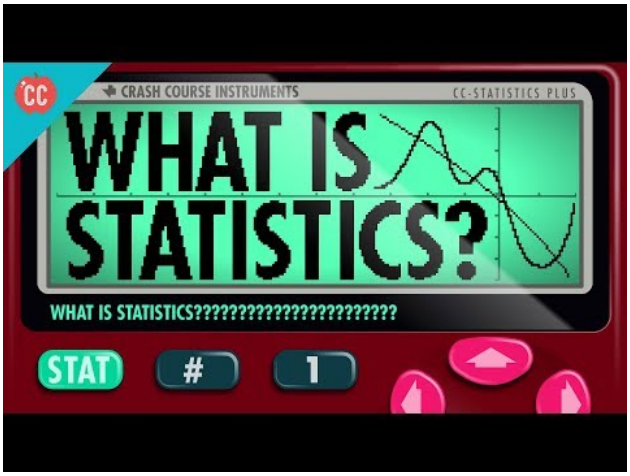


Figure 7.1: What Is Statistics: Crash Course Statistics #1

7.1.1 What a Statistics Course Is Not

Many psychology students dread the idea of taking a statistics course, and more than a few have changed majors upon learning that it is a requirement. That is because many students view statistics as only a math class, which is only partially true. While many of you will not believe this or agree with it, *statistics in psychology isn't just about math.*

Although math is a central component of it, statistics is a broader way of organizing, interpreting, and communicating information in an objective manner. Indeed, great care has been taken

to eliminate as much math from this course as possible (students who do not believe this are welcome to ask the professor what matrix algebra is). Statistics is a way of viewing reality as it exists around us in a way that we otherwise could not.

7.2 Why Do We Study Statistics?

To the surprise of many students, statistics is a fairly significant part of a psychological education. To the surprise of no-one, statistics is very rarely the *favorite* part of one's psychological education. After all, if you really loved the idea of doing statistics, you would probably be enrolled in a statistics class right now, not a psychology class. So, not surprisingly, a pretty large proportion of psychology students probably are not happy about the fact that psychology has so much statistics in it. Maybe you are one of them (I see you, nodding your head vigorously in agreement!). In view of this, the right place to start might be to answer the most common questions that people have about statistics (commonly shortened to stats).

A big part of this issue at hand relates to the very idea of statistics. What are they? What are they there for? And why are scientists so obsessed with them? These are all good questions, when you think about it. Let us start with the last one. As a group, scientists seem to be fixated on running statistical tests on everything. In fact, we use statistics so often that we sometimes forget to explain to people why we do. It is a kind of article of faith among scientists – and especially social scientists – that your findings cannot be trusted until you have done some stats. Undergraduate students might be forgiven for thinking that psychology professors are completely mad, because no-one takes the time to answer one very simple question:

7.2.1 Why Do Statistics? Why Not Just Use "Common Sense"?

The best answer is a really simple one: We do not trust ourselves enough. We recognize that we are human and, therefore, susceptible to all of the biases, temptations, and frailties from which humans suffer. Statistics are basically a safeguard. Using "common sense" (who knows what that is really, anyway?) to evaluate evidence means trusting gut instincts, relying on verbal arguments, on using the raw power of human reason to come up with the right answer. Most scientists do not think this approach is likely to work or be very accurate, and scientists like accuracy and evidence.

Virtually every student of the behavioral sciences takes some form of statistics class. This is because statistics are how we communicate in science. It serves as the link between a research idea and usable conclusions. Without statistics, we would be unable to interpret the massive amounts of information contained in data. Even small datasets contain hundreds—if not thousands—of numbers, each representing a specific observation we made. Without a way to organize these numbers into a more interpretable form, we would be lost, having wasted the time and money of our participants, ourselves, and the communities we serve.

Beyond its use in science, however, there is a more personal reason to study statistics. Like most people, you probably feel that it is important to "take control of your life." But, what does this mean? Partly, it means being able to properly evaluate the data and claims that bombard you every day. If you cannot distinguish good from faulty reasoning, then you are vulnerable to manipulation and to decisions that are not in your best interest. Statistics provides tools that you need in order to react intelligently to information you hear or read. In this sense, statistics is one of the most important things that you can study.

To be more specific, here are some claims that we have heard. (We are not saying that each one of these claims is true!)

- Four out of five dentists recommend Dentyne.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.
- Condoms are effective 94% of the time.
- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes when they work the same job.
- A surprising new study shows that eating egg whites can increase life span.
- People predict that it is very unlikely there will ever be another baseball player with a batting average over .400.
- There is an 80% chance that in a room full of 30 people, at least two people will share the same birthday.
- 79.48% of all statistics are made up on the spot.

All of these claims are statistical in character. We suspect that some of them sound familiar; if not, we bet that you have heard other claims like them. Notice how diverse the examples are. They come from psychology, health, law, sports, business, etc. Indeed, data (and data interpretation) show up in discourse from virtually every facet of contemporary life.

We hope that the discussion above helped explain why science in general is so focused on statistics. However, we are guessing that you have a lot more questions about what role statistics play in psychology, and specifically why psychology classes always devote so many lectures to stats. So, here is an attempt to answer a few of them…

7.2.2 Why Does Psychology Have So Many Statistics?

To be perfectly honest, many different reasons exist, some of which are better than others. The most important reason is that psychology is a statistical science. What I mean by that is that the "things" that we study are *people*. Real, complicated, gloriously messy, infuriatingly perverse people. The "things" of physics include objects like electrons, and while there are all sorts of complexities that arise in physics, electrons do not have minds of their own. They do not have opinions, they do not differ from each other in weird and arbitrary ways, they do not get bored in the middle of an experiment and scroll on their phones, and they do not get angry at the experimenter and then deliberately try to sabotage the data set (yes, this occasionally happens!). Most social sciences are desperately reliant on statistics. Not because we are bad experimenters, but because we have picked a harder problem to solve. We teach you stats because you really, really need it.

7.2.3 Can't Someone Else Besides Me Do the Statistics?

To some extent, perhaps, but not completely. It is true that you do not need to become a fully trained statistician just to do psychology, but you do need to reach a certain level of statistical tolerance and competence. We have found four reasons that every psychology student ought to be able to understand and do basic statistics:

- First, if you want to be good at doing research, then you need to be able to understand psychological literature, right? Here is the catch: Almost every paper in psychological literature reports the results of statistical analyses. So, if you really want to understand psychology, you need to be able to understand what other people did with their data, and that means deciphering a certain amount of statistics.
- Second, if you want to conduct research: Statistics is deeply intertwined with research design. If you want to be good at designing psychological studies, you need to at the very least understand the basics of stats.
- Thirdly, a big practical problem exists with being dependent on other people to do all your statistics: Statistical analysis is *expensive*. If you ever get bored and want to look up how much the Australian government charges for university fees, you would notice something interesting: Statistics is designated as a "national priority" category, and so the fees are much, much lower than for any other area of study. There is a massive shortage of statisticians out there. So, if you were a psychological researcher, the laws of supply and demand are not exactly on your side here! As a result, in almost any real-life situation where you want to do psychological research, the cruel facts will be that you do not have enough money to afford a statistician. The economics of the situation mean that you have to be pretty self-sufficient.

- Lastly, a lot of these reasons generalize beyond researchers. If you want to be a practicing psychologist and stay on top of the field to do right by your clients, it helps to be able to read the scientific literature, which relies pretty heavily on statistics.

7.2.4 I Don't Care About Jobs, Research, or Counseling Work. Do I Still Need Statistics?

Yes, we think it should matter to you, too. Statistics should matter to you personally in the same way that statistics should matter to *everyone* collectively. We live in the 21st century, and data are *everywhere*. Some would say science is under attack, fake news is rampant, AI systems often hallucinate, and our social media feeds us echo chambers of things it already thinks we want to hear. So, we can live blindly, or we can live smartly. Frankly, given the world in which we live these days, a basic knowledge of statistics is pretty much a survival tool!

Please watch Why Study Statistics in Psychology?

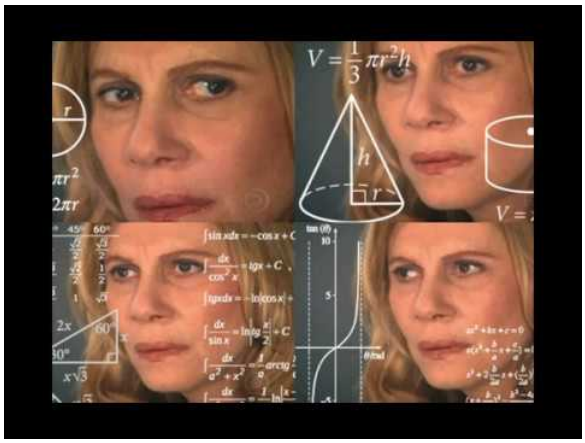


Figure 7.2: Why Study Statistics in Psychology?

Statistics are often presented in an effort to add credibility to an argument or advice. You can see this by paying attention to television advertisements. Many of the numbers thrown about in this way do not represent careful statistical analysis. They can be misleading and push you into decisions that you might find cause to regret. For these reasons, learning about statistics is a long step toward taking control of your life. (It is not, of course, the only step needed to do so.) The purpose of this course, beyond preparing you for a career in psychology, is to help you learn statistical essentials. It will make you into an intelligent consumer of statistical claims.

Please watch How Statistics can be Misleading - Mark Liddell

Figure 7.3: How statistics can be misleading - Mark Liddell

You can take the first step right away. To be an intelligent consumer of statistics, your first reflex must be to understand and question the statistics you encounter. The British Prime Minister Benjamin Disraeli is quoted by Mark Twain as having said, "There are three kinds of lies—lies, damned lies, and statistics." This quote reminds us why it is so important to understand statistics. So, let us invite you to reform your statistical habits from now on. No longer will you blindly accept numbers or findings. Instead, you will begin to think about the numbers, their sources, and most importantly, the procedures used to generate them.

The above section puts an emphasis on defending ourselves against fraudulent claims wrapped up as statistics, but let us also look at a more positive note. Just as important as detecting the deceptive use of statistics is the appreciation of the proper use of statistics. You must also learn to recognize statistical evidence that supports a stated conclusion. Statistics are all around you: sometimes used well, sometimes not. We must learn how to distinguish between the two cases. In doing so, statistics will likely be the course you use most in your day-to-day life, even if you do not ever run a formal analysis or do a research project again.

For those of you who are intending to go into some kind of helping profession with your psychology degree, think about it: Would you want your counselor to "go with their gut" and tell you whatever they feel like is best for you, or would you want a treatment or therapy that scientific studies have found to be helpful for clients with your condition? Would you want your insurance company to charge you whatever they want or give you a fair rate for your car insurance, based on your driving record and personal characteristics? If you wondered if your child had autism or ADHD, would you want your school counselor or school psychologist to use opinion or data and testing to make that diagnosis? Would you want your social worker or case manager to use an approach that has been shown to help people like you and those you care about and to help you find resources? If so, you can see the value of statistics in everyday

professions as well, beyond the research scientists that make a career out of producing scientific manuscripts regarding surveys and experiment results.

7.3 Descriptive Statistics

Descriptive statistics are numbers that are used to summarize and describe data. The word "***data***" refers to the information that has been collected from an experiment, a survey, a historical record, etc. (By the way, *data* are plural. One piece of information is called a ***datum***.) If we are analyzing birth certificates, for example, a descriptive statistic might be the percentage of certificates issued in New York State, or the average age of the mother. Any other number we choose to compute also counts as a descriptive statistic for the data from which the statistic is computed. Several descriptive statistics are often used at one time to give a full picture of the data.

Descriptive statistics are just that: descriptive. They describe and reflect what has happened. They do not involve generalizing or interpreting beyond the data at hand. Making conclusions from our data to another set of cases or individuals is the business of inferential statistics, which we will be studying in another chapter. Here, we focus on simply descriptive statistics.

Some descriptive statistics are shown in Table 7.1. The table shows the average salaries for various occupations in the United States in 1999. Descriptive statistics like these offer insight into American society. It is interesting to note, for example, that the people who educate our children and who protect our citizens are paid a great deal less than people who take care of our feet or our teeth.

Table 7.1: **Table 7.1** Average Salaries for Various U.S. Occupations in 1999[1]

Occupation	**Salary**
Pediatricians	$112,760
Dentists	$106,130
Podiatrists	$100,090
Physicists	$76,140
Architects	$53,410
School, clinical, and counseling psychologists	$49,720
Flight attendants	$47,910
Elementary school teachers	$39,560
Police officers	$38,710
Floral designers	$18,980

For more descriptive statistics, consider Table 7.2. It shows the number of unmarried men per 100 unmarried women in U.S. metro areas in 1990. From this table we see that men

outnumber women most in Jacksonville, North Carolina, and women outnumber men most in Sarasota, Florida. You can see that descriptive statistics can be useful if you were looking for an opposite-sex partner! (These data come from the Information Please Almanac.)

Table 7.2: **Table 7.2** Number of Unmarried Men per 100 Unmarried Women in U.S. Metro Areas in 1990 (Note: Unmarried includes never–married, widowed, and divorced persons, 15 years or older)[2]

Cities with Mostly Men	Men per 100 Women	Cities with Mostly Women	Men per 100 Women
1. Jacksonville, North Carolina	224	1. Sarasota, Florida	66
2. Killeen–Temple, Texas	123	2. Bradenton, Florida	68
3. Fayetteville, North Carolina	118	3. Altoona, Pennsylvania	69
4. Brazoria, Texas	117	4. Springfield, Illinois	70
5. Lawton, Oklahoma	116	5. Jacksonville, Tennessee	70
6. State College, Pennsylvania	113	6. Gadsden, Alabama	70
7. Clarksville–Hopkinsville, Tennessee–Kentucky	113	7. Wheeling, West Virginia–Ohio	70
8. Anchorage, Alaska	112	8. Charleston, West Virginia	71
9. Salinas–Seaside–Monterey, California	112	9. St. Joseph, Missouri	71
10. Bryan–College Station, Texas	111	10. Lynchburg, Virginia	71

These descriptive statistics may make us ponder why the numbers are so disparate in these cities. One potential explanation, for instance, as to why there are more women in Florida than men may involve the facts that older adults tend to move down to the Sarasota region and that women tend to outlive men. Thus, more women might live in Sarasota than men. However, in the absence of proper data, this is only speculation.

You probably know that descriptive statistics are central to the world of sports. Every sporting event produces numerous statistics, such as the shooting percentage of players on a basketball team. For the Olympic marathon (a foot race of 26.2 miles), we possess data that cover more than a century of competition. (The first modern Olympics took place in 1896.) Table 7.3 and Table 7.4 show the winning times for women and men, respectively. (Women have only been allowed to compete since 1984.)

Table 7.3: **Table 7.3** Women's Winning Olympic Marathon Times, 1984–2004[3]

Year	Winner	Country	Time
1984	Joan Benoit	United States	2:24:52
1988	Rosa Mota	Portugal	2:25:40
1992	Valentina Yegorova	Unified Team	2:32:41
1996	Fatuma Roba	Ethiopia	2:26:05
2000	Naoko Takahashi	Japan	2:23:14
2004	Mizuki Noguchi	Japan	2:26:20

Table 7.4: **Table 7.4** Men's Winning Olympic Marathon Times, 1896–2004[4]

Year	Winner	Country	Time
1896	Spyridon Louis	Greece	2:58:50
1900	Michel Théato	France	2:59:45
1904	Thomas Hicks	United States	3:28:53
1906	Billy Sherring	Canada	2:51:23
1908	Johnny Hayes	United States	2:55:18
1912	Kenneth McArthur	South Africa	2:36:54
1920	Hannes Kolehmainen	Finland	2:32:35
1924	Albin Stenroos	Finland	2:41:22
1928	Boughera El Ouafi	France	2:32:57
1932	Juan Carlos Zabala	Argentina	2:31:36
1936	Sohn Kee-chung	Japan	2:29:19
1948	Delfo Cabrera	Argentina	2:34:51
1952	Emil Zátopek	Czechoslovakia	2:23:03
1956	Alain Mimoun	France	2:25:00
1960	Abebe Bikila	Ethiopia	2:15:16
1964	Abebe Bikila	Ethiopia	2:12:11
1968	Mamo Wolde	Ethiopia	2:20:26
1972	Frank Shorter	United States	2:12:19
1976	Waldemar Cierpinski	East Germany	2:09:55
1980	Waldemar Cierpinski	East Germany	2:11:03
1984	Carlos Lopes	Portugal	2:09:21
1988	Gelindo Bordin	Italy	2:10:32
1992	Hwang Young-cho	South Korea	2:13:23
1996	Josia Thugwane	South Africa	2:12:36
2000	Gezahegne Abera	Ethiopia	2:10:10
2004	Stefano Baldini	Italy	2:10:55

There are many descriptive statistics that we can compute from the data in these tables. To

gain insight into the improvement in speed over the years, let us divide the men's times into two pieces, namely, the first 13 races (up to 1952) and the second 13 (starting from 1956). The mean winning time for the first 13 races is 2 hours, 44 minutes, and 22 seconds (written 2:44:22). The mean winning time for the second 13 races is 2:13:18. This is quite a difference (over half an hour). Does this prove that the fastest men are running faster? Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year? We cannot answer this question with descriptive statistics alone. All we can affirm is that the two means are "suggestive."

Examining Table 7.3 and Table 7.4 leads to many other questions. We note that Takahashi (the lead female runner in 2000) would have beaten the male runner in 1956 and all male runners in the first 12 marathons. This fact leads us to ask whether the gender gap will close or remain constant. When we look at the times within each gender, we also wonder how far they will decrease (if at all) in the next century of the Olympics. Might we one day witness a marathon winner finishing in under two hours? The study of statistics can help people make reasonable guesses about the answers to these questions.

It is also important to differentiate what we use to describe populations vs. what we use to describe samples. A population is described by a parameter; the ***parameter*** is the true value of the descriptive in the population, but one that we can never know for sure. For example, the Bureau of Labor Statistics once reported that the average hourly wage of chefs is $23.87. However, even if this number were computed using information from every single chef in the United States (making it a parameter), it would quickly become slightly off as one chef retires and a new chef enters the job market. Additionally, as noted above, there is virtually no way to collect data from every single person in a population. In order to understand a variable, we estimate the population parameter using a sample statistic. Here, the term ***statistic*** refers to the specific number we compute from the data (e.g., the average), not the field of statistics. A sample statistic is an estimate of the true population parameter, and if our sample represents the population well, then the statistic is considered to be a good and reasonable estimator of the parameter.

Even the best sample will be somewhat off from the full population, earlier referred to as ***sampling bias***, and as a result, there will always be a tiny discrepancy between the parameter and the statistic we use to estimate it. This difference is known as ***sampling error***, and, as we will see throughout the course, understanding sampling error is the key to understanding statistics. Every observation we make about a variable, be it a full research study or observing an individual's behavior, is incapable of being completely representative of all possibilities for that variable. Knowing where to draw the line between an unusual observation and a true difference is what statistics as a field of study is all about.

7.4 Inferential Statistics

Descriptive statistics are wonderful at telling us what our data look like. However, what we often want to understand is how our data behave. What variables are related to other variables? Under what conditions will the value of a variable change? Are two groups different from each other, and if so, are people within each group different or similar? These are the questions answered by inferential statistics, and inferential statistics are how we generalize from our sample back up to our population or how we test ideas and questions about the world (otherwise known as ***hypotheses***).

Module 4 is all about ***inferential statistics***, the formal analyses and tests we run to make conclusions about our data and the world. For example, we will learn how to use an r statistic to look at the relationship between two factors. We will learn how to use a t statistic to determine whether people change over time or differ when enrolled in an intervention or experimental condition. We will also use an F statistic to determine if we can predict future values on a variable based on current known values of a variable. There are many types of inferential statistics, each allowing us insight into a different behavior of the data we collect. This course will only touch on a small subset (or a ***sample***) of them, but the principles we learn along the way will make it easier to learn new tests, as most inferential statistics follow the same structure and format.

7.5 Mathematical Notation

As noted earlier, statistics is not just about math. It does, however, rely on logic and on math as a tool for logic. Many statistical formulas involve summing numbers. Fortunately, there is a convenient notation for expressing summation. This section covers the basics of this summation notation, for those whose professors might require hand calculations in their course.

Say we have a variable X that represents the weights (in grams) of 4 grapes:

Table 7.5: **Table 7.5** Weights of Grapes[5]

Grape	***X***
Grape 1	4.6
Grape 2	5.1
Grape 3	4.9
Grape 4	4.4

We label the weight of Grape 1 as X_1 of Grape 2 as X_2, etc. The following formula means to sum up the weights of the four grapes:

$$\sum_{i=1}^{4} X_i$$

The Greek letter Σ (sigma) indicates summation. The "i = 1" at the bottom indicates that the summation is to start with X1, and the 4 at the top indicates that the summation will end with X4. The "Xi" indicates that X is the variable to be summed as i goes from 1 to 4. Therefore:

$$\sum_{i=1}^{4} X_i = \ x_1 + x_2 + x_3 + x_4 = 4.6 + 5.1 + 4.9 + 4.4 = 19$$

As another example, the symbol

$$\sum_{i=1}^{3} x_i$$

indicates that only the first 3 scores are to be summed. The index variable i goes from 1 to 3.

When all the scores of a variable (such as i) are to be summed, it is often convenient to use the following abbreviated notation:

$$\sum X$$

Thus, when no values of i are shown, it means to sum all the values of X.

Many formulas involve squaring numbers before they are summed. This is indicated as:

$$\sum X^2 = 4.6^2 + 5.1^2 + 4.9^2 + 4.4^2 = 21.16 + 26.01 + 24.01 + 19.36 = 90.54$$

Notice that:

$$(\sum X)^2 \ \neq \ \sum X^2$$

because the expression on the left means to sum up all the values of X and then square the sum ($19^2 = 361$), whereas the expression on the right means to square the numbers and then sum the squares (90.54, as shown).

Some formulas involve the sum of cross products (multiplying). Below are the data for variables X and Y. The cross products (XY) are shown in the third column. The sum of the cross products is 3 + 4 + 21 = 28.

Table 7.6: **Table 7.6** The Sum of Cross-Products (Multiplying X and Y)[6]

X	***Y***	***XY***
1	3	3
2	2	4
3	7	21

In summation notation, this is written as:

$$\sum XY = 28$$

7.6 Wrapping it Up

Hopefully, this section has helped you see the importance of statistics as an important tool for scientific thinking for psychology students and practitioners, even if you are not a fan of math and research is not your jam. Statistics help us answer questions and make important decisions, both in our jobs and in everyday life.

7.7 Additional Resources

For tips and demonstrations on setting up and running statistical analyses in SPSS, please see Virginia Wickline: SPSS Statistics Helps (publishing dates vary).

7.8 Media Attributions

[1-6] (CC BY-SA 4.0 by Cote et al., 2021)

7.9 Text Attributions

Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). *Introduction to statistics in the psychological sciences.* Pressbooks. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Navarro, D. J., & Foxcroft, D. R. (2025). *Learning statistics with jamovi: A tutorial for beginners in statistical analysis.* Open Book Publishers. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

7.10 References

CrashCourse (2018, January 24). *What is statistics: Crash course statistics #1.* YouTube. https://www.youtube.com/watch?v=sxQaBpKfDRk

Liddell, M. (2016, January 14). *How statistics can be misleading - Mark Liddell.* TED-Ed. https://www.youtube.com/watch?v=sxYrzzy3cq8

Storage, D. (2019, June 17). *Why study statistics in psychology?* YouTube. https://www.youtube.com/watch?v=nQ_5ta7_jyE

8 Chapter 8: Describing Data Using Distributions and Graphs

Before we can understand our analyses, we must first understand our data. The first step in doing this is using tables, charts, graphs, plots, and other visual tools to see what our data look like. We will not go over every possible graph out there in this chapter, but we will highlight some of the most common ones. Note that the words chart, graph, and plot can be used interchangeably, and we tend to call these figures. ***Figures*** can include any kind of image, but all kinds of graphs would be a kind of figure. Figures are different than ***tables***, which are data collected in rows and columns of information. Tables are not figures, and figures are not tables. Tables and figures both show data, but they are different visual ways of doing so.

8.1 Graphing Categorical Variables

When Apple Computer introduced the iMac computer in August 1998, the company wanted to learn whether the iMac was expanding Apple's market share. Was the iMac just attracting previous Macintosh owners? Or was it purchased by newcomers to the computer market and by previous Windows users who were switching over? To find out, 500 iMac customers were interviewed. Each customer was categorized as a previous Macintosh owner, a previous Windows owner, or a new computer purchaser.

This section examines graphical methods for displaying the results of the interviews. We will learn some general lessons about how to graph data that fall into a small number of categories (otherwise known as ***categorical data*** or ***qualitative data***). The key point about the categorical data in the present section is that they do not come with a pre-established ordering (the way numbers are ordered). For example, there is no natural sense in which the category of previous Windows users comes before or after the category of previous Macintosh users. Categorical data are like bins for sorting trash. They are assigned a number, but the numbering is arbitrary.

Figure 8.1: Photograph of three recycling bins labeled for drink cans, glass and plastic bottles, and paper, each in yellow, red, and blue respectively. Bins are placed under a roof with color-coded images and text to guide proper waste sorting for recycling purposes.

Figure 8.1 Bins for recycling different kinds of materials[1]

A later section will consider how to graph numerical data in which each observation is represented by a number in some range (otherwise known as ***continuous data*** or ***quantitative data***). Continuous data, for example, could include something like height. People of one height are naturally ordered with respect to people of a different height, from shortest to tallest.

Figure 8.2: **Figure 8.2** Original members of the legendary rock band, Kiss, ordered from shortest to tallest[2]

8.1.1 Frequency Distributions

Frequency means how often something happens. It is a count of the total number of responses for a given condition or choice. When we collect data, we often want an idea of who said or did

what, and we want to show those responses as a batch. In other words, we want to create a ***frequency distribution*** that shows each of the responses and how many individuals represent each option. Frequency distributions can be shown in two ways: tables (with rows and columns of data) or figures (graphs or charts).

8.1.2 Frequency Tables

All of the graphical methods shown in this section are first derived from frequency tables. Table 8.1 shows a ***frequency table*** for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the ***relative frequency***, which is the proportion of responses in each category. For example, the relative frequency (.17) for "none" is 85 (people that picked "none") divided by 500 (the total number of respondents): 85/500 = .17.

Table 8.1: **Table 8.1** Frequency Table for the iMac Data[3]

Previous Ownership	**Frequency**	**Relative Frequency**
None	85	.17
Windows	60	.12
Macintosh	355	.71
Total	500	1.00

8.1.3 Pie Charts

The ***pie chart*** in Figure 8.3 shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners (71%), Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.

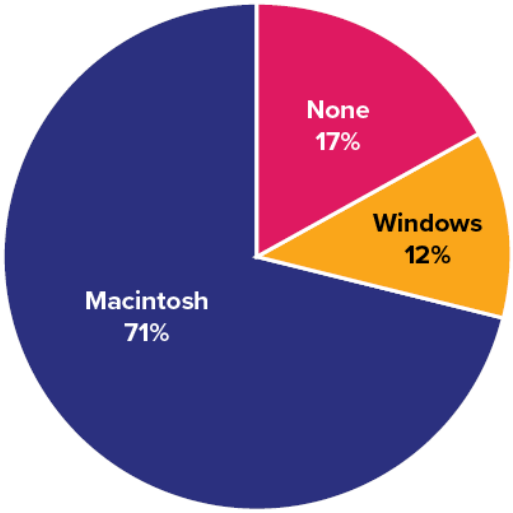


Figure 8.3: **Figure 8.3** Pie chart of iMac purchases illustrating frequencies of previous computer ownership: 71% of purchasers owned a Macintosh before buying their iMac[4]

Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, when you have a large number of categories. Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments. In an influential book on the use of graphs, Edward Tufte (1983) asserted, "The only worse design than a pie chart is several of them" (p. 178).^^

Here is another important point about pie charts. If they are based on a small number of observations, it can be misleading to label the pie slices with percentages. For example, if just 5 people had been interviewed by Apple Computers, and 3 were former Windows users, it would be misleading to display a pie chart with the Windows slice showing 60%. With so few people interviewed, such a large percentage of Windows users might easily have occurred since chance can cause large errors with small samples. In this case, it is better to alert the user of the pie chart to the actual numbers involved. The slices should therefore be labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

8.1.4 Bar Charts

Bar charts can also be used to represent frequencies of different categories. Think of bar charts like separate bins: In which bin do you put your trash versus recycling versus composting? Or how can you sort your laundry into bins to wash it more effectively, with like items in each load (e.g., lights, darks, towels, delicates, dog stuff)? A bar chart of the iMac purchases is shown in Figure 8.4. Notice that there are gaps between the bars because each group is unique. Frequencies are shown on the ***y-axis*** down the side (also known as the ordinate) and the type of computer previously owned is shown on the ***x-axis*** across the bottom (also known as the

abscissa). Typically, the y-axis shows the number of observations (frequency) in each category rather than the percentage of observations in each category as is typical in pie charts.

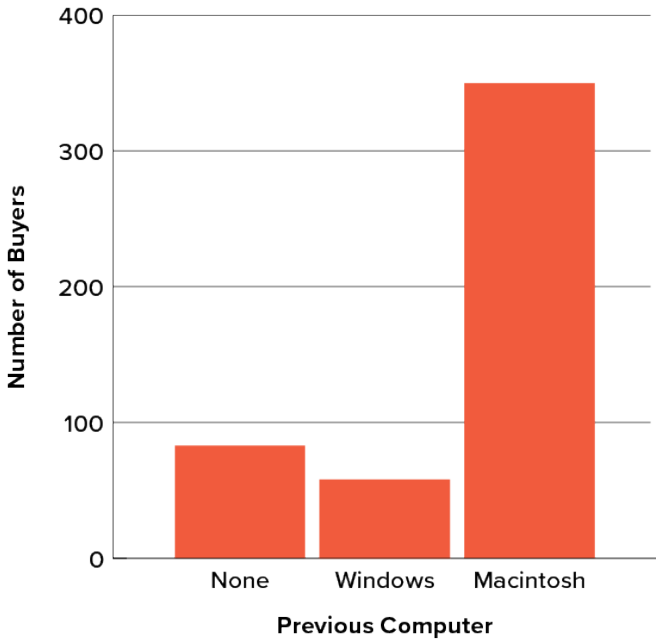


Figure 8.4: **Figure 8.4** Bar chart of iMac purchases as a function of previous computer ownership[5]

8.1.5 Comparing Distributions

Often, we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the ***distributions*** or collections of responses between the surveys or conditions. Bar charts are also excellent for illustrating differences between two distributions. Figure 8.5 shows the number of people playing card games at the Yahoo web site on a Sunday and on a Wednesday in the spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing Hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

The bars in Figure 8.5 are oriented horizontally rather than vertically. The horizontal format is useful with many categories because there is more room for the category labels. There will be more about bar charts when we consider numerical quantities later in this chapter.

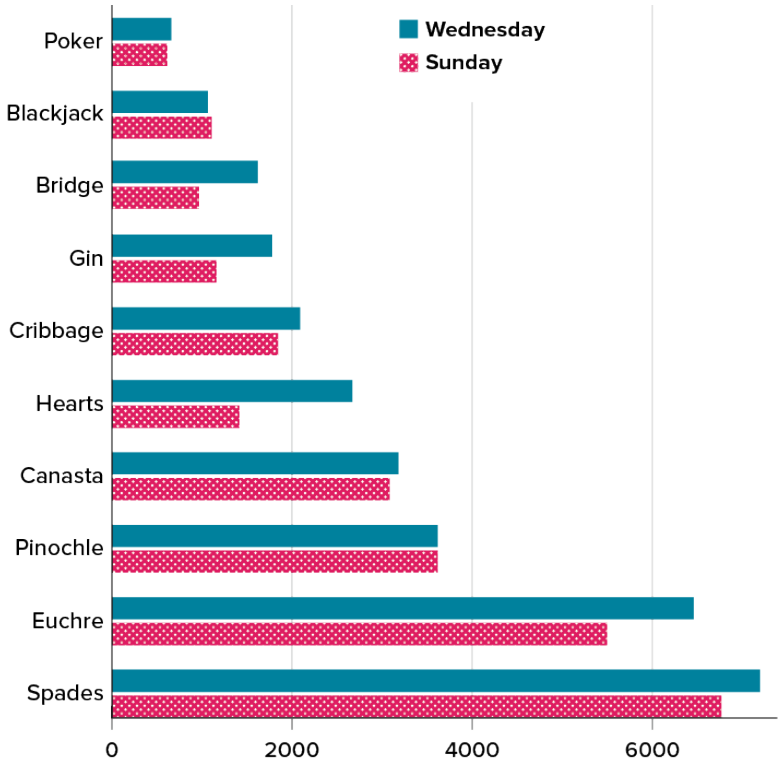


Figure 8.5: **Figure 8.5** A bar chart of the number of people playing different card games on Sunday and Wednesday[6]

Please watch Charts Are Like Pasta - Data Visualization Part 1: Crash Course Statistics #5

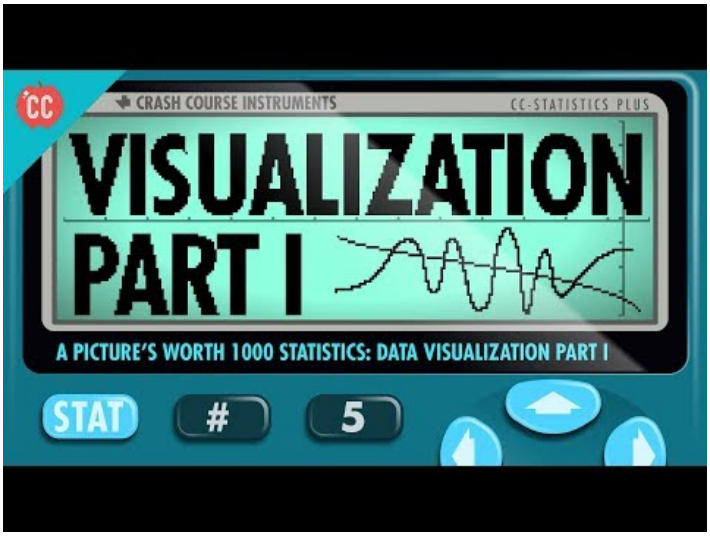


Figure 8.6: Charts Are Like Pasta - Data Visualization Part 1: Crash Course Statistics #5

8.1.6 Some Graphical Mistakes to Avoid

First mistake: Don't get too fancy! Sometimes, more is less. In many instances, people add features to graphs that do not help to convey their information. For example, three-dimensional bar charts such as the one shown in Figure 8.6 are not necessary and are usually not as effective as their two-dimensional counterparts.

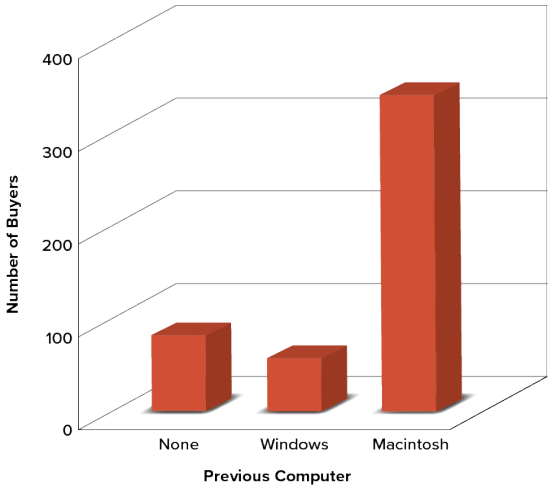


Figure 8.7: **Figure 8.6** A three-dimensional version of Figure 8.4. Charts like this are less effective[7]

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful or intriguing images. For example, Figure 8.7 presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure 8.7 is misleading because the viewer's attention will be captured by areas. The areas can exaggerate the size differences between the groups. In terms of percentages, the ratio of previous Macintosh owners to previous Windows owners is about 6 to 1. But the ratio of the two areas in Figure 8.7 is about 35 to 1. A biased person wishing to hide the fact that many Windows owners purchased iMacs would be tempted to use Figure 8.7 instead of Figure 8.4!

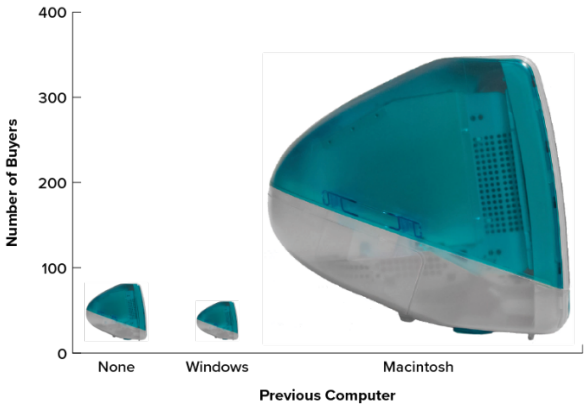


Figure 8.8: **Figure 8.7** A redrawing of Figure 8.4 with a lie factor greater than 8[8]

Tufte (1983) coined the term ***lie factor*** to refer to the ratio of the size of the effect shown in a graph to the size of the effect shown in the data. He suggests that lie factors greater than 1.05 or less than 0.95 produce unacceptable distortion.

Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the y-axis, representing the least number of cases that could have occurred in a category. Normally, but not always, this number should be zero. Figure 8.8 shows the iMac data with a baseline of 50 on the y-axis instead of zero. Once again, the differences in areas suggest a different story than the true differences in percentages. The number of Windows-switchers seems minuscule compared to its true value of 12%. In everyday life, watch out for graphs that do not start their baseline at zero.

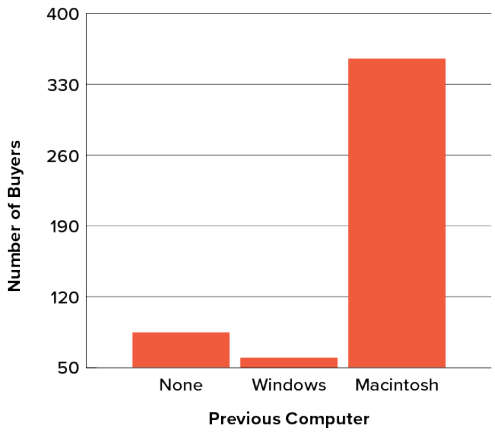


Figure 8.9: **Figure 8.8** A redrawing of Figure 8.4 with a baseline of 50[9]

Finally, we note that it is a serious mistake to use a line graph when the x-axis contains categorical variables. A ***line graph*** is essentially a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). Figure 8.9 inappropriately shows a line graph of the card game data from Yahoo that was presented in Figure 8.5. The drawback to Figure 8.9 is that it gives the false impression that the games are naturally ordered in a numerical way when, in fact, they are ordered alphabetically.

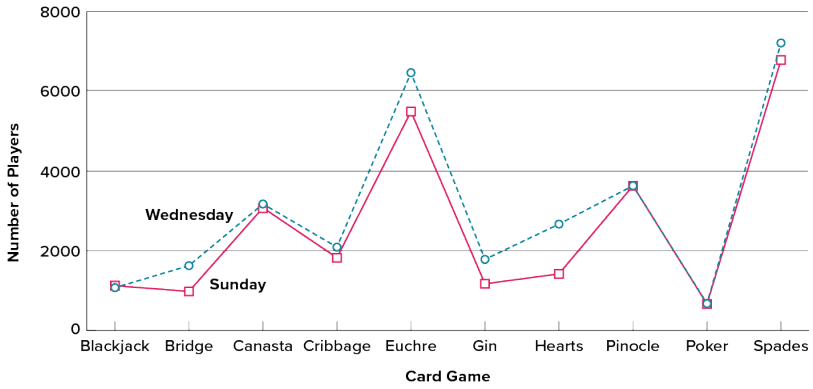


Figure 8.10: A line chart comparing the number of players for various card games on Sunday and Wednesday. Each card game—Blackjack, Bridge, Canasta, Cribbage, Euchre, Gin, Hearts, Pinochle, Poker, and Spades—has two data points: pink squares for Sunday and blue circles for Wednesday. Wednesday values follow a dashed blue line, and Sunday values follow a solid pink line. Player counts vary widely, with peaks at Euchre and Spades and much lower counts for games like Poker and Gin.

Figure 8.9 A line graph used inappropriately to depict the number of people playing different card games on Sunday and Wednesday[10]

Please watch: How to Spot a Misleading Graph – Lea Gaslowitz

Figure 8.11: How to spot a misleading graph - Lea Gaslowitz

8.1.7 Summary

Pie charts and bar charts can both be effective methods of portraying categorical data. Bar charts are better when there are more than just a few categories and for comparing two or more distributions. Be careful to avoid creating misleading graphs.

8.2 Graphing Continuous Variables

Continuous variables are measured on a numeric scale. Height, weight, response time, subjective rating of pain, temperature, and score on an exam are all examples of continuous variables. Continuous variables are distinguished from categorical variables (sometimes also called nominal variables), such as favorite color, religion, city of birth, and favorite sport, in which there is no ordering or measuring involved, even though a number is assigned arbitrarily.

Many types of graphs can be used to portray distributions of continuous variables. The upcoming sections cover the following types of graphs: (1) stem-and-leaf displays, (2) histograms, (3) box plots, (4) bar charts, (5) line graphs, and (6) scatter plots (discussed in the section on correlation, also sometimes spelled scatterplots). Some graph types, such as stem-and-leaf displays, are best-suited for small to moderate amounts of data, whereas others, such as histograms, are best-suited for large amounts of data. Graph types such as box plots or bar charts are good as frequency distributions or for depicting differences between distributions. Scatterplots are used to show the relationship between two variables.

8.2.1 Stem-and-Leaf Displays

A ***stem-and-leaf display*** is a graphical method of displaying data that is particularly useful when data are not too numerous. In this section, we will explain how to construct and interpret this kind of graph.

As usual, we will start with an example. Consider Figure 8.10, which shows the number of touchdown passes (TD passes) thrown by each of the 31 teams in the National Football League during the 2000 season.

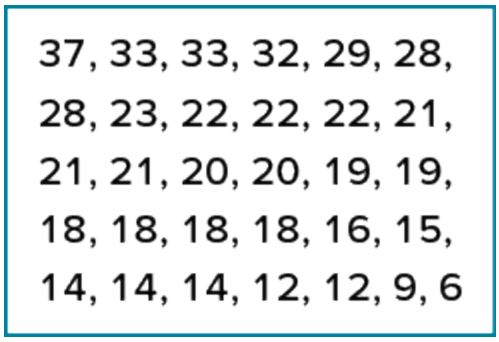


Figure 8.12: **Figure 8.10** Number of touchdown passes[11]

A stem-and-leaf display of the data is shown in Figure 8.11. The left portion of Figure 8.11 contains the stems. They are the numbers 3, 2, 1, and 0, arranged as a column to the left of the bars. Think of these numbers as 10s digits. A stem of 3, for example, can be used to represent the 10s digit in any of the numbers from 30 to 39. The numbers to the right of the bar are leaves, and they represent the 1s digits. Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem.

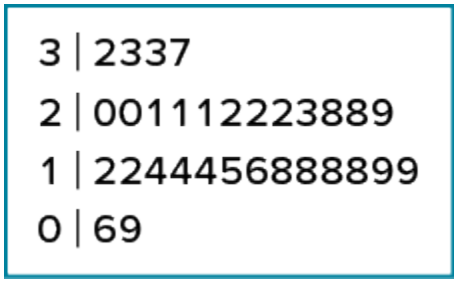


Figure 8.13: **Figure 8.11** Stem-and-leaf display of the number of touchdown passes[12]

To make this clear, let us examine Figure 8.11 more closely. In the top row, the four leaves to the right of stem 3 are 2, 3, 3, and 7. Combined with the stem, these leaves represent the numbers 32, 33, 33, and 37, which are the numbers of TD passes for the first four teams in

Figure 8.10. The next row has a stem of 2 and 12 leaves. Together, they represent 12 data points, namely, two occurrences of 20 TD passes, three occurrences of 21 TD passes, three occurrences of 22 TD passes, one occurrence of 23 TD passes, two occurrences of 28 TD passes, and one occurrence of 29 TD passes. We leave it to you to figure out what the third row represents. The fourth row has a stem of 0 and two leaves. It stands for the last two entries in Figure 8.10, namely 9 TD passes and 6 TD passes. (The latter two numbers may be thought of as 09 and 06.)

One purpose of a stem-and-leaf display is to clarify the shape of the distribution. You can see many facts about TD passes more easily in Figure 8.11 than in Figure 8.10. For example, by looking at the stems and the shape of the plot, you can tell that most of the teams had between 10 and 29 passing TDs, with a few having more and a few having less. The precise numbers of TD passes can be determined by examining the leaves.

There are two things about the football data that make them easy to graph with stems and leaves. First, the data are limited to whole numbers that can be represented with a one-digit stem and a one-digit leaf. Second, all the numbers are positive. If the data include numbers with three or more digits or contain decimals, they can be rounded to two-digit accuracy. Negative values are also easily handled.

Although stem-and-leaf displays are unwieldy for large datasets, they are often useful for datasets with up to 200 observations. Whether your data can be suitably represented by a stem-and-leaf display depends on whether they can be rounded without loss of important information. Also, all values must be two successive digits. Deciding what kind of graph is best suited to displaying your data thus requires good judgment. Statistics is not just recipes!

8.2.2 Box Plots

Box plots (sometimes also called box-and-whisker plots) are useful for identifying extreme scores and for comparing distributions. We will explain box plots with the help of data from a small experiment. Students in Introductory Statistics were presented with a page containing 30 colored rectangles. Their task was to name the colors as quickly as possible. Their times (in seconds) were recorded. We will compare the scores for the 16 men and 31 women who participated in the experiment by making separate box plots for each gender. Such a display is said to involve ***parallel box plots***. The data for the women in our sample are shown in Figure 8.12.

14, 15, 16, 16, 17, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 19, 19, 19
20, 20, 20, 20, 20, 20, 21, 21, 22, 23, 24, 24, 29

Figure 8.14: **Figure 8.12** Women's times, raw data[13]

There are several steps in constructing a box plot. The first relies on the 25th, 50th, and 75th percentiles in the distribution of scores. Figure 8.13 shows how these three statistics are used. For each gender we draw a box extending from the 25th percentile to the 75th percentile. The 50th percentile (otherwise known as the median, see Module 4) is drawn inside the box. Therefore, the bottom of each box is the 25th percentile, the top is the 75th percentile, and the line in the middle is the 50th percentile. Together, this box is otherwise known as the interquartile range, see Module 4).

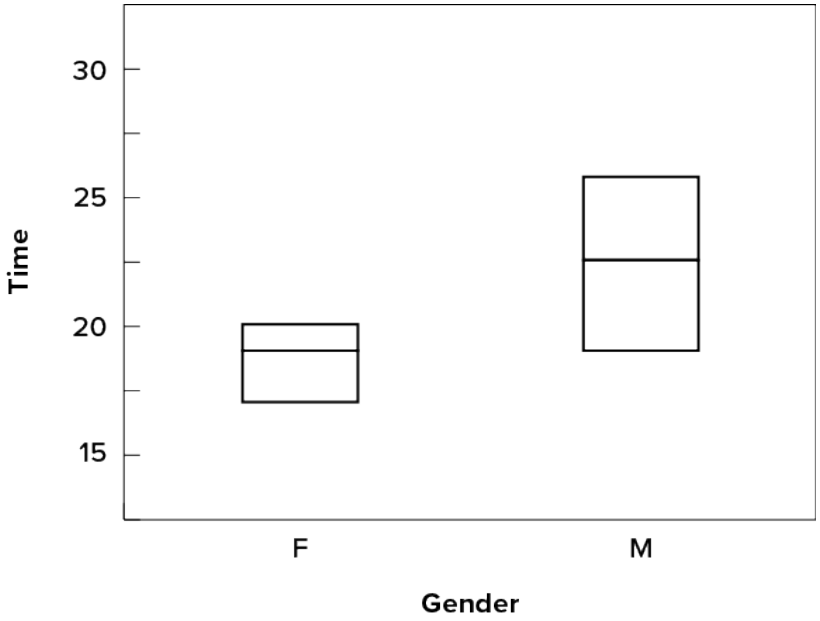


Figure 8.15: **Figure 8.13** The first step in creating box plots[14]

For the data reflecting the women's times, the 25th percentile is 17, the 50th percentile is 19, and the 75th percentile is 20. For the men (whose data are not shown), the 25th percentile is 19, the 50th percentile is 22.5, and the 75th percentile is 25.5.

Continuing with the box plots, we put "whiskers" above and below each box to give additional information about the spread of data. ***Whiskers*** are vertical lines that end in a horizontal stroke. Whiskers are drawn at 1.5 times the ***interquartile range***, which is the 75th percentile minus the 25th percentile.

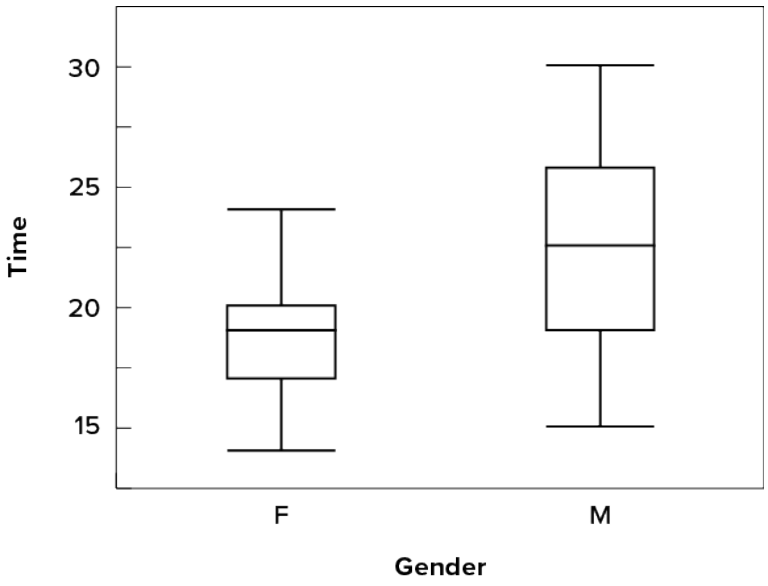


Figure 8.16: **Figure 8.14** The box plots with the whiskers drawn[15]

Although whiskers are not drawn all the way to outside or far-out values, we still wish to represent them in our box plots. This is achieved by adding additional marks beyond the whiskers. Specifically, outside values are indicated by small circles (^{0}), and far out values are indicated by asterisks (*). In our data, there are no far-out values and just one outside value. This outside value of 29 is for the women and is shown in Figure 8.15.

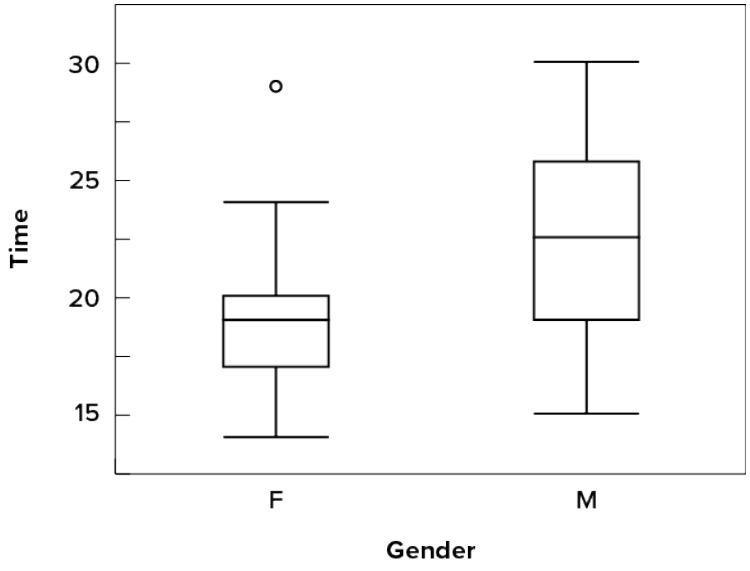


Figure 8.17: **Figure 8.15** The box plots with the outside value shown[16]

Box plots provide basic information about a distribution. For example, a distribution with a that is heavier on the upper end would have a longer whisker in the positive direction than in the negative direction. Box plots are good at portraying extreme values and are especially good at showing differences between distributions. However, many of the details of a distribution are not revealed in a box plot; to examine these details one should create a histogram and/or a stem-and-leaf display.

8.2.3 Bar Charts

In the section on categorical variables, we saw how bar charts could be used to illustrate the frequencies of different categories. For example, as we saw earlier in this chapter, the bar chart shown in Figure 8.4 shows how many purchasers of iMac computers were previous Macintosh users, previous Windows users, and new computer purchasers.

In this section we show how bar charts can be used with two variables to present other kinds of continuous information, not just frequency counts. The bar chart in Figure 8.16 shows the percent increases in the Dow Jones, Standard & Poor 500 (S&P), and Nasdaq stock indexes from May 24, 2000, to May 24, 2001. Notice that both the S&P and the Nasdaq had "negative increases" which means that they decreased in value. In this bar chart, the y-axis is not frequency but rather another variable, the *percentage increase*.

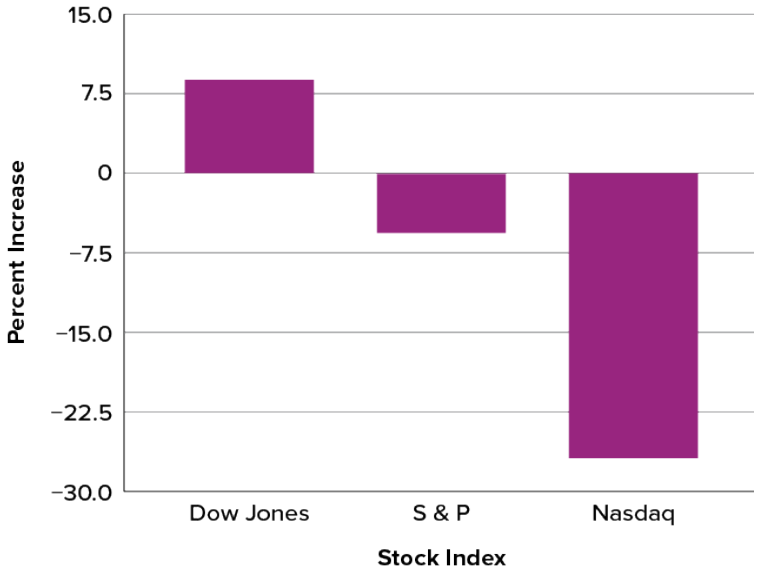


Figure 8.18: **Figure 8.16** Percent increase in three stock indexes from May 24, 2000, to May 24, 2001[17]

Bar charts are particularly effective for showing change over time. Figure 8.17 for example,

shows the percent increase in the Consumer Price Index (CPI) over four three-month periods. The fluctuation in inflation is apparent in the graph.

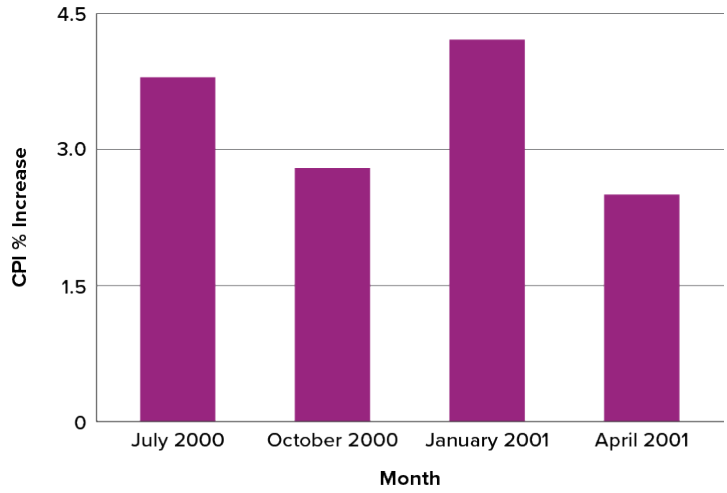


Figure 8.19: **Figure 8.17** Percent change in the Consumer Price Index (CPI) over time. Each bar represents percent increase for the three months ending at the date indicated[18]

Bar charts are often used to compare the means of different experimental conditions. Figure 8.18 shows the mean time it took one person to move the cursor to either a small target or a large target. On average, more time was required for small targets than for large ones.

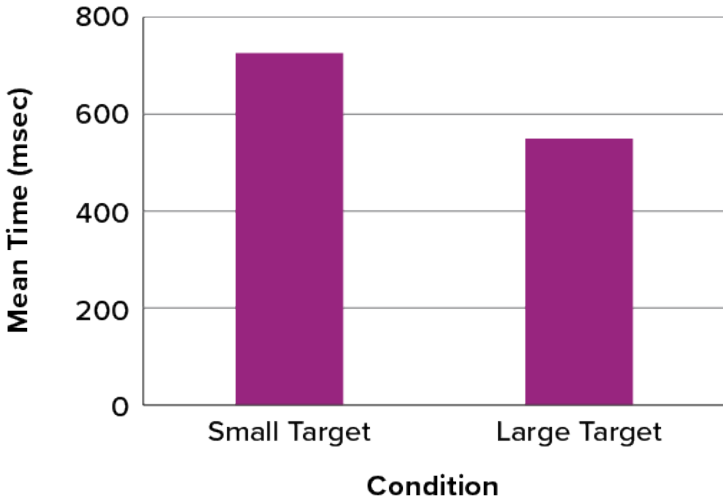


Figure 8.20: **Figure 8.18.** Bar chart showing the means for the two conditions.[19]

8.2.4 Line Graphs

A line graph is like a bar graph but with the tops of the bars represented by points joined by lines (and the rest of the bar is suppressed). For example, Figure 8.17, which was presented in the section on bar charts, shows changes in the Consumer Price Index (CPI) over time. A line graph of these same data is shown in Figure 8.19. Although the figures are similar, the line graph emphasizes the change from period to period.

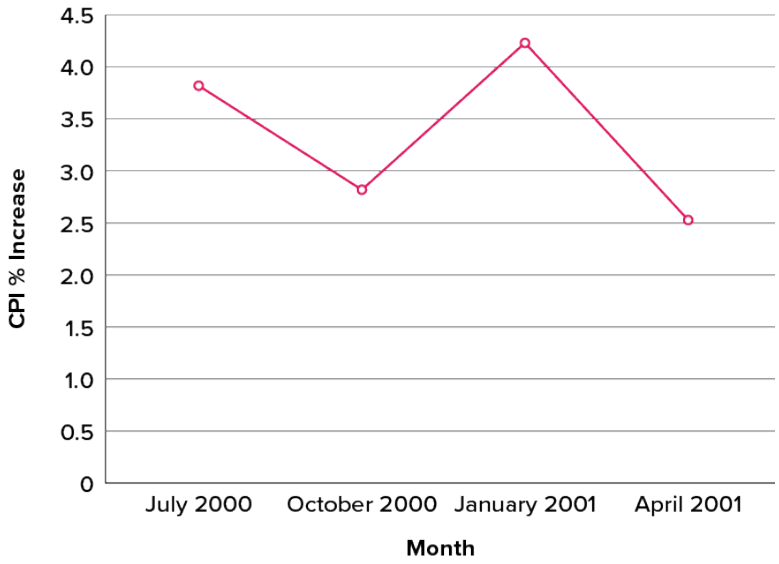


Figure 8.21: **Figure 8.19** A line graph of the percent change in the CPI over time. Each point represents percent increase for the three months ending at the date indicated[20]

Line graphs are appropriate only when both the x– and y-axes display continuous, ordered (rather than categorical) variables. Although bar charts can also be used in this situation, line graphs are generally better at comparing changes over time. Figure 8.20, for example, shows percent increases and decreases in five components of the CPI. The figure makes it easy to see that medical costs had a steadier progression than the other components. Although you could create an analogous bar chart, its interpretation would not be as easy.

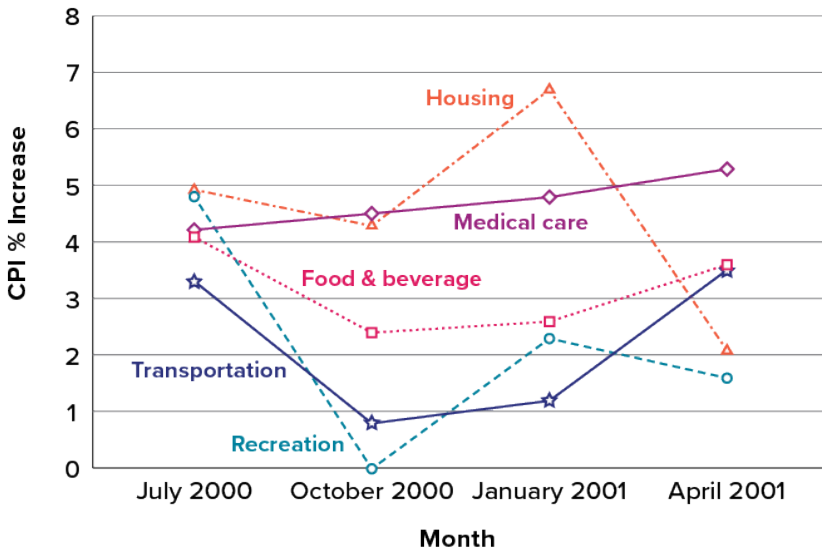


Figure 8.22: **Figure 8.20** A line graph of the percent change in five components of the CPI over time[21]

Let us stress again that it is misleading to use a line graph when the x-axis contains merely categorical variables. As we saw earlier in this chapter, Figure 8.9, inappropriately shows a line graph of the card game data from Yahoo, discussed in the section on qualitative variables. The defect in Figure 8.9 is that it gives the false impression that the games are naturally ordered in a numerical way, whereas Figure 8.20 is appropriate because we are looking at differences across time.

8.2.5 Histograms

A ***histogram*** is a graphical method for displaying the shape of a distribution of continuous scores (interval or ratio data). It is particularly useful when there are a large number of observations. We begin with an example consisting of the scores of 642 students on a psychology test. The test consists of 197 items, each graded as "correct" or "incorrect." The students' scores, therefore, ranged from 46 to 167.

The first step is to create a frequency table. Unfortunately, a simple frequency table would be too big, containing over 100 rows. To simplify the table, we group scores together as shown in Table 8.2, called a ***grouped frequency table***.

Table 8.2: **Table 8.2** Grouped Frequency Distribution of Psychology Test Scores[22]

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

To create this table, the range of scores was broken into bins or intervals, called ***class intervals***. The first interval is from 39.5 to 49.5, the second from 49.5 to 59.5, etc. Next, the number of scores falling into each interval was counted to obtain the class frequencies. There are 3 scores in the first interval, 10 in the second, etc.

Class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too "choppy." More information on choosing the widths of class intervals is presented later in this section. Placing the limits of the class intervals midway between two numbers (e.g., 49.5) ensures that every score will fall in an interval rather than on the boundary between intervals.

In a histogram, the class frequencies are represented by bars. The height of each bar corresponds to its class frequency. Although a histogram uses bars, it is not called a bar chart because it shows continuous rather than categorical data, so the bars are connected, rather than distinct groups. A histogram of these psychology test score data is shown in Figure 8.21.

The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes. You can also see that the distribution is not symmetric: the scores extend farther to the right than they do to the left. The distribution is therefore said to be ***skewed***. (We'll have more to say about shapes of distributions below.)

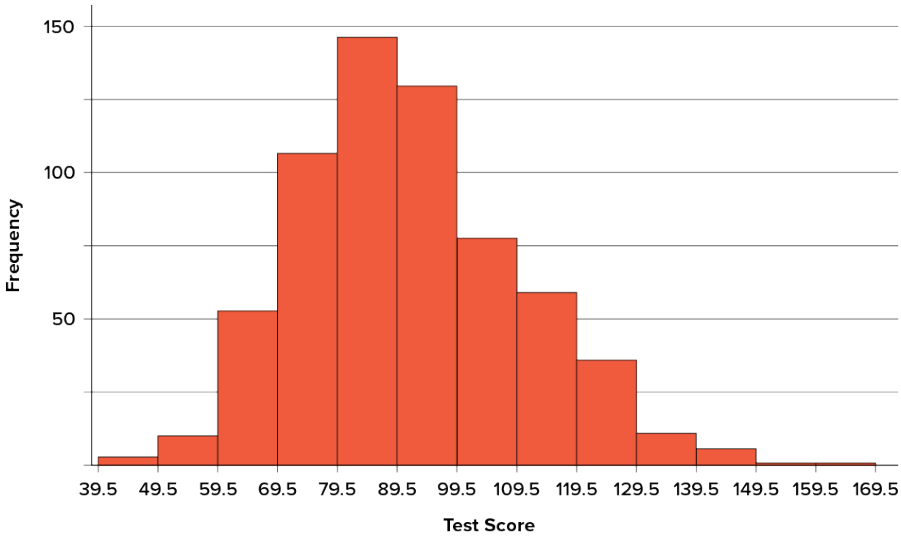


Figure 8.23: **Figure 8.21** Histogram of scores on a psychology test[23]

In our example, the observations are whole numbers. Histograms can also be used when the scores are measured on a more continuous scale with decimals such as the length of time (in milliseconds) required to perform a task. In this case, there is no need to worry about fence sitters since they are improbable. (It would be quite a coincidence for a task to require exactly 7 seconds, measured to the nearest thousandth of a second.) We are therefore free to choose whole numbers as boundaries for our class intervals, for example, 4000, 5000, etc. The class frequency is then the number of observations that are greater than or equal to the lower bound, and strictly less than the upper bound. For example, one interval might hold times from 4000 to 4999 milliseconds. Using whole numbers as boundaries avoids a cluttered appearance and is the practice of many computer programs that create histograms. Note also that some computer programs label the middle of each interval rather than the end points.

Histograms can be based on relative frequencies instead of actual frequencies. Histograms based on ***relative frequencies*** show the proportion of scores in each interval rather than the number of scores. In this case, the y-axis runs from 0 to 1 (or somewhere in between if there are no extreme proportions). You can change a histogram based on frequencies to one based on relative frequencies by (a) dividing each class frequency by the total number of observations, and then (b) plotting the quotients on the y-axis (labeled as proportion).

There is more to be said about the widths of the class intervals, sometimes called ***bin widths***. Your choice of bin width determines the number of class intervals. This decision, along with the choice of starting point for the first interval, affects the shape of the histogram. The best advice is to experiment with different choices of width, and to choose a histogram according to how well it communicates the shape of the distribution.

Please watch: Plots, Outliers, and Justin Timberlake: Data Visualization Part 2: Crash Course Statistics #6

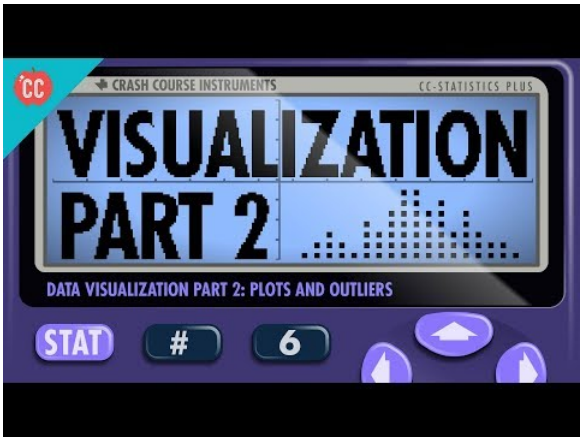


Figure 8.24: Plots, Outliers, and Justin Timberlake: Data Visualization Part 2: Crash Course Statistics #6

8.2.6 Shapes of Continuous Distributions

Finally, it is useful to discuss how we describe the shapes of distributions, which we will revisit in future chapters to learn how different shapes affect our numerical descriptors of data and distributions.

The primary characteristic we are concerned about when assessing the shape of a distribution is whether the distribution is symmetrical or skewed. A ***symmetrical distribution***, as the name suggests, when cut down the center forms two mirror images. Consider the two sides of butterfly wings, which look almost identical. Although in practice we will never get a perfectly symmetrical distribution, we would like our data to be as close to symmetrical as possible, otherwise known as ***approximately normal***, for reasons we delve into in future sections. Many types of distributions are symmetrical, but by far the most common and pertinent distribution at this point is the normal distribution, shown in Figure 8.22. Notice that although the symmetry is not absolutely perfect (for instance, the bar just to the right of the center is taller than the one just to the left), the two sides are roughly the same shape. The normal distribution has a single peak, known as the center, and two ends that extend out equally, forming what is known as a bell shape or ***bell curve***.

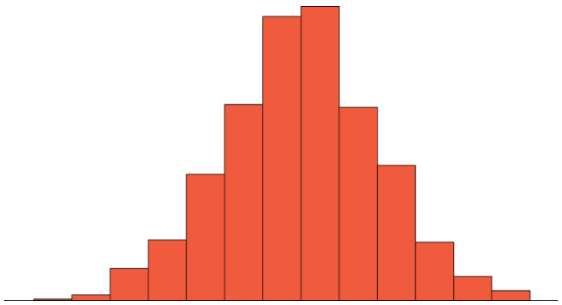


Figure 8.25: **Figure 8.22** A symmetrical distribution[24]

Symmetrical distributions can also have multiple peaks. Figure 8.23 shows a ***bimodal distribution***, named for the two peaks that lie roughly symmetrically on either side of the center point (think of a Bactrian camel, with two humps rather than one). As we will see in future chapters, this is not a particularly desirable characteristic of our data, and, worse, this is a relatively difficult characteristic to detect numerically. Thus, it is important to visualize your data by looking at the distribution before moving ahead with any formal ***data analyses***, which means summarizing and answering questions with your data.

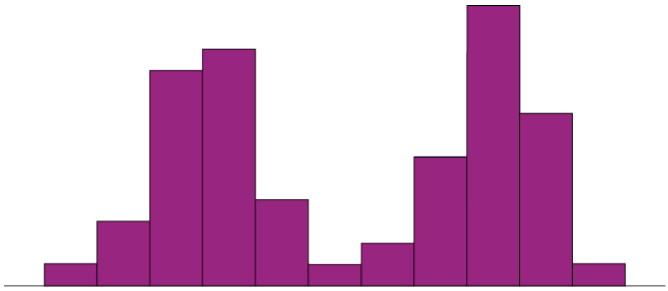


Figure 8.26: **Figure 8.23** A bimodal distribution[25]

8.2.7 Skew and Kurtosis

There are two more descriptive statistics that are sometimes reported in the psychological literature: skew and kurtosis. In practice, neither one is used anywhere near as frequently as the measures of central tendency and variability. Distributions that are not symmetrical also come in many forms, more than can be described here. The most common asymmetry to be encountered is referred to as ***skew***, in which one of the two ***tails*** (ends) of the distribution is disproportionately longer than the other. This property can affect the value of the averages we use in our analyses and make them an inaccurate representation of our data, which causes many problems. Skew can either be ***positive*** (right skew, to the right) or ***negative*** (left skew, to the left), based on which tail is longer. It is very easy to get the two confused at first; many students want to describe the skew by where the bulk of the data (larger portion of the

histogram, known as the ***body***) is placed, but the correct determination is based on which tail is longer. You can think of the tail as an arrow or ski slope; whichever direction the arrow or slope is pointing is the direction of the skew. Figure 8.24 shows positive (right) and negative (left) skew, respectively.

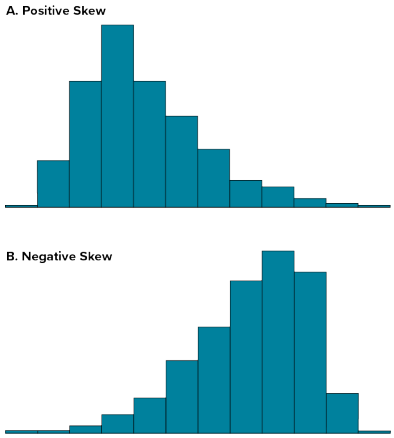


Figure 8.27: **Figure 8.24** Positively skewed (A) and negatively skewed (B) distributions[26]

As another example, Figure 8.25 illustrates that if the data tend to have a lot of extreme small values (i.e., the lower tail is "longer" than the upper tail) and not so many extremely large values (left panel), the data are negatively skewed, with a tail stretching to the left. On the other hand, if there are more extremely large values than extremely small ones (right panel), the data are positively skewed, with a tail stretching to the right. A symmetric distribution has a skewness of 0. The skewness value for a positively skewed distribution is positive, and a negative value for a negatively skewed distribution.

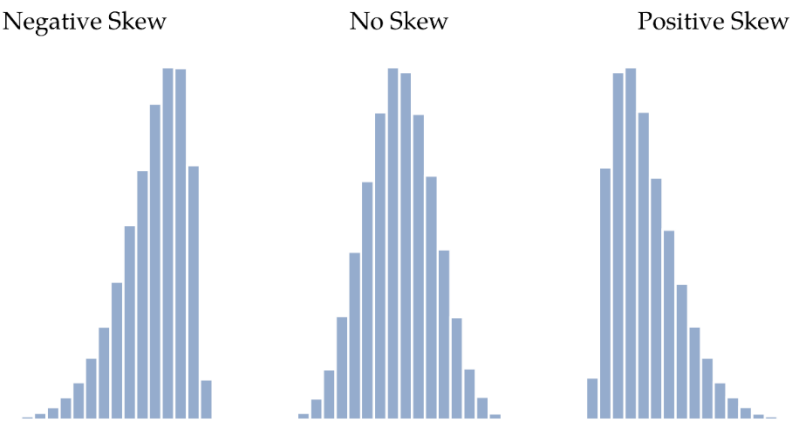


Figure 8.28: **Figure 8.25** Distributions with negative skew (left), no skew (middle), and positive skew (right)[27]

The final measure we will review in distributions is ***kurtosis*** of a data set. Put simply, kurtosis is a measure of how thin or thick the ***tails*** (ends) of a distribution are, as illustrated in Figure 8.26.

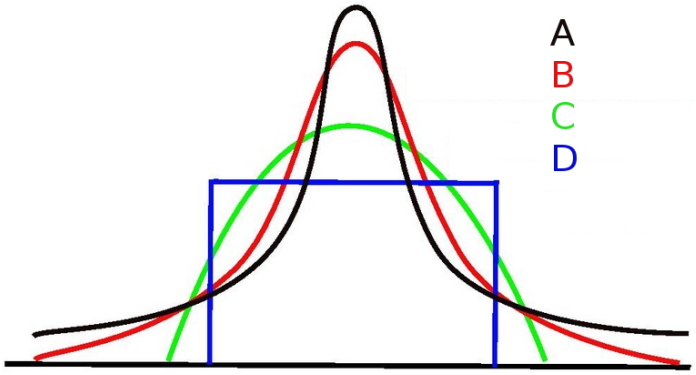


Figure 8.29: **Figure 8.26** Different shapes of distributions including leptokurtic (A), mesokurtic (B), platykurtic (C) and uniform[28]

The first distribution (A, in black) has fat tails with more extreme scores, so the kurtosis is positive, and we say that the data is ***leptokurtic***. Often, these distributions also appear to

have a tall peak, so you might notice they look like they are leaping up. The second distribution (B, in red) is ***mesokurtic***, with a kurtosis value almost exactly 0 and tails are neither thin nor fat. Think of a mesa in the desert or mesa in Spanish, which means "table". The third distribution (C, in green) is ***platykurtic***, meaning that the distribution is fairly evenly spread out and has thin tails (very few extreme scores), so the kurtosis value is negative. Think of a plate or a platypus tail – very flat. Finally, the fourth distribution (D, in blue) is uniform. This is a special kind of platykurtic distribution that is really evenly spread out, with no tails. This is summarized in Table 8.3.

Table 8.3: **Table 8.3** Thin to Fat Tails to Illustrate Kurtosis[29]

Informal Term	**Description**	**Mathematical Kurtosis Value**
platykurtic	tails too thin, few outliers	negative
mesokurtic	tails neither thin nor fat	zero
leptokurtic	tails too fat	positive
uniform	Spread out, no tails	close to zero

Please watch: Randomness: Crash Course Statistics #17 (which covers skew and kurtosis)

Figure 8.30: Randomness: Crash Course Statistics #17

8.3 Summary

Hopefully, this section has helped you see the importance of tables and graphs for visualizing what is happening in a sample of data. It is important to be able to 'read' tables and graphs to test claims that others are making about what is happening or what is good for us, and it is good for scientists to use table and graph representations to communicate their findings with their audience swiftly and effectively.

8.4 Additional Video Helps & Resources

For tips and demonstrations on setting up and running visual data analyses in SPSS, please see Virginia Wickline: SPSS Statistics Helps (publishing dates vary).

8.5 Media Attributions

[1] (Posted by Terrence Ong via Wikimedia Commons licensed under CC BY -SA 3.0.)

[2] (Public domain in the U.S., without copyright notice. Posted by Fma 12 via Wikimedia Commons.)

[3] (CC BY-SA 4.0 by Cote et al., 2021)

[4] ("Mac Pie Chart" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[5] ("Mac Bar Chart" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[6] ("Card Game Bar Chart" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[7] ("Mac Bar Chart 3D" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[8] ("Mac Bar Chart Lie Factor" by Judy Schmitt is licensed under CC BY-NC-SA 4.0. "Apple iMac G3 (1998)" by albaco/Flickr is licensed under CC BY-NC-SA 2.0; image was brightened and background was removed.)

[9] ("Mac Bar Chart Baseline 50" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[10] Figure 8.9. ("Line Chart Inappropriately Used" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[11] ("Touchdown Passes Raw Data" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[12] ("Touchdown Passes Stem and Leaf" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[13] ("Women's Times Raw Data" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[14] ("Box Plot First Step" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[15] ("Box Plot Whiskers" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[16] ("Box Plot Outside Value" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[17] ("Percent Increase in Stock Indexes" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[18] ("Percent Change in CPI" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[19] ("Means of Two Conditions" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[20] ("Percent Change in CPI Line Graph" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[21] ("Percent Change in CPI x5 Line Graph" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[22] (CC BY-SA 4.0 by Cote et al., 2021)

[23] ("Psychology Test Scores Histogram" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[24] ("Symmetrical Distribution" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[25] ("Bimodal Distribution" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[26] ("Skewed Distributions" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[27] (CC BY-SA 4.0 by Navarro & Foxcroft, 2025)

[28] (CC BY 3.0 by Hunsvotti via Wikimedia Commons)

[29] (CC BY-SA 4.0 by Navarro & Foxcroft, 2025)

8.6 Text Attributions

Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). *Introduction to statistics in the psychological sciences.* Pressbooks. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Navarro, D. J., & Foxcroft, D. R. (2025). *Learning statistics with jamovi: A tutorial for beginners in statistical analysis.* Open Book Publishers. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

8.7 References

CrashCourse (2018, May 23). *Randomness: Crash course statistics #17*. https://www.youtube.com/watch?v=jL9en6NvQfk

CrashCourse (2018, February 21). *Charts are like pasta - data visualization part 1: Crash course statistics #5*. https://www.youtube.com/watch?v=hEWY6kkBdpo

CrashCourse (2018, February 28). Plots, outliers, and Justin Timberlake: Data visualization part 2: Crash course statistics #6. https://www.youtube.com/watch?v=HMkllhBI91Y

Gaslowitz, L. (2017, July 6). *How to spot a misleading graph – Lea Gaslowitz*. TED-Ed. https://www.youtube.com/watch?v=E91bGT9BjYk

Tufte, E. R. (1983). *The visual display of quantitative information*. Graphics Press.

9 Chapter 9: Central Tendency and Variability

9.1 Measures of Central Tendency

In most situations, the first thing that a researcher will want to calculate for a sample of data is a measure of *central tendency*. That is, they would like to know something about what the most typical response is. The three most commonly used measures of central tendency are the mean, median, and mode.

The ***mean*** of a set of observations is another way of saying the average. Add all of the values up, and then divide by the total number of values. If the number of books that a group of five kindergartners read last month were 56, 31, 56, 8, and 32, the mean of these observations would be: (56+31+56+8+32) / 5 = 183 / 5 = 36.60. Of course, this definition of the mean probably is not new to you. Averages (i.e., means) are used so often in everyday life that this is pretty familiar stuff.

Figure 9.1: **Figure 9.1** Row of mathematics books[1]

The second measure of central tendency that is often used is the ***median***. The median of a set of observations is just the middle value, like the median on the highway separates the two lanes of traffic. As before, let us imagine how many books a group of kindergarteners read: 56, 31, 56, 8 and 32. To figure out the median, sort these numbers into ascending order: 8, 31, **32,** 56, 56. Once the numbers are ordered, it is easy to see that the median value of these five observations is 32 since that is the middle score in the sorted list (we have put it in bold to make it even more obvious). Easy enough, right? However, what should we do if we are interested in six kindergarteners rather than five? If a sixth kindergartener read 14 books, our

sorted list is now 8, 14, **31, 32**, 56, 56. There are two middle numbers, 31 and 32. The median is defined as the average of those two numbers, which is 31.5. It would be very tedious to do this by hand if you had lots of numbers.

Figure 9.2: **Figure 9.2** Highway with a median that divides the road in half[2]

The ***mode*** of a sample is very simple. It is the value that occurs most frequently. Things that are fashionable are "in the mode" and pie with ice cream (also very popular) is called pie á la mode. One last point to make regarding the mode. While the mode is most often calculated when data are nominal, because means and medians are useless for those sorts of variables, there are some situations in which one really does want to know the mode of an ordinal, interval, or ratio scale variable. Consider this scenario: A friend of yours is offering a bet and they pick a football game at random. Without knowing who is playing, you have to guess the exact winning margin. If you guess correctly, you win $50. If you do not, you lose $1. There are no consolation prizes for "almost" getting the right answer. You have to guess exactly the right margin. For this bet, the mean and the median are completely useless to you. It is the mode that you should bet on. Or, what if you were asking how old people are in your statistics classes and there were two most popular answers, 19 and 22? In that case, having a bimodal distribution would be another good reason to focus on the mode(s) for interval or ratio data.

Figure 9.3: Apple pie à la mode - so popular!

Figure 9.4: **Figure 9.3** Apple pie à la mode - so yummy and popular![3]

Figure 9.4 Tokyo Fashion Week 2010 collections. The mode is so popular, everyone will want it[4]

Please watch: Math Antics – Mean, Median and Mode.

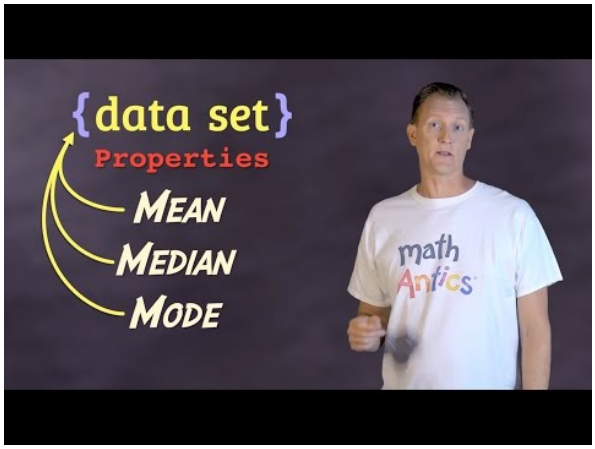


Figure 9.5: Math Antics - Mean, Median and Mode

9.2 Measures of Variability

The statistics that we have discussed so far all relate to central tendency. That is, they all talk about which value(s) are most typical or common in the data. However, central tendency is not the only type of summary statistic that we want to calculate. The second thing that we really want is a measure of the ***variability*** of the data, that is, how spread out the data are. Metaphorically, do we have a small puddle or a large puddle coming out from the center? In other words, whatever we have chosen for our measure of central tendency (mean, median, or mode), we also want to know how far away from this the observed values tend to be.

The ***range*** for a given variable is very simple. It is the biggest value minus the smallest value. Although the range is the simplest way to quantify the notion of variability, it is also one of the worst because it is the least precise. If the data set has some extreme (possibly untrustworthy) values in it, we would like our summary statistics to not be unduly influenced by these cases. For example, let us say we asked nine people how many cookies they ate yesterday and we got these responses: 0, 1, 2, 3, 4, 5, 6, 10, 100. In a variable containing an extreme outlier, it is clear that the range is not very robust. This variable has a range of 100, but if the outlier were removed, it would have a range of only 10. One has to use the range if they have selected the mode for central tendency, but for continuous data, we probably want a more representative answer.

The ***interquartile range (IQR)*** is like the range, but instead of the difference between the biggest and smallest value, the difference between the 25th percentile and the 75th percentile is calculated. If you do not already know what a ***percentile*** is, the 10th percentile of a data set is the given number (let's say, x) such that 10% of the data is less than x. In fact, we have already come across this idea indirectly. The median of a data set is the value that is at the 50th percentile!

While it is clear how to interpret the range, how to interpret the IQR is a little less obvious. The simplest way to think about it is like this: The interquartile range is the range spanned by the middle half or middle 50% of the scores. That is, one quarter of the data falls below the 25th percentile and one quarter of the data is above the 75th percentile, leaving the middle half of the data lying in between the two, so the IQR is the range covered by that middle half. If one has opted for the median as the best measure of central tendency, one would select the IQR as the best measure of variability.

Another kind of variability that we would use with the mean is called the ***variance***, which is the average squared deviation (distance) of each score from the mean. The variance of a data set 𝑋 is sometimes written as Var(𝑋), but it is more commonly denoted 𝑋2 (the reason for this will become clearer shortly).

How about a concrete example? Say our friend, Jamisa, was a basketball player, and we looked at how many points she scored in her last five games. If we calculate the mean, she is averaging 36.6 points per game (great job, Jamisa!). We would end up with the information shown in Table 9.1.

Table 9.1: **Table 9.1** Measures of Variability for Jamisa's Last Five Basketball Games[5]

English	Math	Value	Deviation from Mean	Squared Deviation
notation	i	X_i	$X_i - \overline{X}$	$(X_i - \overline{X})^2$
	1	56	19.4	376.36
	2	31	-5.6	31.36
	3	56	19.4	376.36
	4	8	-28.6	817.96

English	Math	Value	Deviation from Mean	Squared Deviation
	5	32	-4.6	21.16

That last column contains all of our squared deviations, so all we have to do is average them, dividing by the sample size (N). If we do that by hand, using a calculator, we end up with a variance of 324.64.

If we were using this to estimate a population value, rather than just describing the sample, we would divide by (N-1). There is a subtle distinction between describing a sample and making guesses about the population from which the sample came. Regardless of whether one is describing a sample or drawing inferences about the population, the mean is calculated exactly the same way. Not so for the variance, or the standard deviation, or for many other measures. What was outlined to you initially (i.e., take the actual average, and thus divide by) assumes that you literally intend to calculate the variance of the sample. Most of the time, however, researchers are not terribly interested in the sample in and of itself. Rather, the sample exists to tell us something about the world. If so, you are actually starting to move away from calculating a sample ***statistic*** (gathered from the data you collected) and towards the idea of estimating a population ***parameter*** (an estimate for the population at large).

This section so far may have read a bit like a mystery novel. We have calculated the variance and described the " −1" thing to estimate a population, but we need to review the single most important thing: how to interpret the variance. Descriptive statistics are supposed to describe things, after all, and right now the variance might seem like a gibberish number. Unfortunately, the reason why we have not given you the human-friendly interpretation of the variance is that there really is not one. This is the most serious problem with the variance. Although it has some elegant mathematical properties that suggest that it really is a fundamental quantity for expressing variation, it is completely useless if you want to communicate with an actual human. Variances are completely uninterpretable in terms of the original variable! All the numbers have been squared, so they do not mean anything practical anymore. This is a huge issue. For instance, according to Table 9.1, the margin in game 1 was "376.36 points-squared higher than the average margin." This is exactly as goofy as it sounds, and so when we calculate a variance of 324.64, we are in the same situation. If you have watched a good bit of basketball, you will notice at no time has anyone ever referred to "points squared." It is not a real unit of measurement.

Okay, suppose that you like the idea of using the variance, but since you are a human and not a robot or math fanatic, you would like to have a measure that is expressed in the same units as the data itself (i.e., points scored, not points squared). What should you do? The solution to the problem is to take the square root of the variance, known as the ***standard deviation***, also called the root mean squared deviation, or RMSD. This solves our problem fairly neatly. While nobody has a clue what "a variance of 324.64 points-squared" really means, it is much easier to understand "a standard deviation of 18.01 points" since it is expressed in

the original units. It is traditional to refer to the standard deviation of a sample of data as s, though "sd" and "std dev." are also used at times, and you might see the Greek letter sigma (s) if we are talking populations. If you are using a statistical program to analyze your data, it almost always calculates a version that divides by $n - 1$ rather than n.

Interpreting standard deviation is slightly more complex. Because the standard deviation is derived from the variance, and the variance is a quantity that has little to no meaning that makes practical sense, the standard deviation does not have a simple interpretation. As a consequence, most of us just rely on a simple rule of thumb. In general, if the data are normally distributed, one should expect 68% of the data to fall within 1 standard deviation of the mean, 95% of the data to fall within 2 standard deviations of the mean, and 99.7% of the data to fall within 3 standard deviations of the mean. This rule tends to work pretty well most of the time, but it is not exact. It is actually calculated based on an assumption that the histogram is symmetric and "bell shaped".

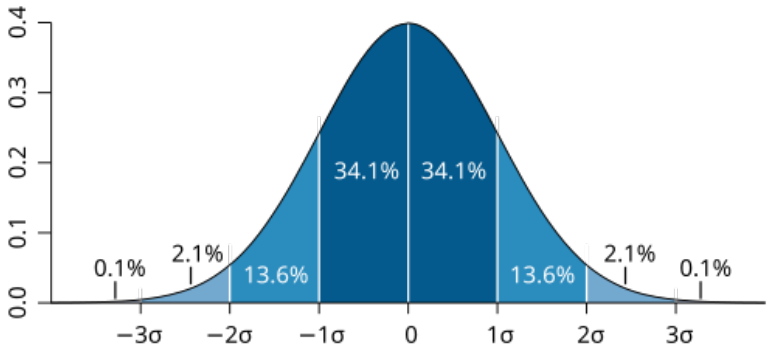


Figure 9.6: **Figure 9.5** A symmetrical bell curve showing that 68.2% of scores fall within one standard deviation (SD), and 95.4% within 2 SDs, and 99.6% within 3 SDs[6]

Here is a quick summary about variability:

- *Range*. Describes the full spread of the data. It is very vulnerable to outliers, and as a consequence, it is not often used unless one is complementing the mode for nominal data or there are good reasons to care about the extremes in the data.
- *Interquartile range*. Indicates where the middle half of the data sits. It is pretty robust and complements the median nicely. This is used a lot.
- *Variance*. The average squared deviation from the mean. It is mathematically elegant and is probably the "right" way to describe variation around the mean, but it is challenging because it does not use the same units as the data. Almost never used except as a mathematical tool, when it is buried within other statistical calculations.

- *Standard deviation.* This is the square root of the variance. It is fairly elegant mathematically, and it is expressed in the same units as the data so it can be understood and interpreted pretty well. In situations where the mean is the measure of central tendency, this is the default. This is by far the most popular measure of variation. In short, the IQR and the standard deviation are easily the two most common measures used to report the variability of the data.

Please watch: Measures of Variability (Range, Standard Deviation, Variance).

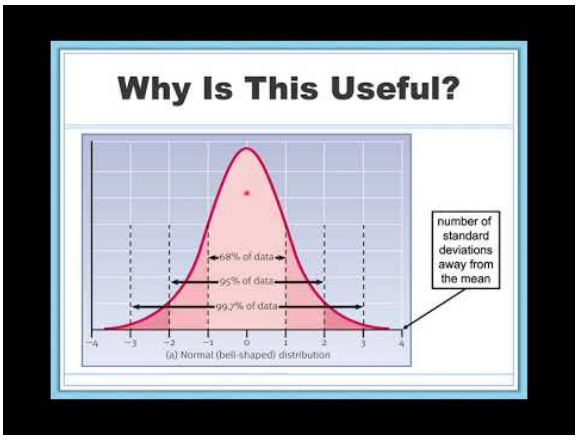


Figure 9.7: Measures of Variability (Range, Standard Deviation, Variance)

9.3 Mean, Median, or Mode – Which to Choose?

Knowing how to calculate mean, median, and mode is only a part of the story. One also needs to understand what each one is saying about the data, and what that implies for when each one should be used. This is illustrated in Figure 9.6.

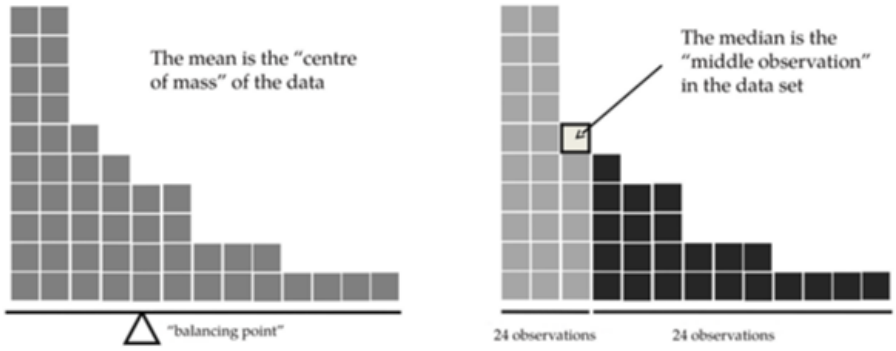


Figure 9.8: **Figure 9.6** An illustration of the difference of how the mean and the median should be interpreted [7]

The mean is basically the 'center of gravity' of the data set. If you imagine that the histogram of the data is a solid object, then the point on which you could balance it (as if on a see-saw or old-fashioned scale) is the mean. In contrast, the median is the middle observation, with half of the scores smaller than it and half of the observations larger than it.

The mean is the balance point of the data set, whereas the median is the dividing point in the data, and the mode is the most popular score or scores. What this implies, as far as which one should be used, depends a little on what type of data one has and what they are trying to achieve. Consider your best measure of central tendency and variability as dancing partners: The mode always goes with the range, the median always goes with the IQR, and the mean always goes with the standard deviation. Do not try to make them dance with other partners!

Figure 9.9: **Figure 9.7** Like dancing partners, the mode always goes with the range, the median always goes with the IQR, and the mean always goes with the standard deviation[8]

As a rough guide:

- If the data are on a *nominal* scale, statisticians say the mean and the median cannot or should not be used. Both the mean and the median rely on the idea that the numbers assigned to values are meaningful and have order. If the numbering scheme is arbitrary like it is for nominal (categorical) data, then the mode must be used. As mentioned previously, the mode would also be the best choice if the data are bimodal.
- If the data are on an *ordinal* scale, use the median instead of the mean. The median only makes use of the order information in the data (i.e., which numbers are bigger) but does not depend on the precise numbers involved. That is exactly the situation that applies when data are on an ordinal scale. The mean, on the other hand, makes use of the precise numeric values assigned to the observations, so it is not really appropriate for ordinal data.
- For data on an *interval and ratio* scale, either the median or mean is generally acceptable. Which one you pick depends a bit on what you are trying to achieve. The mean has the advantage that it uses all the information in the data (which is useful when not much data exists). However, the mean is very sensitive to extreme, outlying values called ***outliers***. Extreme scores are two or more standard deviations from the mean, and outliers are three or more standard deviations. Think of outliers as outlaws in an old Western movie: They kind of do their own thing, and they often cause chaos in a group of data.

Figure 9.10: **Figure 9.8** A mural of an outlaw on his horse. Watch out for outliers in your data set – they can wreak havoc[9]

Expanding on that last part a little, one consequence is that there are systematic differences between the mean and the median or mode when the histogram is asymmetric (has abnormal skew and kurtosis, as discussed in a previous chapter). This is illustrated in Figure 9.6 above. Notice that the median (right-hand side) is located closer to the body of the histogram, whereas the mean (left-hand side) gets dragged towards the tail (where the extreme values are).

To give a concrete example, suppose Bob (income $50,000), Li (income $60,000) and Keisha (income $65,000) are sitting at a table. The average income at the table is $58,333 and the median income is $60,000. Then José sits down with them (income $100,000,000). The average income has now jumped to $25,043,750, but the median rises only to $62,500. If you are interested in looking at the overall income for the people at the table, the mean might be the right answer. However, if you are really interested in the *typical* income at the table, the median would be a much better, more realistic, and (some would say) more truthful choice.

Another demonstration is illustrated in Figure 9.9. If a data set is normally distributed, the mean, median, and mode will all be at (or relatively at) the same place in the distribution, so it does not matter which one you pick (although we would go with the mean if we can). When a data set is positively skewed by a large outlier or outliers to the right (like the high-income example above), the mean is pulled toward those extreme scores, making the mean higher than the median and mode, making the median the most reliable indicator of what is typical for the dataset. If the data set is negatively skewed because there is an outlier or outliers in the low end of the distribution (a very low score), the mean will be lower in value than the median and mode, again, making the median the most reliable indicator of what is typical for the group.

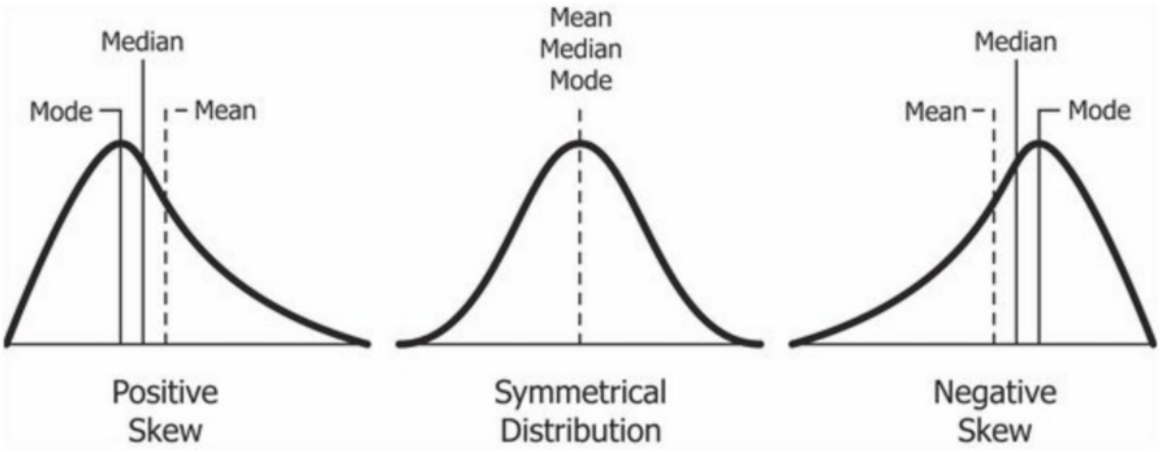


Figure 9.11: Diagram showing three types of data distributions: positive skew, symmetrical distribution, and negative skew. Each distribution curve includes labeled vertical lines for mode, median, and mean, illustrating their relative positions and differences in skewed versus symmetrical data.

Figure 9.9 How the mean, median, and mode line up on skewed and normal distributions[10]

Please watch: Why Averages Lie To You.

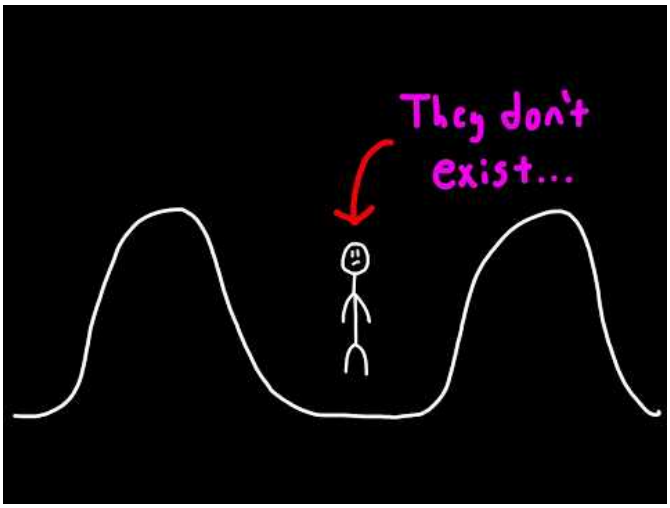


Figure 9.12: Why Averages Lie To You

9.3.1 A Metaphor and a Real-Life Example

When the chapter author's daughter was in elementary school, the highlight for the end-of-year celebration was field day, with the very last event being a tug-of-war competition for each grade. Being in a very small school, they kept children of like abilities together, so my daughter always had the same classmates. In her class was a child they called "Big Mike" (not his real name). Big Mike was four feet tall at age 8 – no joke – over a foot taller than anyone else in his class and matching height with his teachers by fourth grade. Guess who won the tug-of-war every year? My daughter's class! Well, let's be honest…Big Mike. In this case, Big Mike was like the outlier, pulling the flag from the middle to his team's side every time. This is just what an outlier does to the mean – it pulls it off center in the distribution. Following the metaphor, Big Mike masked the strength of typical children because he overpowered them. An outlier does the same thing to the mean.

Figure 9.13: **Figure 9.10** Junior high students in China play tug-of-war. Like an outlier, someone super strong will pull the flag (the mean) from the middle to their side to win the game[11]

To try to get an even better sense of why you need to pay attention to the differences between the mean and the median, consider also this real-life example. This is an article by Michael Janda from the ABC news website from 24 September, 2010:

> Senior Commonwealth Bank executives have travelled the world in the past couple of weeks with a presentation showing how Australian house prices, and the key price to income ratios, compare favourably with similar countries. "Housing affordability has actually been going sideways for the last five to six years," said Craig James, the chief economist of the bank's trading arm, CommSec.

This probably comes as a huge surprise to anyone with a mortgage, or who wants a mortgage, or pays rent, or is not completely oblivious to what has been going on in the housing market over recent decades. Back to the article:

> CBA has waged its war against what it believes are housing doomsayers with graphs, numbers and international comparisons. In its presentation, the bank rejects arguments that Australia's housing is relatively expensive compared to incomes. It says Australia's house price to household income ratio of 5.6 in the major cities, and 4.3 nationwide, is comparable to many other developed nations. It says San Francisco and New York have ratios of 7, Auckland's is 6.7, and Vancouver comes in at 9.3.

More excellent news! Except, the article goes on to make the observation that:

> Many analysts say that has led the bank to use misleading figures and comparisons. If you go to page four of CBA's presentation and read the source information at

> the bottom of the graph and table, you would notice there is an additional source on the international comparison– Demographia. However, if the Commonwealth Bank had also used Demographia's analysis of Australia's house price to income ratio, it would have come up with a figure closer to 9 rather than 5.6 or 4.3.

That is a rather serious discrepancy. One group of people say 9, another says 4-5. Should we just split the difference and say the truth lies somewhere in between? Absolutely not! This is a situation where there is a right answer and a wrong answer. Demographia is correct, and the Commonwealth Bank is wrong. As the article points out:

> [An] obvious problem with the Commonwealth Bank's domestic price to income figures is they compare average incomes with median house prices (unlike the Demographia figures that compare median incomes to median prices). The median is the mid-point, effectively cutting out the highs and lows, and that means the average is generally higher when it comes to incomes and asset prices, because it includes the earnings of Australia's wealthiest people. To put it another way: the Commonwealth Bank's figures count Ralph Norris' multi-million dollar pay packet on the income side, but not his (no doubt) very expensive house in the property price figures, thus understating the house price to income ratio for middle-income Australians.

Notice: The way that Demographia calculated the ratio is correct. The way that the Bank did it is incorrect. As for why an extremely quantitatively sophisticated organization such as a major bank made such an elementary mistake, well…we cannot say for sure. However, the article itself does happen to mention the following facts, which may or may not be relevant:

> [As] Australia's largest home lender, the Commonwealth Bank has one of the biggest vested interests in house prices rising. It effectively owns a massive swathe of Australian housing as security for its home loans as well as many small business loans.

Sometimes, it is good to be skeptical of the source of information, as organizations can have agendas, just like individuals.

9.4 Summary

This section has helped illustrate why both what is typical (central tendency) and how spread out the data are (variability) are important in understanding what is happening in a data sample. Each measure of central tendency has its variability partner when reporting data. One must be careful in considering which measure of central tendency and variability are best to report, considering scale of measurement and shape of the distribution when making their choice; the average is not always best just because it is the average. In fact, sometimes the average can be misleading.

9.5 Additional Video Helps & Resources

For tips and demonstrations on setting up and running descriptive statistical analyses in SPSS, please see Virginia Wickline: SPSS Statistics Helps (publishing dates vary).

9.6 Media Attributions

[1] (CC BY-SA 4.0 by AKibombo via Wikimedia Commons)

[2] (CC0 1.0 Universal by Vknyshov via Wikimedia Commons)

[3] (CC BY 3.0 by Dwight Burdette via Wikimedia Commons)

[4] (CC BY-SA 2.0 by Tokyographer via Wikimedia Commons)

[5] (CC BY-SA4.0 by Navarro & Foxcroft, modified by current authors.)

[6] (CC BY 2.5 by Jeremy Kemp via Wikimedia Commons)

[7] (CC BY-SA 4.0 by Navarro & Foxcroft, 2025)

[8] (CC BY-SA4.0 by Todd Combs via Wikimedia Commons)

[9] (CC BY-SA 3.0 by Franco Baresi via Wikimedia Commons)

[10] (CC BY-SA4.0 by Diva Jain via Wikimedia Commons)

[11] (CC BY-SA 2.0 by ▯▯▯▯ ▯▯▯ (WC-QHS) via Wikimedia Commons)

9.7 Text Attributions

Navarro, D. J., & Foxcroft, D. R. (2025). *Learning statistics with jamovi: A tutorial for beginners in statistical analysis.* Open Book Publishers. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

9.8 References (including Video Attributions)

Foundations of Math (2026, January 14). Why averages lie to you. https://www.youtube.com/watch?v=8LXynB2mPls

MathAntics (2017, March 3). *Math antics – Mean, median and mode.* https://www.youtube.com/watch?v=B1HEzNTGeZ4

Storage, D. (2019, June 18). *Measures of variability (range, standard deviation, variance).* https://www.youtube.com/watch?v=s7WTQ0H0Acc

10 Chapter 10: Cronbach's alpha (Reliability Analysis)

10.1 Cronbach's alpha (a): A Measure of Internal Consistency

If you are using a survey measure in your study, you will want to see how well the factors are measured using a scale that combines the observed variables. In statistics, we use reliability analysis to provide information about how consistently a scale measures a psychological ***construct*** (see Module 2), which is a variable we are interested in studying. Many kinds of ***reliability*** exist to make sure a measure is consistent. ***Internal consistency*** is what we are concerned with here, which refers to the consistency across all the individual items that make up a measurement scale. In other words, we want to make sure that all the variables play nice together, that they all correlate with one another. We can calculate a statistic that tells us how internally consistent these items are in measuring the underlying construct.

One popular statistic used to check the internal consistency of a scale is ***Cronbach's alpha*** (Cronbach, 1951). Cronbach's alpha is a measure of equivalence (whether different sets of scale items would give the same measurement outcomes). Equivalence is tested by dividing the scale items into two groups (a "split-half") and seeing whether analysis of the two parts gives comparable results. Of course, there are many ways a set of items could be split, but if all possible splits are made then it is possible to produce a statistic that reflects the overall pattern of split-half coefficients. Cronbach's alpha (a) is such a statistic: a function of all the split-half coefficients for a scale. If a set of items that measure a construct (e.g., an Extraversion scale) has an a = .80, then the proportion of error variance in the scale is 0.20. In other words, a scale with an a = .80 includes approximately 20% error.

10.2 Cronbach's a: Let's Cook!

Why do we need a measure to be internally consistent? Let's use the metaphor of cooking. People cook so they can eat. Some people really like to cook. Almost everyone really likes to eat. Many cultures around the world enjoy eating some form of chicken soup. Take a second and imagine a big steaming cup of chicken noodle soup like your Mama or Daddy (or whomever cooked for you) used to make.

Figure 10.1: **Figure 10.1** Cup of chicken noodle soup with carrots and green beans. Yum![1]

Although chicken noodle soup is a worldwide phenomenon, not everyone makes chicken noodle soup the same way. There are a thousand different recipes for a thousand different chefs. Imagine your ingredients. What elements would you need to make good chicken soup?

Chicken? Yes

Noodles? Yes

Broth? Yes

Carrots? Yes

Garlic? Yes

Parsley? Yes

Onion? Yes

Salt? Yes

Pepper? Yes

Turmeric? Yes

Sardines? Yes

Chocolate? Yes

Figure 10.2: **Figure 10.2** Sardines grilled and served with lemon[2]

Figure 10.3: **Figure 10.3** Dark chocolate and mint chocolate chips[3]

Whoa…whoa…whoa…hold up. Sardines?!? Chocolate?!? We probably had you nodding and maybe even your stomach growling until these last two ingredients. Maybe you like chocolate, but it does not belong in chicken noodle soup. Maybe you hate sardines. If you put spoiler items in your recipe, it might ruin it, and you could need to throw it out and start over. For some other ingredients, if they are missing, you still have a recipe, but one that is not as good. Yet other ingredients are an absolute must. Do you have soup if you do not add broth? No, you have a casserole or stir fry instead.

That is really the big picture of what Cronbach's a does: It makes sure we have a good 'recipe' for our survey measure. It makes sure all the items correlate with one another. Enough ingredients, not too few or too many. Nothing irrelevant. Nothing that makes you retch or start over. Ready to serve for a dinner party. Ready to use in our research study and trust our outcomes.

10.3 Cronbach's a: How Reliable is 'Enough' or 'Too Much'?

Cronbach's a is expressed as a number somewhere between 0 and 1. Those are its mathematical limits.

If you ever get a negative number for Cronbach's a, this indicates that you seriously messed something up. The smoke alarm is going off. The pan is charred. The dog won't even eat it. Start over. Sometimes the items are just plain wrong and do not belong, like if your measure of self-esteem also included people's age or how much they like peanut butter. Sometimes if you get a negative number, you have an item that is relevant, but it is backwards and needs to be ***reverse scored***, which is where you turn a scale from, say, 1 to 5 instead of 5 to 1. For example, if you were doing a measure of attitudes toward mask-wearing for COVID, where a score of 1 = Strongly Disagree and 5 = Strongly Agree. If you had an item that said, "Wearing masks in public is really annoying," it is relevant to your idea, but it is written in the opposite of what you are looking for. Someone who scores highly on this item does *[not]* like wearing masks. In this case, instead of throwing the item out, you would reverse it so a high score becomes a low score…then it will play nicely (mathematically) with your other items.

We also want to know when our measure is good enough to use in research. If we were cooking, we would want to know when our recipe is good enough to serve to guests for dinner. Tavakol and Dennick (2011) note that mathematicians differ a little on what stands for 'good enough,' with estimates somewhere between .70 and .95. Across authors we have seen, an a of .70 is generally considered 'sufficient' or 'good' for research purposes. Below that, and it will typically be called 'questionable', with the lower the number goes shifting to 'poor' or 'unreliable.' In other words, an a below .70 calls into question whether you have good data and, therefore, might not really be measuring what you think you are in your research, no matter what you are trying to find, study, or test. Maybe you should start over and get another measure. If you have already run your project, this could lead to researchers not to trust your whole study, and you could have trouble getting your research published or funded (if you were seeking grants). Thus, it is important to make sure when you are *planning* your study that you have good survey measures that are reliable. It is usually better, therefore, to use measures that have been created and checked by other authors in their studies, rather than making up questions on your own. If you do make up your own questions, the burden of proof is on you – you need to show that you have made a good batch (or recipe) of items that go well together and make sense.

An alpha level can also be too high. Tavakol and Dennick (2011) recommend .90, other authors say .95. If a is too high, then we are being redundant, like saying "add noodles," "add noodles," "add noodles" when we are cooking. We already *said* to add noodles, so we do not need to say it again. If you ask participants the same or nearly identical question over and over on a survey, they will tend to get irritated and less cooperative, because you are wasting their time.

10.4 Cronbach's a: Some Caveats

When Cronbach's a is high, it is a good starting point, but it is not necessarily a measure of ***unidimensionality*** (i.e., an indicator that a scale is measuring a single factor or construct rather than multiple related constructs). Scales that are multidimensional will cause alpha to

be under-estimated (low) if not assessed separately for each dimension, but a high value for alpha is not always an indicator of unidimensionality. So, a = .80 does not mean that 80% of a single underlying construct is accounted for. It could be that the 80% comes from more than one underlying construct. In other words, a strong Cronbach's a score is ***necessary but not sufficient*** for determining unidimensionality. If Cronbach's a is low, the measure cannot be unidimensional. The items do not go well together. If the Cronbach's a is high, the measure *might* be unidimensional, but it is not necessarily so.

Revisiting our chicken noodle soup measure for illustration. First, the sardines. You're cooking. If you are rolling right along, building your chicken noodle soup, and throw in an ingredient you do not like that does not go well with the others (say sardines or perhaps rotten tomatoes), it will ruin the recipe. Cronbach's a will be low, and you will need to start all over, taking the stinker ingredient (your bad survey item) out.

If we have all of our ingredients, which are things that we like, and we also add another thing that we like (chocolate), it will not necessarily ruin the statistic. Instead of measuring how well everything tastes together, we might just be measuring "things I like to eat," irrespective of whether we like to eat them [together], in this recipe. So even if the numbers play nice, we also need to read and review the items to make sure they fit together logically. In this case, reliability is also a necessary but not sufficient condition for ***validity*** (accuracy). If something is going to be measured right (valid), it must be consistent (reliable)…but just because it is reliable does not make it valid.

Another illustration that reliability is not validity can be seen in the following recipe. You decide to make chicken noodle soup, so you throw together ice cream, hot fudge sauce, peanut butter sauce, sprinkles, whipped cream, and a cherry. Voilá! Chicken noodle soup! By now you are looking pretty confused and saying, "That is not chicken noodle soup…that is an ice cream sundae." And right you are! But mathematically, you would still have a strong Cronbach's a, because all of these items go together well. They are reliable but not valid items for a chicken noodle soup recipe.

Cronbach's a can also be affected by the length of the survey (Tavakol & Dennick 2011). If there are too few items, Cronbach's a will not be as strong. (Consider if you just made soup with chicken, noodles, and broth. It would be mid, or just OK.). You will need to add more ingredients (more survey items).

Further, another feature of a is that it tends to be sample specific: It is not a characteristic of the scale, but rather a characteristic of the sample in which the scale has been used. A biased, unrepresentative, or small sample could produce a very different a coefficient than a large, representative sample. The a can even vary from large sample to large sample. The picture of chicken noodle soup above shows mushrooms and green beans. Maybe you did not grow up eating chicken noodle soup with mushrooms and beans but instead had celery and onions. In Mexico, chicken noodle soup will almost certainly have cilantro instead of parsley. In China, chicken noodle soup typically has rice noodles or wonton noodles instead of egg noodles, while in Japan it will have soba or udon noodles. How you make the recipe depends on who and

where you ask the questions. Your oma or grandma's recipe for chicken noodle soup might not go over as well in China, Japan, or Mexico, where the cultural expectations for the recipe are a little different. Similarly, the questions we ask on our survey might not go over as well with a new group of people of a different age in a different place. That is why researchers would check the Cronbach's a for their study even if they are using a measure that has worked well in other studies. Even if we believe it is a good recipe, we need to sip and taste it before serving it up.

Please watch: Cronbach's Alpha (Simply Explained).

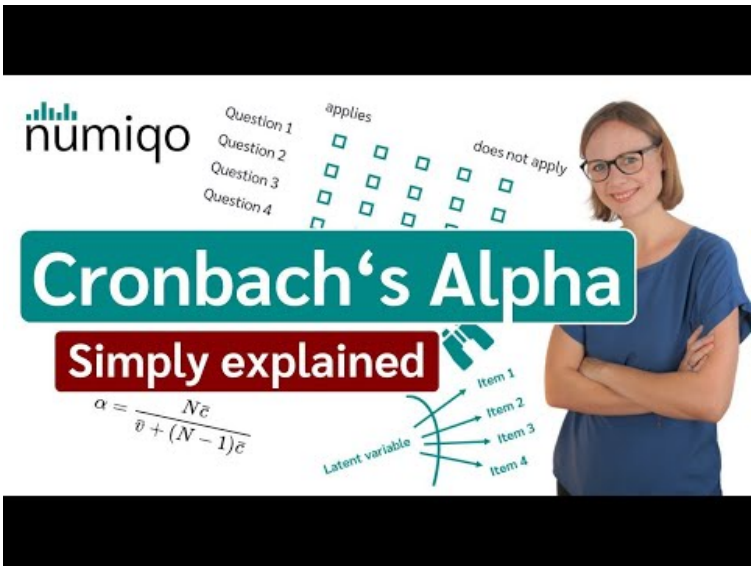


Figure 10.4: Cronbach's Alpha (Simply explained)

10.5 Summary

Despite some of its limitations, Cronbach's a has been popular in psychology for estimating internal consistency reliability—for making sure that the items go together well. It is fairly easy to calculate, understand and interpret, so it can be a useful initial check on scale performance when you administer a survey with a different sample from a different setting, population, or age group, for example, than the original authors who created the measure.

10.6 Additional Resources

For tips and demonstrations on setting up and running statistical analyses in SPSS, please see Virginia Wickline: SPSS Statistics Helps (publishing dates vary).

10.7 Media Attributions

[1] (CC BY-SA 2.0 by Casja Lillihook via Wikimedia Commons)

[2] (CC BY-SA 2.0 by Boxley via Wikimedia Commons)

[3] (CC BY-SA 2.0 by TaurusEmerald via Wikimedia Commons)

10.8 Text Attributions

Navarro, D. J., & Foxcroft, D. R. (2025). *Learning statistics with jamovi: A tutorial for beginners in statistical analysis.* Open Book Publishers. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified extensively by the current authors.

10.9 References

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16(3)*, 297–334. https://doi.org/10.1007/BF02310555

Numiqo (2021, October 27). *Cronbach's alpha (simply explained)*. https://www.youtube.com/watch?v=W9uPvAmtTOk

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's a. *International Journal of Medical Education, 2,* 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Part IV

Module 4: Analyzing Data - Calculating Inferential Statistics

11 Chapter 11: The Basics of Hypothesis Testing

After reviewing descriptive statistics, it is now time to turn our attention to a really big idea in statistics, which is ***hypothesis testing***. In its most abstract form, hypothesis testing is really a very simple idea. The researcher has some idea about the world and wants to determine whether or not the data actually support that idea. However, the details are messy, and most people find the theory of hypothesis testing to be the most frustrating part of statistics. First, we will explore how hypothesis testing works in a fair amount of detail, using a simple running example to show you how a hypothesis test is "built". Then, we will spend a bit of time talking about the various dogmas, rules, and heresies that surround the theory of hypothesis testing.

11.1 Logic and Purpose of Hypothesis Testing

A ***hypothesis*** is a prediction that is tested in a research study. Here we will present an example based on a James Bond movie where the superspy insisted that martinis should be shaken rather than stirred. Consider a hypothetical experiment to determine whether Mr. Bond can tell the difference between a shaken martini and a stirred martini. Suppose we gave Mr. Bond a series of 16 taste tests. In each test, we flipped a fair coin to determine whether to stir or shake the martini. Then, we presented the martini to Mr. Bond and asked him to decide whether it was shaken or stirred. If Mr. Bond was correct on 13 of the 16 taste tests, does this prove that Mr. Bond has at least some ability to tell whether the martini was shaken or stirred?

Figure 11.1: **Figure 11.1** That is one classy-looking martini! We hope you approve, Mr. Bond[1]

This result does not prove that he does; it could be he was just lucky and guessed right 13 out of 16 times. But, how plausible is the explanation that he was just lucky? To assess its plausibility, we determine the probability that someone who was just guessing would be correct 13/16 times or more. Using a probability table for assistance, this probability can be computed to be .011. This is a pretty low probability, and therefore someone would have to be very lucky to be correct 13 or more times out of 16 if they were just guessing. So, either Mr. Bond was very lucky and it was just a fluke, or he really can tell whether the drink was shaken or stirred. The hypothesis that he was guessing is not proven false, but considerable doubt is cast upon it. Therefore, there is strong evidence that Mr. Bond can tell whether a drink was shaken or stirred.

Our example aligns with what science philosopher Karl Popper (1959) called ***falsifiability***: The goal of science is to rule out the bad or unreasonable answers, so that whatever we are left with could be ***plausible***, which means apparently true or reasonable. Those who were fans of the TV show *Mythbusters* (see Figure 11.2) might remember that they had three outcomes for any of the fantabulous myths they tested: confirmed, busted, or plausible.

Figure 11.2: **Figure 11.2** Mythbusters Grant Imahara, Jamie Hyneman, and Adam Savage at the 2009 San Diego ComicCon Convention[2]

If your proposition is, "Everyone loves chocolate," it only takes one chocolate-hater or one person with a chocolate allergy to prove you wrong. However, if everyone you happen to ask loves chocolate, that also does not prove without a doubt that *everyone* in the whole world loves chocolate. It just makes it plausible. You would need to encounter all the world's inhabitants from *all* times and *all* places to prove your point without any doubt, which would be impossible.

11.2 The Probability Value

It is very important to understand precisely what probability values mean. In the James Bond example, the computed ***probability*** of .0106 is the likelihood that he would be correct on 13 or more taste tests (out of 16) if he were just guessing. It is easy to mistake this probability of .0106 as the probability he cannot tell the difference. *This is not at all what it means.*

The probability of .0106 is the probability of a certain outcome (13 or more out of 16) assuming a certain state of the world (James Bond was only guessing). It is not the probability that a state of the world is true. Although this might seem like a distinction without a difference, consider the following example. An animal trainer claims that a trained bird can determine whether or not numbers are evenly divisible by 7. In an experiment assessing this claim, the bird is given a series of 16 test trials. On each trial, a number is displayed on a screen and the bird pecks at one of two keys to indicate its choice. The numbers are chosen in such a way that the probability of any number being evenly divisible by 7 is .50. The bird is correct on 9/16 choices. We can compute that the probability of being correct nine or more times out of 16 if one is only guessing is .40. Since a bird who is only guessing would do this well 40% of the time, these data do not provide convincing evidence that the bird can tell the difference

between the two types of numbers. As a scientist, you would be very skeptical that the bird had this ability. Would you conclude that there is a .40 probability that the bird can tell the difference? Certainly not! You would think the probability is much lower than .0001.

To reiterate, the probability value is the probability of an outcome (9/16 or better) and not the probability of a particular state of the world (the bird was only guessing). In statistics, it is conventional to refer to possible states of the world as hypotheses since they are hypothesized states of the world. Using this terminology, the ***probability value (p-value)*** is the likelihood of an outcome given the hypothesis. It is not the probability of the hypothesis given the outcome.

This is not to say that we ignore the probability of the hypothesis. If the probability of the outcome given the hypothesis is sufficiently low, we have evidence that the hypothesis is false. However, we do not compute the probability that the hypothesis is false. In the James Bond example, the hypothesis is that he cannot tell the difference between shaken and stirred martinis. The probability value is low (.0106), thus providing evidence that he can tell the difference. However, we have not computed the probability that he can tell the difference.

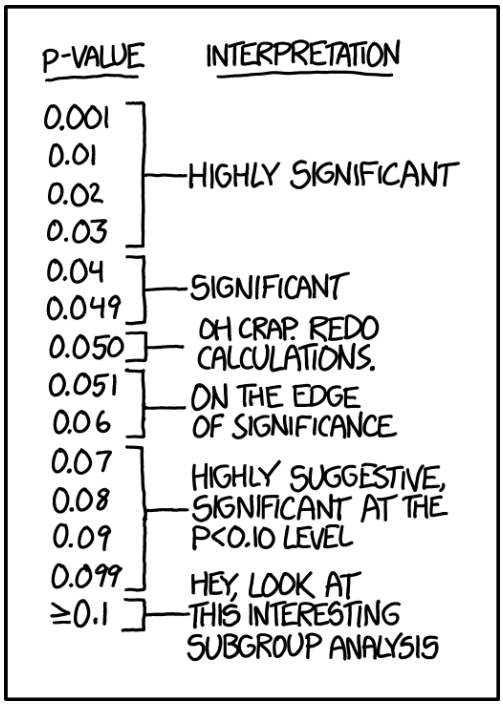


Figure 11.3: **Figure 11.3** While p-values are really just labeled as significant or not depending on whether the p-value is smaller than the alpha level or not, this cartoon adds (hopefully) a little humor to your learning experience[3]

11.3 A Menagerie of Hypotheses

Suppose we are interested in studying the phenomenon of extrasensory perception (ESP). Often displayed in sci-fi movies and TV shows, ESP is an ability to perceive information without any of the typical senses. Our first study is a simple one in which we seek to test whether clairvoyance exists: an awareness of objects that are hidden from view. Each participant sits down at a table and is shown a card by an experimenter. The card is black on one side and white on the other. The experimenter takes the card away and places it on a table in an adjacent room. The card is placed black side up or white side up completely at random, with the randomization occurring only after the experimenter has left the room with the participant. A second experimenter comes in and asks the participant which side of the card is now facing upwards. It is purely a one-shot experiment. Each person sees only one card and gives only one answer, and at no stage is the participant actually in contact with someone who knows the right answer. The data set, therefore, is very simple. I have asked the question of N people and some number (X) of these people have given the correct response. To make things concrete, suppose that we have tested $N = 100$ people and $X = 62$ of these got the answer right. A surprisingly large number, sure, but is it large enough for me to feel safe in claiming we have found evidence for ESP? This is the situation where hypothesis testing comes in useful. However, before we talk about how to test hypotheses, we need to be clear about what we mean by hypotheses.

11.3.1 Research Hypotheses Versus Statistical Hypotheses

The first distinction to keep clear is between research hypotheses and statistical hypotheses. In our ESP study, the overall scientific goal is to demonstrate that clairvoyance exists. In this situation there is a clear research goal: We are hoping to discover evidence for ESP. In other situations we might actually be a lot more neutral than that, so we might say that the research goal is to determine whether or not clairvoyance exists. Regardless of how we want to portray it, the basic point here is that a ***research hypothesis*** involves making a substantive, testable, scientific claim. If you are a psychologist, then your research hypotheses are fundamentally about psychological constructs. Any of the following would count as research hypotheses:

- *Listening to music reduces your ability to pay attention to other things.* This is a claim about the causal relation between two psychologically meaningful concepts (listening to music and paying attention to things), so it is a perfectly reasonable research hypothesis.
- *Intelligence is related to personality.* Like the last one, this is a relational claim about two psychological constructs (intelligence and personality), but the claim is weaker: correlational not causal.
- *Intelligence is the speed of information processing.* This hypothesis has a quite different character. It is not actually a relational claim at all. It is an ontological claim about the fundamental character of intelligence. It is usually easier to think about how to construct experiments to test research hypotheses of the form "does X affect Y?" than

it is to address claims like "what is X?" And in practice what usually happens is that scientists find ways of testing relational claims that follow from the ontological ones. For instance, if I believe that intelligence is the speed of information processing in the brain, my experiments will often involve looking for relations between measures of intelligence and measures of speed. As a consequence, most everyday research questions do tend to be relational in nature, but they are almost always motivated by deeper ontological questions about the state of nature.

Notice that, in practice, research hypotheses could overlap a lot. The ultimate goal in the ESP experiment might be to test an ontological claim like "ESP exists", but we might operationally restrict myself to a narrower hypothesis like "Some people can 'see' objects in a clairvoyant fashion". That said, there are some things that really do not count as proper research hypotheses in any meaningful sense:

- *Love is a battlefield*. This is too vague to be testable. While it is okay for a research hypothesis to have a degree of vagueness to it, it has to be possible to ***operationalize*** the theoretical ideas (that is, to figure out how the ideas are going to be measured). It might be challenging to see how this interesting idea can be converted into any concrete research design. If so, then this is not a scientific research hypothesis – it is just a catchy pop song. A lot of deep questions that humans have fall into this category. Maybe one day science will be able to construct testable theories of love, or to test to see if God exists, and so on. Until then, there are some good questions that science just cannot address.
- *The first rule of tautology club is the first rule of tautology club.* This is not a substantive claim of any kind. It is true by definition. No conceivable state of nature could possibly be inconsistent with this claim. We say that this is an unfalsifiable hypothesis, and as such it is outside the domain of science. Whatever else one does in science, our claims must have the possibility of being wrong.
- *More people in my experiment will say "yes" than "no".* This one fails as a research hypothesis because it is a claim about the data set, not about the psychology (unless of course your actual research question is whether people have some kind of "yes" bias!). Actually, this hypothesis is starting to sound more like a statistical hypothesis than a research hypothesis.

As you can see, research hypotheses can be somewhat messy at times, and ultimately, they are scientific claims. ***Statistical hypotheses*** are neither of these two things. Statistical hypotheses must be mathematically precise, and they must correspond to specific claims about the characteristics of the data generating mechanism (i.e., the "population"). Even so, the intent is that statistical hypotheses bear a clear relation to the substantive research hypotheses that we care about! For instance, in our ESP study, our research hypothesis is that some people are able to see through walls or whatever. What I want to do is to "map" this onto a statement about how the data were generated. So, what would that statement be? The

quantity that we are interested in within the experiment is P(correct), the true-but-unknown ***probability*** (or likelihood) with which the participants in my experiment answer the question correctly. Let's use the Greek letter q (*theta*) to refer to this probability. Here are four different statistical hypotheses:

- If ESP does not exist and if my experiment is well designed, then our participants are just guessing. So, we should expect them to get it right half of the time. Our statistical hypothesis is that the true probability of choosing correctly is q = .50.
- Alternatively, suppose ESP does exist and participants can see the card. If that is true, people will perform better than chance. The statistical hypothesis is that q > .50.
- A third possibility is that ESP does exist, but the colors are all reversed, and people do not realize it (okay, that is a little wacky, but you never know). If that is how it works, then we would expect people's performance to be below chance. This would correspond to a statistical hypothesis that q < .50.
- Finally, suppose ESP exists but I have no idea whether people are seeing the right colour or the wrong one. In that case the only claim I could make about the data would be that the probability of making the correct answer is not equal to 0.5. This corresponds to the statistical hypothesis that q =/= .50.

All of these are legitimate examples of a statistical hypothesis because they are statements about a population parameter and are meaningfully related to the experiment.

What this discussion hopefully reveals is that when attempting to construct a statistical hypothesis test, the researcher actually has two quite distinct hypotheses to consider. First, there is a research hypothesis (a claim about psychology), and this then corresponds to a statistical hypothesis (a claim about the data generating population). In our ESP example these might be as shown in Table 11.1

Table 11.1: **Table 11.1** Research and Statistical Hypotheses[4]

Type of Hypothesis	Claim Made
Our research hypothesis:	ESP exists
Our statistical hypothesis:	$\theta \neq 0$

A key thing to recognize is this: *A statistical hypothesis test is directly a test of the statistical hypothesis, NOT the research hypothesis.* If a study is badly designed, then the inherent link between the research hypothesis and the statistical hypothesis is broken. To give a silly example, suppose that our ESP study was conducted in a situation where the participant can actually see the card reflected in a window. If that happens, we could find very strong statistical evidence that $\theta \neq 0$, but this would tell us nothing about whether "ESP exists." Thus, a study needs to

be designed well, with few flaws or errors (called ***confounds***). To refer to a previous chapter, having a good design with no flaws or problems is otherwise known as ***internal validity***.

Please watch 6 Steps to Formulate a STRONG Hypothesis

Figure 11.4: 6 Steps to Formulate a STRONG Hypothesis | Scribbr

11.3.2 Null Hypotheses and Alternative Hypotheses

So far, so good. We have a research hypothesis that corresponds to what we want to believe about the world, and we can map it onto a statistical hypothesis that corresponds to what we want to believe about how the data were generated. It is at this point that things get somewhat counter-intuitive for a lot of people. Because what we do is invent a new statistical hypothesis (the ***null hypothesis, H_0***) that corresponds to the exact *opposite* of what we want to believe or think will happen. The null hypothesis suggests that an apparent effect is due to chance**.** then focus exclusively on that almost to the neglect of the thing we actually interested in. Keep in mind that null means nothing, nada, zip, zero. It signifies an empty state, nothingness, or a thing negated. If we made a deal that was null, our deal is no good. If an Elvis chapel marriage is annulled after the parties sober up in Vegas, it is canceled as if it never happened.

If the null hypothesis is that in the population of physicians, the mean time expected to be spent with obese patients is equal to the mean time expected to be spent with average-weight patients where the Greek letter m (or "mu") is used for the population means. This null hypothesis can be written as:

$$H_0 : \ \mu_{obese} - \ \mu_{average} = 0$$

The null hypothesis in a correlational study of the relationship between high school grades and college grades would typically be that the population correlation is 0. This can be written as:

$$H_0 : \ \rho = 0$$

where r (Greek letter "rho") is the population correlation value (we will cover correlation in a later chapter).

Although the null hypothesis is usually that the value of a parameter is 0, there are occasions in which the null hypothesis is a value other than 0. For example, if we are working with mothers in the U.S. whose children are at risk of low birth weight, we can use 7.47 pounds, the average birth weight in the U.S., as our null value and test for differences against that. In this case, using birth weight as an example, our null hypothesis takes the form:

$$H_0 : \ \mu = 7.47$$

The number on the right-hand side is our null hypothesis value that is informed by our research question. Notice that we are testing the value for m, the population parameter or mean, *not* the sample statistic's mean (M). This is for two reasons: (1) once we collect data, we know what the value of M is—it is not a mystery or a question, it is observed and used for the second reason, which is (2) we are interested in understanding the population, not just our sample.

Again, keep in mind that *the null hypothesis is typically the opposite or absence of the researcher's hypothesis*. It is what we do [not] expect to happen. In the example about doctors and patients, the researchers hypothesized that physicians would expect to spend less time with obese patients. The null hypothesis that the two types of patients are treated identically is put forward with the hope that it can be discredited and therefore rejected. If the null hypothesis were true, a large difference in the sample would be very unlikely to occur.

In general, the null hypothesis is the idea that nothing is going on: there is no effect of our treatment, no relationship between our variables, and no difference in our sample mean from what we expected about the population mean. This is always our baseline starting assumption, and it is what we seek to reject. If we are trying to treat depression, we want to find a difference in average symptoms between our treatment and control groups. If we are trying to predict job performance, we want to find a relationship between conscientiousness and evaluation scores. However, until we have evidence against it, we must use the null hypothesis as our starting point.

In our ESP example, the null hypothesis is that $\theta = 0$, since that is what we would expect if ESP does not exist. Our hope if we are studying it, of course, is that ESP is real and so the alternative to this null hypothesis is $\theta \neq 0$. In essence, what we are doing is dividing up the possible values of q into two groups: those values that we really hope are not true (the null), and those values that we would be happy with if they turn out to be right (the alternative). Having done so, the important thing to recognize is that the goal of a hypothesis test is not to show that the alternative hypothesis is (probably) true. The goal is to show that the null hypothesis is (probably) false. Most people find this pretty weird.

The best way to think about it might be to imagine that a hypothesis test is like an American criminal trial. The null hypothesis is the defendant, the researcher is the prosecutor, and the statistical test itself is the judge. Just like a criminal trial, there is a presumption of innocence. The null hypothesis is deemed to be true unless the researcher can prove beyond a reasonable doubt that it is false. You are free to design your experiment however you like (within reason, obviously!) and your goal when doing so is to maximize the chance that the data will yield a conviction for the crime of being false. The catch is that the statistical test sets the rules of the trial and those rules are designed to protect the null hypothesis, specifically to ensure that if the null hypothesis is actually true the chances of a false conviction are guaranteed to be low. This is pretty important. After all, the null hypothesis does not get a lawyer, and given that the researcher is trying desperately to prove it to be false, someone has to protect it.

Figure 11.5: **Figure 11.4** Continuing to report on hypothesis testing is controversial to some statisticians, as humorized here[5]

If the null hypothesis is rejected, then we will need some other explanation, which we call the alternative hypothesis, H_{A} or H_1. The ***alternative hypothesis*** is simply the reverse of the null hypothesis, and there are three options, depending on where we expect the difference to lie. Thus, our alternative hypothesis is the mathematical way of stating our research question.

In most cases, the alternative hypothesis is known as ***non-directional***: We expect a difference or a relationship, but we are not sure which way the relationship or difference is going to go. In that case, our statistics will be what is called ***two-tailed.*** In other words a difference or relationship that was big could be either on the high end or on the low end of a distribution of possible scores.

$$H_A: \mu \neq 7.47$$

If we expect our obtained sample mean to be above or below the null hypothesis value, which we call a ***directional hypothesis***, then our alternative hypothesis takes the form:

$H_A: \mu < 7.47$ or $H_A: \mu > 7.47$

based on the research question itself and is otherwise known as ***one-tailed***. We should only use a directional hypothesis if we have good reason, based on prior observations or research, to suspect a particular direction.

We will set different criteria for rejecting the null hypothesis based on the directionality (greater than, less than, or not equal to) of the alternative. To understand why, we need to see where our criteria come from and how they relate to z scores and distributions.

11.3.3 Two Types of Errors

Before going into details about how a statistical test is constructed, it is important to understand the philosophy behind it. We hinted at it when pointing out the similarity between a null hypothesis test and a criminal trial, but we should now be explicit. Ideally, we would like to construct our test so that we never make any errors. Unfortunately, since the world is messy, this is never possible. Sometimes you are just really unlucky or, as Abelson (1995) said in his First Law of Statistics, chance is really lumpy. For instance, suppose you flip a coin 10 times in a row and it comes up heads all 10 times. That feels like very strong evidence for a conclusion that the coin is biased, but of course, there is a 1 in 1,024 chance that this would happen, even if the coin were totally fair. In other words, in real life we always have to accept that there is a chance that we made a mistake and just got a really unlikely outcome. As a consequence, the goal behind statistical hypothesis testing is not to eliminate errors, but to minimize them.

At this point, we need to be a bit more precise about what we mean by errors. First, we will state the obvious. It is either the case that the null hypothesis is true or that it is false, and our test will either ***retain*** the null hypothesis or ***reject*** it. So, as Table 11.2 illustrates, after we run the test and make our choice, one of four things can happen:

Table 11.2: **Table 11.2** Null Hypothesis Statistical Testing (NHST)[6]

In the Real World…	**…and We Retain Null**	**… and We Reject Null**
H_0 is true	correct decision	error (Type I)
H_0 is false	error (Type II)	correct decision

As a consequence, there are actually two different types of error here. If we reject a null hypothesis that is actually true, then we have made a ***Type I error*** (otherwise known as a ***false positive***). In other words, we have said "something happens" when in real life "nothing happens. In real life, this would be like convicting an innocent person, or misreading a mammogram and saying someone has breast cancer when they do not. On the other hand, if we retain the null hypothesis when it is in fact false, then we have made a ***Type II error (false negative)***. In other words, we have said"nothing happens" when in real life, "something happens." In real life, this would be like letting a guilty person go free or misreading a mammogram and saying someone does not have breast cancer when they do. Both Type I and and Type II errors can have grave consequences, and we do not want to make either of these kinds of errors. Please note that this is not saying the data were collected incorrectly, recorded incorrectly, or made up – those kinds of mistake by the researcher would be entirely different (and, some would say, worse) kinds of errors.

Remember how we said that statistical testing was kind of like an American criminal trial? A criminal trial requires that establishing "beyond a reasonable doubt" that the defendant did it. All of the evidential rules are (in theory, at least) designed to ensure that there is (almost) no chance of wrongfully convicting an innocent defendant. The trial is designed to protect the rights of a defendant. In other words, a criminal trial does not treat the two types of error in the same way. Punishing the innocent is deemed to be *much* worse than letting the guilty go free. A statistical test is pretty much the same. The single most important design principle of the test is to control the probability of a type I error, to keep it below some fixed probability. This probability, which is denoted a, is called the ***significance level*** of the test. A hypothesis test is said to have significance level a if the type I error rate is no larger than a.

So, what about the type II error rate? We would also like to keep those under control, and we denote this probability by b. However, it is much more common to refer to the ***power*** of the test, that is, the probability with which we reject a null hypothesis when it really is false, which is 1- b. To help keep this straight, here is the same table again but with the relevant numbers alphanumeric numbers added (see Table 11.3).

Table 11.3: **Table 11.3** Null Hypothesis Statistical Testing (NHST) – Additional Detail[7]

In the Real World…	**…and We Retain Null**	**… and We Reject Null**
H_0 is true	1 - a (probability of correct retention)	a (type I error rate)
H_0 is false	b (type II error rate)	1 - b (power of the test)

A powerful hypothesis test is one that has a small value of b, while still keeping a fixed at some (small) desired level. By convention, scientists make use of three different a levels: .05, .01, and .001. Notice the asymmetry here; the tests are designed to ensure that the level is kept small but there's no corresponding guarantee regarding b.

Please watch Why 0.05? The Number That Decides What's True

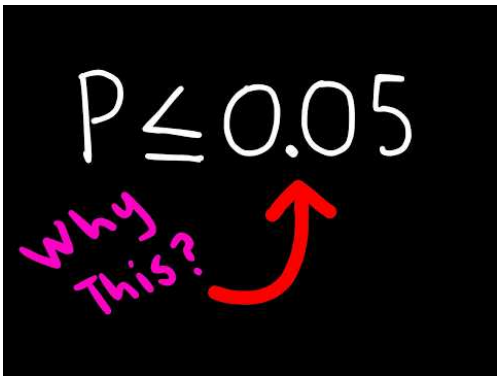


Figure 11.6: Why 0.05? The Number That Decides What's True

11.4 Significance Level, Critical Values, and *p* Values

A low probability value indicates that the sample outcome (or an outcome more extreme) would be very unlikely if the null hypothesis were true. A low probability value casts doubt on the null hypothesis. How low must the probability value be in order to conclude that the null hypothesis is false? Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than .05. More conservative or picky researchers conclude the null hypothesis is false only if the probability value is less than .01. When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis. The probability value below which the null hypothesis is rejected is called the a-level or simply a (Greek letter "alpha"). It is also called the ***significance level***. If a is not explicitly specified, assume that a = .05. This is the value we will use repeatedly for the course of the semester, since we are getting used to the language and use of statistical testing.

The significance level is a threshold we set before collecting data in order to determine whether or not we should reject the null hypothesis. We set this value beforehand to avoid biasing ourselves by viewing our results and then determining what criteria we should use. If our data produce values that meet or exceed this threshold, then we have sufficient evidence to reject the null hypothesis; if not, we fail to reject the null (we never "accept" the null).

There are two criteria we use to assess whether our data meet the thresholds established by our chosen significance level, and they both have to do with our discussions of probability. Recall that probability refers to the likelihood of an event, given some situation or set of conditions. In hypothesis testing, that situation is the assumption that the null hypothesis value is the correct value, or that there is no effect. The value laid out in H_0 is our condition under which we interpret our results. To reject this assumption, and thereby reject the null hypothesis, we need results that would be very unlikely if the null was true.

Let us say we were using a ***one-sample z-test***, which compares a sample of a mean to a population to see if they are similar or different. In this case, we would use the following formula to calculate the z-test:

$$z = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}}$$

In this case M = mean of the sample, m = mean of the population, s = population standard deviation, and n = the sample size. If a z-score is zero, it indicates that the mean of the sample is the same as the mean of the population (average). If a z-score is positive, it indicates that the mean of the sample is greater than the mean of the population (above average). If a z-score is negative, it indicates that the mean of the sample is greater than the mean of the population (below average). After we calculate the z-score, we would compare the z to a standard normal distribution to see how unlikely this particular z-score is. Values of z which fall in the tails of the standard normal distribution represent unlikely values. That is, the proportion of the area under the curve as extreme as z—or more extreme than z—is very small as we get into the tails of the distribution. Our significance level corresponds to the area in the tail that is exactly equal to a. If we use our normal criterion of a = .05, then 5% of the area under the curve becomes what we call the ***critical region*** (also called the rejection region) of the distribution. This is illustrated in Figure 11.5. The shaded rejection region takes us 5% of the area under the curve. Any result that falls in that region is sufficient evidence to reject the null hypothesis.

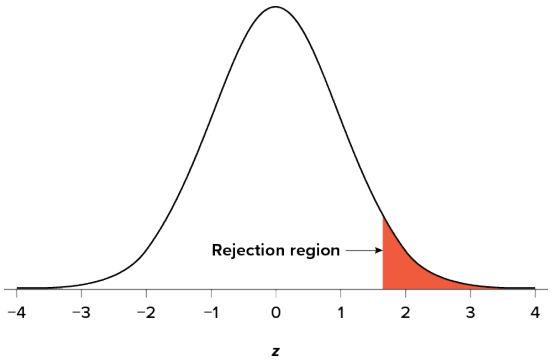


Figure 11.7: **Figure 11.5** The rejection region for a one-tailed z-test[8]

The rejection region is bounded by a specific z value, as is any area under the curve. In hypothesis testing, the value corresponding to a specific rejection region is called the ***critical value*** (hence the other name "critical region"). If we go to a ***unit normal table*** for probabilities, we will find that the z score corresponding to 5% of the area under the curve is equal to 1.645 (z = 1.64 corresponds to .0505 and z = 1.65 corresponds to .0495, so .05 is exactly in between them) if we go to the right and −1.645 if we go to the left. The direction must be determined by your alternative hypothesis, and drawing and shading the distribution is helpful for keeping directionality straight.

Suppose, however, that we want to do a non-directional test. We need to put the critical region in both tails, but we do not want to increase the overall size of the rejection region (for reasons we will see later). To do this, we simply split it in half so that an equal proportion of the area under the curve falls in each tail's rejection region. For a = .05, this means 2.5% of the area is in each tail, which, based on the z table, corresponds to critical values of $z = \pm 1.96$. This is shown in Figure 11.6.

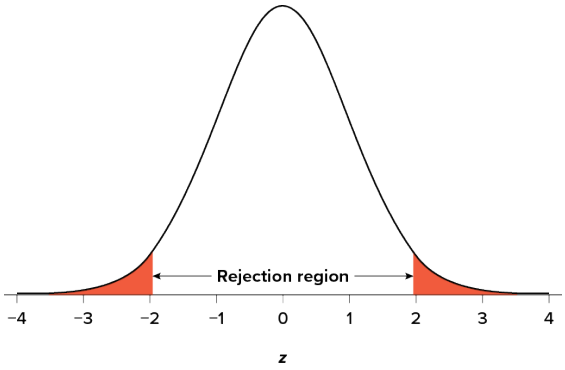


Figure 11.8: **Figure 11.6** Two-tailed rejection region for a z-test[9]

Thus, any z score falling outside ± 1.96 (greater than 1.96 in absolute value) falls in the rejection region. It is an extreme, unusual, or unlikely score. When we use z-scores in this way, the obtained value of z (sometimes called z obtained and abbreviated z_{obt}) is something known as a ***test statistic***, which is simply an inferential statistic used to test a null hypothesis.

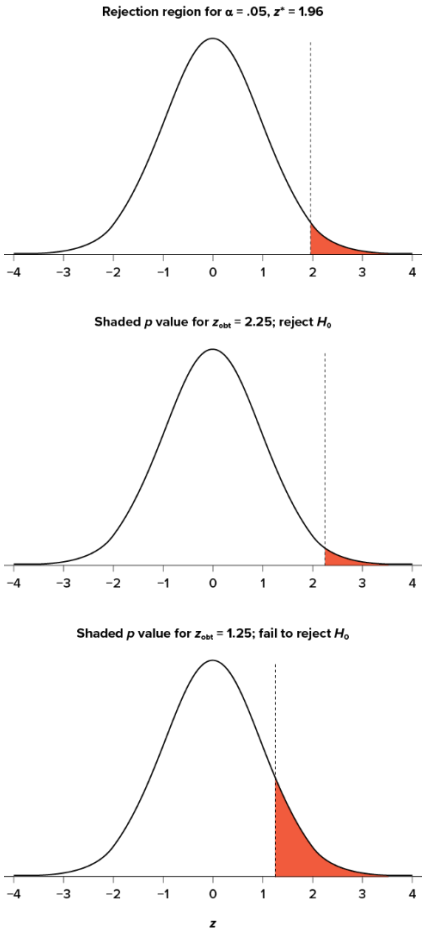


Figure 11.9: **Figure 11.7** Relation between a, z_{obt}, and p[10]

As illustrated in Figure 11.7, to formally test our hypothesis, we compare our obtained z statistic (z_{obt}) to our critical z value (z_{crit}). If $z_{obt} > z_{crit}$, that means it falls in the rejection region (to see why, draw a line for $z = 2.5$ on Figure 11.5 or 11.6. Thus, we ***reject*** H_0 – it looks like something interesting happened, rather than nothing. If $z_{obt} < z_{crit}$, we ***fail to reject*** the null hypothesis, or in other words, we ***retain*** (hang onto) the null hypothesis. Remember that as z gets larger, the corresponding area under the curve beyond z gets smaller. Thus, the proportion, or p value, will be smaller than the area for a, and if the area is smaller, the probability gets smaller. Specifically, the probability of obtaining that result, or a more extreme result, under the condition that the null hypothesis is true gets smaller.

The z statistic is very useful when we are doing our calculations by hand. However, when we use computer software, it will report to us a ***probability value (p-value)***, which is simply the proportion of the area under the curve in the tails beyond our obtained z statistic. We can directly compare this p value to a to test our null hypothesis: if $p <$ a, we reject H_0, but

if $p >$ a, we fail to reject. Note also that the reverse is always true. If we use critical values to test our hypothesis, we will always know if p is greater than or less than a. If we reject, we know that $p <$ a because the obtained z statistic falls farther out into the tail than the critical z value that corresponds to a, so the proportion (p value) for that z statistic will be smaller. Conversely, if we fail to reject, we know that the proportion will be larger than a because the z statistic will not be as far into the tail. This is illustrated for a one-tailed test in Figure 11.7.

When the null hypothesis is rejected, the effect is said to have ***statistical significance*** or to be ***statistically significant***. It is important to keep in mind that statistical significance means only that the null hypothesis of exactly no effect is rejected; it does not mean that the effect is important, which is what "significant" usually means. When an effect is significant, you can have confidence the effect is not exactly zero. Finding that an effect is significant does not tell you about how large or important the effect is.

Do not confuse statistical significance with practical significance (importance). A small effect can be highly significant if the sample size is large enough.

Why does the word "significant" in the phrase "statistically significant" mean something so different from other uses of the word? Interestingly, this is because the meaning of "significant" in everyday language has changed. It turns out that when the procedures for hypothesis testing were developed, something was "significant" if it signified something. Thus, finding that an effect is statistically significant signifies that the effect is likely to be real and not likely due to chance. Over the years, the meaning of "significant" changed, leading to the potential misinterpretation. Also, if a finding is not statistically significant, it is called ***non-significant*** (not statistically likely) rather than ***insignificant*** (which would suggest it is unimportant).

11.5 The Hypothesis Testing Process

11.5.1 A Four-Step Procedure

The process of testing hypotheses follows a simple four-step procedure. This process will be what we use for the remainder of the textbook and course, and although the hypotheses and statistics we use will change, this process will not. Commit these steps of hypothesis testing to memory, as we will revisit them for every inferential statistical test that we run this semester.

11.5.2 Step 1: Look at the Data and State the Null and Alternative Hypotheses

Our hypotheses are the first thing we need to lay out. Otherwise, there is nothing to test! We have to state the null hypothesis (which is what we test) and the alternative hypothesis (which is what we expect). These should be stated mathematically as they were presented above *and* in words, explaining in everyday language what each one means in terms of the

research question. Think of this as translating from math to English especially if you, like many of the authors, have ever thought that math was like a foreign language!

11.5.3 Step 2: Check Assumptions and Set the Critical Value(s)

Next, we consider any mathematical assumptions that exist for the particular test we want to use and formally lay out the criteria we will use to test our hypotheses. A ***critical value*** is the test statistic that exists at the particular probability value we decide is far enough away from the mean to be ***statistically significant,*** or unlikely to have happened by chance alone. There are three pieces of information that inform our critical value or values:

- a, which determines how much of the area under the curve composes our rejection region
- the directionality of the test, which determines what the critical region or regions will be (one-tailed or two-tailed), and
- N, or sample size.

The sample size helps us calculate the degrees of freedom for the test, which will either be equal to N - 1 or N - 2, depending on the test we are running. ***Degrees of freedom*** are the number of scores that can vary in an analysis without violating any established limits or assumptions of the test. Once you have all of these pieces, you can look up the critical value(s) in a probability table for that particular test. If you are using a statistical program to run your analyses for a hypothesis test, the program calculates and utilizes the critical value(s), but it does not tend to show it in the output.

11.5.4 Step 3: Calculate & Report the Descriptive Statistics, Test Statistic, and Effect Size

Once we have our hypotheses and the standards we use to test them, we can collect data and calculate our test statistic—in this case, z. This step is where the vast majority of differences in future chapters will arise: Different tests used for different data are calculated in different ways, but the way we use and interpret them remains the same. As part of this step, we will also calculate ***effect size***, or the magnitude of the difference between our groups or the relationship between variables. Does the statistic we found show no effect? Or is it small, medium, or large? Although effect size is not considered part of hypothesis testing, reporting it as part of the results is approved convention, especially in the American Psychological Association standards that guide and inform our work in psychology. A low probability value can occur with small effect sizes, particularly if the sample size is large.

11.5.5 Step 4: Interpret Your Findings in Relation to the Hypothesis (Translate Math to English)

Finally, once the test statistic is obtained, we can compare it to our critical value and decide whether we should reject or fail to reject (retain) the null hypothesis. When we do this, we must interpret the decision in relation to our research question, stating what we concluded, what we based our conclusion on, and the specific statistics we obtained.

Please watch: Hypothesis Testing

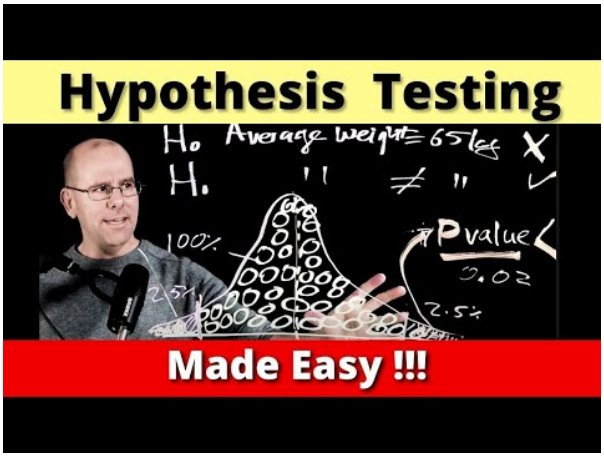


Figure 11.10: Hypothesis testing

11.6 An Example: Movie Popcorn

Figure 11.11: **Figure 11.8** A popcorn seller processing popcorn[11]

Let us see how hypothesis testing works in action by working through an example. Say that a movie theater owner likes to keep a close eye on how much popcorn goes into each bag sold, so she knows that the average bag has 8 cups of popcorn and that this varies a little bit, about half a cup. That is, the known population mean is m = 8.00 (cups) and the known population standard deviation is s = 0.50 (cups). The owner wants to make sure that the newest employee is filling bags correctly, so over the course of a week she randomly assesses 25 bags filled by the employee to test for a difference ($N = 25$). She does not want bags over-filled (the theater would lose money) or under-filled (movie-goers would complain), so she looks for differences in both directions. This scenario has all of the information we need to begin our hypothesis testing procedure.

11.6.1 Step 1: State the Hypotheses

Our manager is looking for a difference in the mean weight of popcorn bags compared to the population mean of 8. We will need both a null and an alternative hypothesis written both mathematically and in words. We will always start with the null hypothesis:

$$H_0 : \textit{There is no difference in the weight of popcorn bags from this employee.}$$

$$H_0 : \ \mu = 8.00.$$

Notice that we phrase the hypothesis in terms of the population parameter (m), which in this case would be the true average weight of bags filled by the new employee. Our assumption of no difference, the null hypothesis, is that this employee's mean is exactly the same as the known population mean value that we want it to match: 8.00.

Next is the alternative hypothesis:

$$H_A : \textit{There is a difference in the weight of popcorn bags from this employee.}$$

$$H_A : \ \mu \ \neq 8.00.$$

In this case, we do not know if the bags will be too full or not full enough, so we utilized a two-tailed alternative hypothesis that there is a difference, without specifying the direction.

11.6.2 Step 2: Find the Critical Values

Our critical values are based on three things: the directionality of the test, the level of significance, and (for some tests) the sample size. We decided in Step 1 that a two-tailed test is the appropriate directionality. We were given no information about the level of significance, so we assume that a = .05 is what we will use. As stated earlier in the chapter, the critical values for a two-tailed z test at a = .05 are $z = \pm 1.96$. These will be the criteria we use to test our hypothesis. We can now draw out our distribution, as shown in Figure 11.9, so we can visualize the rejection region and make sure it makes sense.

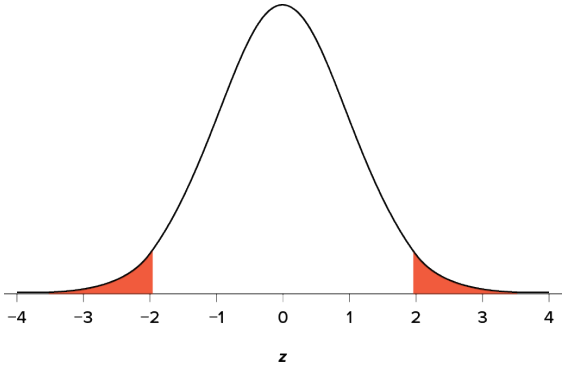


Figure 11.12: **Figure 11.9** Rejection region for z* = ±1.96[12]

11.6.3 Step 3: Calculate & Report the Descriptive Statistics, Test Statistic, and Effect Size

Now we come to our formal calculations. Let us say that the manager collects data and finds that the average weight of this employee's popcorn bags is $M = 7.75$ cups. We can now plug this value, along with the values presented in the original problem, into our equation for z:

$$z = \frac{7.75 - 8.00}{\frac{0.50}{\sqrt{25}}} = \frac{-0.25}{0.10} = -2.50$$

So, our calculated test statistic is $z = -2.50$, which we can draw onto our rejection region distribution as shown in Figure 11.10.

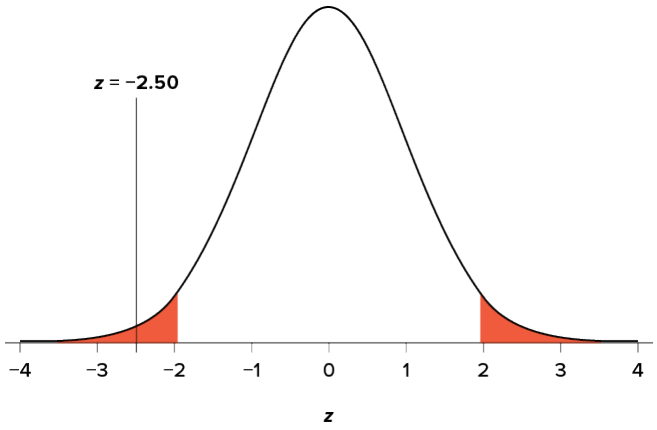


Figure 11.13: **Figure 11.10** Test statistic (z-score) location in the z-distribution[13]

11.6.3.1 Effect Size

When we reject the null hypothesis, we are stating that the difference we found was statistically significant, but we have mentioned several times that this tells us nothing about practical significance. To get an idea of the actual size of what we found, we can compute a new statistic called an effect size. ***Effect size*** gives us an idea of how large, important, or meaningful a statistically significant effect is. Each kind of test we run will have a different effect size. For mean differences like we calculated here, our effect size is Cohen's d:

$$d = \frac{M - \mu}{\sigma}$$

This is very similar to our formula for z, but we no longer take into account the sample size (since overly large samples can make it too easy to reject the null). Cohen's d is interpreted in units of standard deviations, just like z. For our example:

$$d = \frac{7.75 - 8.00}{0.50} = \frac{-0.25}{0.50} = -0.50$$

Cohen's d is interpreted as small, moderate, or large. Specifically, $d = 0.20$ is small, $d = 0.50$ is moderate (medium), and $d = 0.80$ is large (this is true whether Cohen's d is positive or negative). Obviously, values can fall in between these guidelines, so we should use our best judgment and the context of the problem to make our final interpretation of size. Our effect size happens to be exactly equal to one of these, so we say that there is a moderate effect.

Effect sizes are incredibly useful and provide important information and clarification that overcomes some of the weakness of hypothesis testing. Any time we perform a hypothesis test,

whether statistically significant or not, we should always calculate and report effect size as well.

11.6.4 Step 4: Interpret Your Findings in Relation to the Hypothesis (Translate Math to English)

Looking at Figure 11.10, we can see that our obtained z statistic falls in the rejection (critical) region. We can also directly compare it to our critical value: in terms of absolute value, $-2.50 > -1.96$, so we reject the null hypothesis. If we looked it up a unit normal table, we would see that a z-score of -2.50 has a probability of .005 (rounded to three decimals). We can now write our conclusion. When we write our conclusion, we write out the words to communicate what it actually means, but we also include the average sample size we calculated, the z statistic and calculated p-value, and the effect size. Our very last sentence should help translate math to English, or as one of the authors likes to say: "Now, tell my nana, my nephew, my neighbor—someone who does not 'speak math'—what the numbers mean in everyday language."

> Based on the sample of 25 bags, the average popcorn bag from this employee is smaller ($M = 7.75$ cups) than the average size of popcorn bags at this movie theater ($\mu = 8.00$ cups), $z = -2.50$, $p = .005$, $d = 0.50$. We would reject the null hypothesis; the alternative hypothesis is supported. The effect size was moderate. This employee seems to provide less popcorn than most employees.

11.7 Misconceptions in Hypothesis Testing

Misconceptions about significance testing are common. This section lists three important ones.

1. *Misconception:* The probability value (p-value) is the probability that the null hypothesis is false.
   Proper interpretation: The probability value (p-value) is the probability of a result as extreme or more extreme given that the null hypothesis is true. It is the probability of the data given the null hypothesis. It is not the probability that the null hypothesis is false.

2. *Misconception:* A low probability value indicates a large effect.
   Proper interpretation: A low probability value indicates that the sample outcome (or an outcome more extreme) would be very unlikely if the null hypothesis were true. A low probability value can occur with small effect sizes, particularly if the sample size is large.

3. *Misconception:* A non-significant outcome means that the null hypothesis is probably true.
   Proper interpretation: A non-significant outcome means that the data do not conclusively demonstrate that the null hypothesis is false.

11.8 Confidence Intervals

Previously, we discussed how scientists rely on statistics to make guesses about population parameters on the basis of a sample of data. Every data set leaves us with some uncertainty, so our estimates are never going to be perfectly accurate.

The thing that has been missing from this discussion is an attempt to quantify the amount of uncertainty that attaches to our estimate. It is not enough to be able guess that the mean intelligence quotient (IQ) of undergraduate psychology students is, say, 115 (yes, we arbitrarily made that number up). We also want to be able to say something that expresses the degree of certainty that we have in our guess. For example, it would be nice to be able to say that there is a 95% chance that the true mean lies between 109 and 121. The name for this estimate is a ***confidence interval*** for the mean.

Constructing a confidence interval for the mean is actually pretty easy. Here is how it works. Suppose the true population mean is m, and the standard deviation is s. Imagine we just finished running a study that has N participants, and the mean IQ among those participants is M *(*also sometimes written as X with a bar over top, as in the formula below*)*. We know from a phenomenon known as the central limit theorem that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution in Module 3 that there is a chance that a normally-distributed quantity will fall within about two standard deviations of the true mean.

To be more precise, the more correct answer is that there is a 95% chance that a normally distributed quantity will fall within 1.96 standard deviations of the true mean. Of course, there is nothing special about the number 1.96. It just happens to be the multiplier we need to use if we want a 95% confidence interval. If we wanted a 70% confidence interval, we would have used 1.04 as the magic number rather than 1.96. As long as N is sufficiently large (large enough for us to believe that the sampling distribution of the mean is normal), then we can calculate our 95% confidence interval with this formula:

$$CI_{95} = \overline{X} \pm (1.96 \times \frac{\sigma}{\sqrt{N}})$$

11.8.1 Interpreting a Confidence Interval

The hardest thing about confidence intervals is understanding what they mean. Whenever people first encounter confidence intervals, the first instinct is almost always to say that "there is a 95% probability that the true mean lies inside the confidence interval". It's simple and it seems to capture the common sense idea of what it means to say that I am "95% confident." Unfortunately, this interpretation is not quite right. The intuitive definition relies very heavily on your own personal beliefs about the value of the population mean. I say that I am 95% confident because those are my beliefs. In everyday life, this it perfectly okay, but in statistics, we should be more precise. Our interpretation of a 95% confidence interval must have something to do with replication.

Specifically, if we replicated the experiment over and over again and computed a 95% confidence interval for each replication, then 95% of those intervals would contain the true mean. More generally, 95% of all confidence intervals constructed using this procedure should contain the true population mean. This idea is illustrated in Figure 11.10, which shows 50 confidence intervals constructed for a "measure 10 IQ scores" experiment (top panel) and another 50 confidence intervals for a "measure 25 IQ scores" experiment (bottom panel).

We would expect that around 95 of our confidence intervals would contain the true population mean, and that is what we found in Figure 11.11.

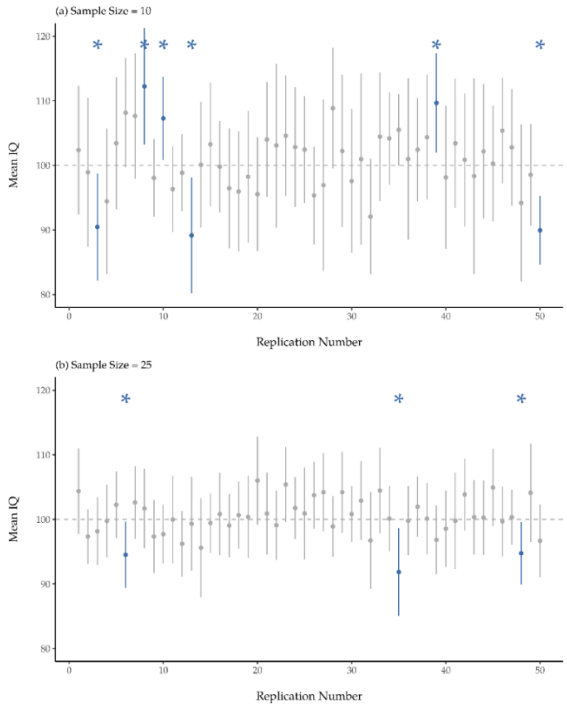


Figure 11.14: **Figure 11.11** The 95% confidence intervals for 10 versus 25 samples of intelligence quotient (IQ) scores[14]

In Figure 11.11, the top panel (a) shows 50 simulated replications of an experiment in which we measure the IQs of 10 people. The dot marks the location of the sample mean and the line shows the 95% confidence interval. Most of the 50 confidence intervals do contain the true mean (i.e., 100), but a few – in blue and marked with asterisks – do not. The lower graph (panel b) shows a similar simulation, but this time we simulate replications of an experiment that measures the IQs of 25 people.

When we are approaching hypothesis testing by looking at frequencies, the population mean is fixed, and no probabilistic claims can be made about it. Confidence intervals, however, are repeatable, so we can replicate experiments. Therefore, a *frequentist* is allowed to talk about the probability that the *confidence interval* (a random variable) contains the true mean but is not allowed to talk about the probability that the *true population mean* (not a repeatable event) falls within the confidence interval. I know that this seems a little pedantic, but it does matter. It matters because the difference in interpretation leads to a difference in the math.

Please watch Confidence Intervals: Crash Course Statistics #20

Figure 11.15: Confidence Intervals: Crash Course Statistics #20

11.9 Summary

Hypothesis testing (including the null and alternative hypothesis) is one of the most pervasive elements to statistical theory and research – in other words, it is practically everywhere in all things statistics. The vast majority of scientific papers report the results of some hypothesis test or another. As a consequence, it is almost impossible to get by in science without having at least a cursory understanding of what a p-value means, even if you are a statistician who argues that science should more away from using p-values at all or exclusively in understanding statistics (e.g., Ahmed & Butt, 2025; American Statistical Association, 2016; Wang & Long, 2022).

11.10 Additional Video Helps

- How P-Values Help Us Test Hypotheses: Crash Course Statistics #21
- P-Value Problems: Crash Course Statistics #22
- Playing with Power: P-Values Pt 3: Crash Course Statistics #23
- What Confidence Intervals Actually Mean

11.11 Additional Resources

- Unit Normal Table for Z-scores: See Appendix A of Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). *Introduction to statistics in the psychological sciences*.

11.12 Media Attributions

[1] (CC attribution 3.0 unported by Tobascoman77 via Wikimedia Commons)

[2] (CC attribution 2.0 generic by Dave Fayram via Wikimedia Commons)

[3] ("P-Values" by Randall Munroe/xkcd.com is licensed under CC BY-NC 2.5)

[4] (CC BY-SA 4.0 by Navarro & Foxcroft, 2025)

[5] ("Null Hypothesis" by Randall Munroe/xkcd.com is licensed under CC BY-NC 2.5)

[6-7] (CC BY-SA 4.0 by Navarro & Foxcroft, 2025)

[8] (Rejection Region for One-Tailed Test" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[9] ("Rejection Region for Two-Tailed Test" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[10] ("Relationship between alpha, z-obt, and p" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[11] (CC BY-NC-SA 4.0 by Rwebogora via Wikimedia Commons)

[12] ("Rejection Region z+-1.96" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[13] ("Test Statistic Location z-2.50" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[14] (CC BY-SA 4.0 by Navarro & Foxcroft, 2025)

11.13 Text Attributions

Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). *Introduction to statistics in the psychological sciences.* Pressbooks. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Navarro, D. J., & Foxcroft, D. R. (2025). *Learning statistics with jamovi: A tutorial for beginners in statistical analysis.* Open Book Publishers. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

11.14 References

Abelson, R. P. (1995). *Statistics as principled argument.* Psychology Press (a Taylor & Francis Group).

Ahmed, E. S., & Butt, M. N. (2025). The misunderstood p-value: Why statistical significance is not enough in clinical practice. *British Journal of Anaesthesia, 134*(4), 909-913. https://doi.org/10.1016/j.bja.2025.01.008

American Statistical Association (2016, March 7). *American Statistical Association releases statement on statistical significance and p-values.* https://www.amstat.org/asa/files/pdfs/p-valuestatement.pdf

CrashCourse (2018, June 13). *Confidence intervals: Crash course statistics #20.* https://www.youtube.com/watch?v=yDEvXB6ApWc

Foundations of Math. (2025, September 10). *Why 0.05? The number that decides what's true.* https://www.youtube.com/watch?v=CS7nGURwdC4

Global Health with Greg Martin (2022, February 22). *Hypothesis testing.* https://www.youtube.com/watch?v=S2eKynREGM4

Popper, K. (1959). *The logic of scientific discovery.* Hutchinson.

Scribbr (2019, November 15). *6 steps to formulate a STRONG hypothesis.* https://www.youtube.com/watch?v=PCgLjDDD4ek

Wang, M., & Long, Q. (2022). Addressing common misuses and pitfalls of p-values in biomedical research. *Cancer Research, 82*(15), 2674-2677. https://doi.org/10.1158/0008-5472.CAN-21-2978

12 Chapter 12: Correlation and Regression

Correlation and regression are both used to examine relations among variables, but they answer different kinds of questions. ***Correlation*** describes the strength and direction of the relation between two variables. ***Regression*** uses one or more predictor variables to predict or explain variation in a continuous outcome variable.

A key difference between correlation and regression is that correlation treats the two variables symmetrically. We ask whether two variables are related, but we do not label one variable as the predictor and the other as the outcome. Regression is inherently directional in its setup: one variable is the outcome, and one or more variables are used as predictors.

Table 12.1 *Overview of Correlation and Regression Designs*[1]

Analysis	Variables and Design	Research Question
Pearson correlation	Two continuous (interval or ratio) variables, normally distributed	Are the two variables linearly related?
Spearman correlation	One or both ordinal (including Likert scales) and/or non-normally distributed variables	Are the two variables related in rank order?
Simple linear regression	One predictor variable and one continuous (interval or ratio) outcome variable	Does one variable predict the outcome?
Multiple regression	Two or more predictor variables and one continuous (interval or ratio) outcome variable	How well do the predictors work together to predict the outcome?
Regression with categorical predictors	Continuous outcome with categorical and/or continuous predictors	Do groups differ on the outcome after being represented in a regression model?
Hierarchical regression	Predictors entered in blocks or steps	Does adding a new set of predictors improve the model?

We will continue using the same four-step hypothesis-testing process introduced in the previous chapter: look at the data and state the null and research hypotheses; check assumptions and set critical value(s); calculate and report the descriptive statistics, test statistic, and effect size; and interpret the results in relation to the hypotheses (translate math to English). The main

difference is that we are now asking about relations among variables rather than comparing group differences.

One important caution applies throughout this chapter: relations do not automatically imply causation. A correlation or regression coefficient can show that variables are related, but causal conclusions depend on the research design, measurement quality, and alternative explanations.

12.1 Correlation

Correlation is used to describe the relation between two variables. In this section, we focus mostly on the ***Pearson correlation***, which is used when we want to examine the linear relation between two continuous (interval or ratio) variables.

The ***correlation coefficient, r,*** developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable (x) and the dependent variable (y).

The correlation coefficient is calculated as

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2}}$$

where n = the number of data points, x = the value of one variable, and y = the value of a second variable.

A correlation can tell us three things about a relation:

1. ***Direction***: whether the relation is positive or negative.
2. ***Strength***: how closely the variables are related.
3. ***Statistical significance***: whether the observed relation is unlikely to have occurred by chance if there were no relation in the population.

Correlation coefficients range from -1 to +1. Anything beyond -1 or +1 is an impossible value. You cannot, for instance, have a correlation coefficient equal to -1.67 or +25.3.

- A ***positive correlation*** means that higher scores on one variable tend to go with higher scores on the other variable.
- A ***negative correlation*** means that higher scores on one variable tend to go with lower scores on the other variable.
- A correlation close to **0** means there is little or no linear relation between the variables.

Correlation does not show causation. A statistically significant correlation tells us that two variables are related, but it does not tell us whether one variable caused the other.

Pearson correlations are standardized, so they are also effect sizes. As a rough heuristic, values around (|.10|) are often described as small, values around (|.30|) as medium, and values around or beyond (|.50|) as large. These are only general guidelines. The importance of a correlation depends on the research context.

12.1.1 Scatterplots (or Scatter Plots)

The way we visualize a relation is by a type of graph called a scatterplot. A ***scatterplot*** looks at the relationship between two factors. Each dot in the graph is a respondent (participant) who has provided answers to both of the questions at hand. For the correlations we are doing, we want to be able to see or imagine a roughly straight line, rather than a wide cloud of dots (or what you might think of as a big chocolate chip cookie). The stronger the correlation, the more the dots will fall into or cluster around the regression line, otherwise known as the ***line of best fit***, a visual representation of the relationship between the two factors. Remember: A correlation cannot be greater than +1 or lower than -1. Those are its limits.

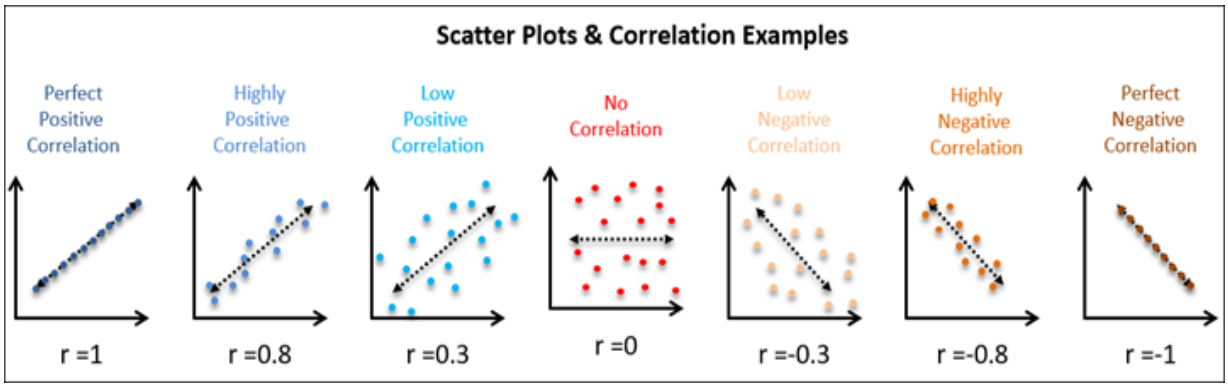


Figure 12.1: Seven scatterplots showing different r values ranging from -1 to 0 to 1.

Figure 12.1 *Scatter Plots and Correlation Examples Ranging from -1 to 0 to 1*[2]

Please watch: What Are Correlations? (Daniel Storage)

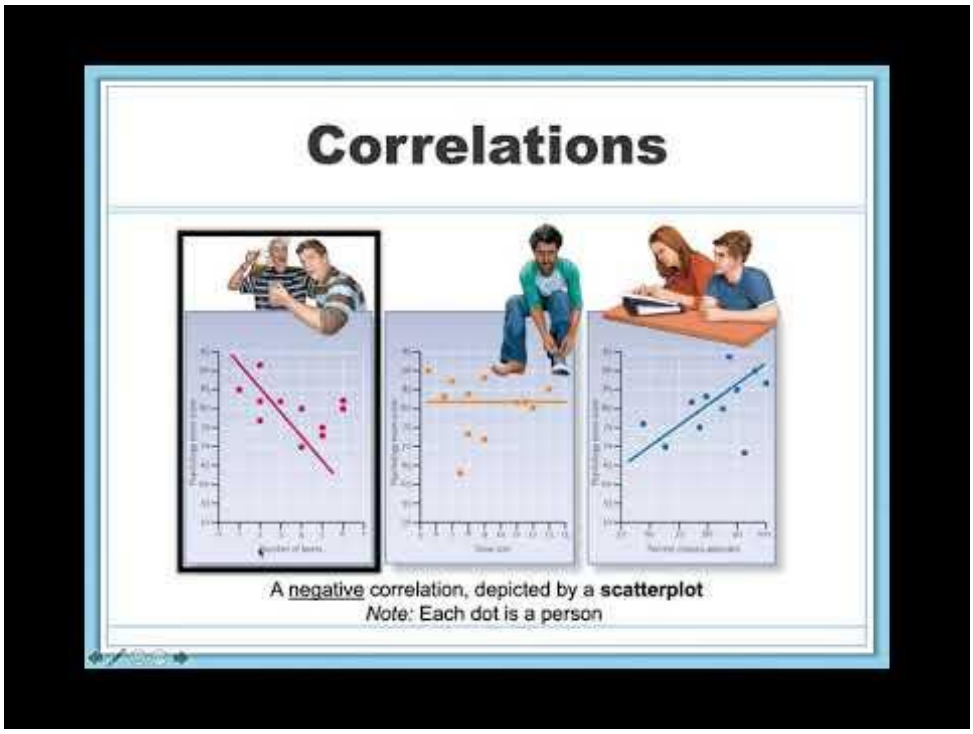


Figure 12.2: What Are Correlations?

Please watch: Pearson's r correlation (Statistics Lectures)

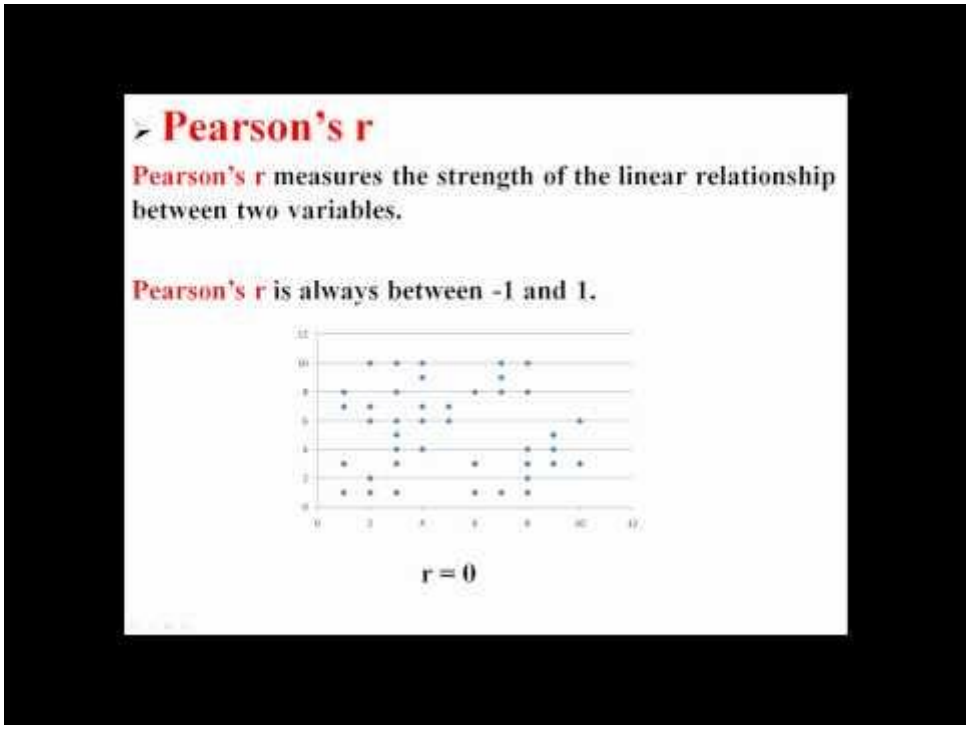


Figure 12.3: Pearson's r Correlation

12.1.2 Step 1: Look at the Data and State the Null and Research Hypotheses

Let's turn to a topic close to every parent's heart: sleep. The data set we'll use is fictitious but based on real events. Suppose we're curious to find out how much an infant's sleeping habits affect a parent's mood. Let's say that the parent rates their grumpiness very precisely, on a scale from 0 (not at all grumpy) to 100 (grumpy as a very, very grumpy old man or woman). Let's also assume that the parent has been measuring their grumpiness, their sleeping patterns, and their infant's sleeping patterns for quite some time now. Let's say, for 100 days. The main variables we will use are:

- dan.grump: the parent's grumpiness, measured from 0 to 100
- dan.sleep: the parent's sleep quality
- baby.sleep: the baby's sleep quality

The dataset also includes day, which records the day of the study from 1 to 100.

12.1.2.1 Data Set-Up

To conduct a Pearson correlation in a statistical program, the dataset needs two continuous variables. Each row should represent one participant, case, or unit of analysis, and each case should have a score on both variables.

In this example, each row represents one day. The variables dan.grump, dan.sleep, and baby.sleep are continuous variables.

	ID	dan.sleep	baby.sleep	dan.grump	day
1	1	7.59	10.18	56	1
2	2	7.91	11.66	60	2
3	3	5.14	7.92	82	3
4	4	7.71	9.61	55	4
5	5	6.68	9.75	67	5
6	6	5.99	5.04	72	6
7	7	8.19	10.45	53	7
8	8	7.19	8.27	60	8
9	9	7.40	6.06	60	9
10	10	6.58	7.09	71	10

Figure 12.4: Example data from 10 days with values for parent's sleep quality, baby's sleep quality, and parent's grumpiness.

Figure 12.2 *Example Dataset with the Parent's Sleep Quality, Baby's Sleep Quality, Parent's Grumpiness, and Day of Assessment*[3]

12.1.2.2 Describe the Data

For correlation, we should examine each variable individually and then examine the relation between the variables.

The descriptive statistics show the sample size, missing data, means, medians, standard deviations, minimum values, maximum values, skew, and kurtosis. These help us understand the scale and distribution of each variable.

We should also create scatterplots for the pairs of variables we plan to correlate, which helps us see the direction, form, strength, and possible outliers in the relation. This is important because a Pearson correlation summarizes a linear relation. If the relation is strongly curved or affected by an extreme outlier(s), the Pearson correlation may be misleading, and a more flexible option like a Spearman correlation should be used instead.

Descriptives

	dan.sleep	baby.sleep	dan.grump	day
N	100	100	100	100
Missing	0	0	0	0
Mean	6.97	8.05	63.71	50.50
Median	7.03	7.95	62.00	50.50
Standard deviation	1.02	2.07	10.05	29.01
Variance	1.03	4.30	101.00	841.67
Minimum	4.84	3.25	41	1
Maximum	9.00	12.07	91	100
Skewness	−0.30	−0.02	0.45	0.00
Std. error skewness	0.24	0.24	0.24	0.24
Kurtosis	−0.65	−0.61	−0.04	−1.20
Std. error kurtosis	0.48	0.48	0.48	0.48
Shapiro-Wilk W	0.98	0.98	0.98	0.95
Shapiro-Wilk p	0.069	0.256	0.122	0.002

Figure 12.5: Table including descriptive statistics for parent's sleep quality, baby's sleep quality, parent's grumpiness, and day of assessment, including sample size, percentage of missing values, mean, median, standard deviation, variance, minimum value, maximum value, skewness, standard error of skewness, kurtosis, Shapiro-Wilk W, and Shapiro-Wilk p.

Figure 12.3 *Descriptive Statistics for the Parent's Sleep Quality Example*[4]

12.1.2.3 Specify Your Hypotheses

For a Pearson correlation, the null hypothesis is that there is no linear relation between the two variables. The alternative hypothesis depends on whether the research question is directional or non-directional.

For a two-tailed correlation:

- H_0: There is no linear relation between the two variables. $(r = 0)$
- H_1: There is a linear relation between the two variables. $(r \neq 0)$

For a directional positive correlation:

- H_0: The variables are not positively related.$(r \leq 0)$
- H_1: The variables are positively related. $(r > 0)$

For a directional negative correlation:

- H_0: The variables are not negatively related. $r \geq 0$)
- H_1: The variables are negatively related. $(r < 0)$

In this example, we will focus on whether the parent's grumpiness is related to the parent's sleep quality and the baby's sleep quality. Because we are not specifying a direction before looking at the results, we will use two-tailed hypotheses.

Because we are examining more than one pair of variables, each correlation has its own hypothesis test.

12.1.3 Step 2: Check Assumptions and Set the Critical Value(s)

The Pearson correlation has several assumptions:

1. ***Continuous Variables***: Both variables are measured at the interval or ratio level, meaning they are continuous.
2. ***Independence***: The observations are paired and independent. Each case should have one score on each variable, and one case should not determine another case.
3. ***Linearity***: The relation between the two variables is ***linear***, a straight line. The data points should fall along an approximate straight-line pattern when plotted as $(X,\ Y)$ data points on a scatterplot.
4. ***Normality***: The variables are approximately normally distributed. This implies that there are more Y values scattered closer to the line than are scattered farther away.
5. ***Homoscedasticity***: The standard deviations of the population Y values about the line are equal for each value of X. In other words, each of these normal distributions of Y values has the same shape and spread about the line (homo = same, scedasticity = scatter). The relation should not be driven by extreme outliers.

We cannot test the first two assumptions using the output alone; those assumptions are based on how the data were measured and collected. However, we can evaluate normality, linearity, and outliers using descriptive statistics and graphs.

12.1.3.1 Testing Normality

For Pearson correlation, we evaluate the normality of both continuous variables included in the correlation. We can use any of these four normality checks: Shapiro-Wilk tests, Q-Q plots, skew and kurtosis z-scores, and visual inspection of each variable's distribution (histogram).

In this example, the main variables of interest are dan.grump, dan.sleep, and baby.sleep. The variable day is different because it simply records the day of the study from 1 to 100. It has a uniform distribution by design, so it is not useful for judging the normality of the main psychological variables.

Overall, the normality checks for the main variables do not raise major concerns.

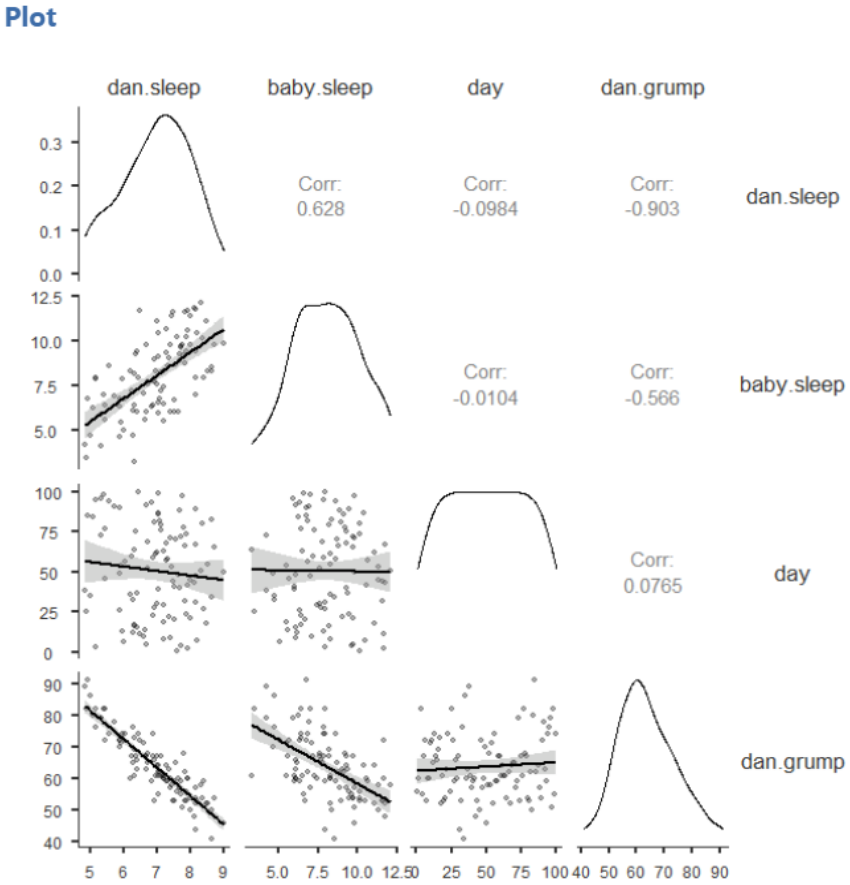


Figure 12.6: Normality checks for parent's sleep quality, baby's sleep quality, parent's grumpiness, and day.

Figure 12.4 *Normality Checks for All Variables in the Parent Sleep Quality Example*[5]

12.1.3.2 Testing Linearity and Looking for Outliers

To check linearity, examine a scatterplot for each pair of variables being correlated. The points should show a roughly straight-line pattern. The relation does not need to be perfect, but it should not show a strong curve, which is also called ***non-linear*** or, more specifically, ***curvilinear***.

We should also look for extreme outliers. An outlier can strongly affect a correlation, especially in small samples. If one unusual point appears to drive the relation, the correlation should be interpreted cautiously.

In our current example, the scatterplots do not suggest strong non-linear relations or extreme outliers, so the linearity assumption appears reasonable.

The image below illustrates the difference between linear and non-linear relations, including a curvilinear relationship.

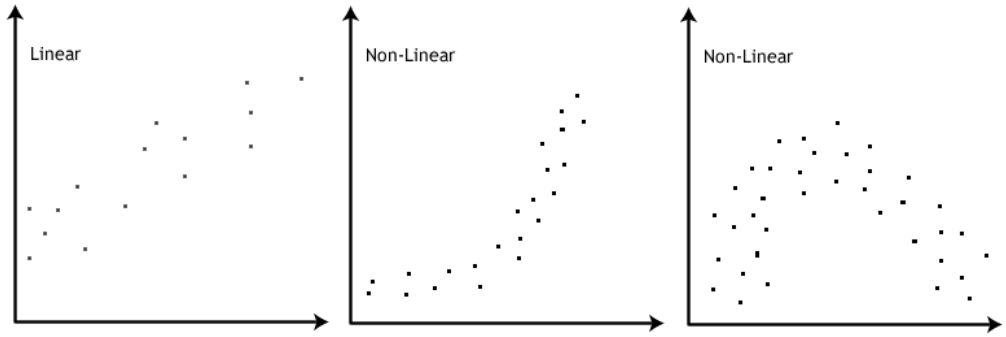


Figure 12.7: Three scatterplots demonstrating linear and non-linear correlations between variables.

Figure 12.5 *Scatterplots of Linear Versus Nonlinear Relations*[6]

12.1.3.3 Setting the Critical Value(s)

Once we know the sample size (N), degrees of freedom (d.f.), and probability level (), we use this information by looking at a probability table for the test we are using (in this case Pearson's r) to see what critical r-value(s) we will need to label a test as statistically significant. Please see Appendix A for links to probability tables. In this particular case, let's use a standard = .05 (which we will do all semester, unless your professor requires otherwise). We are fairly confident we know how these correlations will go: less sleep = more grumpy, more baby sleep = more parent sleep. Thus, we will use a one-tailed test, so we will have one critical value for

each of our correlations. For a sample size of 10, the degrees of freedom are $N - 2$, or 10-2 = 8. The critical value is .549 (or -.549 for sleep and grumpiness).

APPENDIX D: CRITICAL VALUES FOR PEARSON'S R (CORRELATION TABLE)

	LEVEL OF SIGNIFICANCE FOR ONE-TAILED TEST			
	.05	.025	.01	.005
	LEVEL OF SIGNIFICANCE FOR TWO-TAILED TEST			
df = n − 2	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.875
6	.621	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.715	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708

Figure 12.8: Table showing critical values for Pearson's r correlation coefficients based on sample size (df = n-2) and significance levels for one-tailed and two-tailed tests. Highlighted values include a circled significance level of 0.05 and a marked critical value of 0.549 for degrees of freedom of 8, indicating a threshold for statistical significance.

Figure 12.6 *Critical Values for Pearson's R, Sleep and Grumpiness Example*[7]

12.1.4 Step 3: Calculate and Report the Descriptive Statistics, Test Statistic, and Effect Size

If the variables are continuous, approximately normal, and linearly related, use the Pearson correlation. If one or both variables are ordinal, seriously non-normal, or better described by a monotonic rank-order relation, use Spearman's rank correlation instead. Kendall's tau is another rank-based correlation. It can be useful in some settings, especially with small samples or many tied ranks, but we will not use it in this course. For this example, we will be using Pearson's correlation coefficient. Depending on your professor, you might calculate the correlation by hand or using statistical software such as SPSS, SAS, R, Excel, or jamovi. Once your data are calculated, we will combine Step 3 with Step 4 below.

12.1.5 Step 4: Interpret Your Findings in Relation to the Hypothesis (Translate Math to English)

Once we are satisfied that the assumptions for Pearson correlation are reasonably met, we can interpret the results.

The degrees of freedom for a Pearson correlation are calculated as (n - 2). With 100 days of data, ($df = 98$).

Correlation Matrix

Correlation Matrix

		dan.sleep	baby.sleep	day	dan.grump
dan.sleep	Pearson's r	—			
	p-value	—			
baby.sleep	Pearson's r	0.63	—		
	p-value	< .001	—		
day	Pearson's r	−0.10	−0.01	—	
	p-value	0.330	0.918	—	
dan.grump	Pearson's r	−0.90	−0.57	0.08	—
	p-value	< .001	< .001	0.449	—

Figure 12.9: Correlation table for the variables parent's sleep quality, baby's sleep quality, parent's grumpiness, and day of assessment.

Figure 12.7 *Correlations Between All Variables in the Parent Sleep Quality Example*[8]

The ***correlation matrix*** is a table that shows the correlation coefficient and p-value for each pair of variables. The sign of the correlation tells us the direction of the relation, and the absolute value tells us the strength of the relation.

In this example, the parent's grumpiness is negatively correlated with the parent's sleep quality. This means that days with better parent sleep tend to be days with lower grumpiness. The parent's grumpiness is also negatively correlated with the baby's sleep quality. This means that days with better baby sleep also tend to be days with lower parent grumpiness.

The parent's sleep quality and the baby's sleep quality are positively correlated, meaning that days with better baby sleep tend to also be days with better parent sleep.

12.1.5.1 Write Up the Results in American Psychological Association (APA) Style

An APA-style results section should remind the reader of the research question, summarize the relevant descriptive statistics, report the inferential test and effect size, and interpret the result.

> Pearson correlations indicated that the parent's grumpiness was negatively related to the parent's sleep quality, $r(98) = -.90$, $p < .001$, and the baby's sleep quality, $r(98) = -.57$, $p < .001$. The parent's sleep quality was positively related to the baby's sleep quality, $r(98) = .63$, $p < .001$. All correlations showed large effects. When the baby gets better sleep, the parent also gets better sleep and reports feeling less grumpy.

This write-up focuses on the correlations. If the descriptive statistics are important for the research question, you should also report the means and standard deviations for each variable before reporting the correlations.

> Pearson correlations indicated that the parent's grumpiness ($M = 63.71$, $SD = 10.05$) was negatively related to the parent's sleep quality ($M = 6.97$, $SD = 1.02$), $r(98) = -.90$, $p < .001$, and the baby's sleep quality($M = 8.05$, $SD = 2.07$), $r(98) = -.57$, $p < .001$. The parent's sleep quality was positively related to the baby's sleep quality, $r(98) = .63$, $p < .001$. All correlations showed large effects. When the baby gets better sleep, the parent also gets better sleep and reports feeling less grumpy.

12.1.5.2 Visualize the Results

For correlation, the most useful visualization is usually a scatterplot. A ***scatterplot*** shows each case as a point, with one variable on the x-axis and the other variable on the y-axis. This helps readers see the direction, form, strength, and possible outliers in the relation.

A correlation matrix plot can be useful when we are examining several variables at once. However, when writing about a specific correlation, a single scatterplot is often clearer.

For this example, a scatterplot of dan.sleep and dan.grump would show the strong negative relation between the parent's sleep quality and grumpiness. A fitted line (also called a line of best fit or regression line) can help summarize the linear trend, but the points themselves are important because they show the actual data.

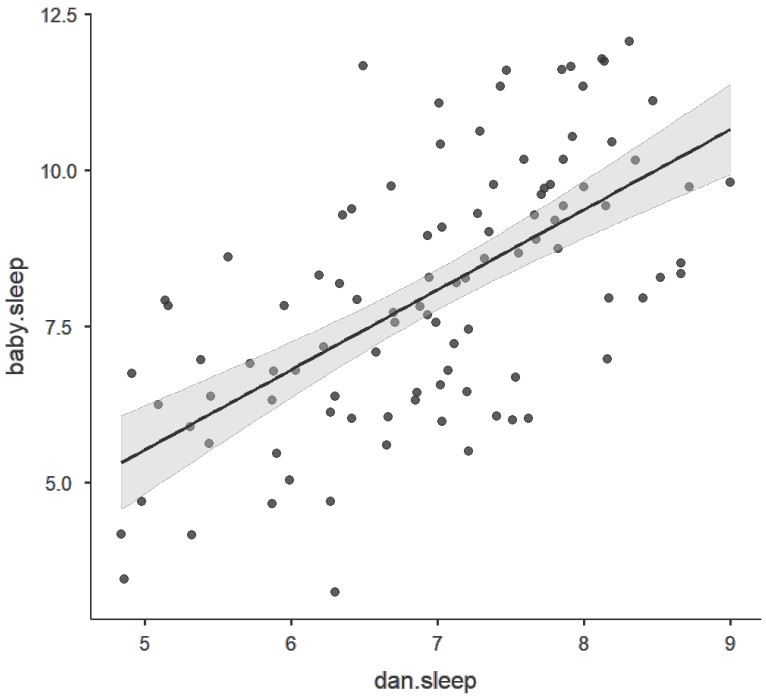


Figure 12.10: Scatterplot of the relation between parent's sleep quality and baby's sleep quality, which is positive.

Figure 12.8 *Scatterplot of Parent's Sleep Quality and Baby's Sleep Quality*[9]

12.1.5.3 Shared Variance

The variable R^2 is called the ***coefficient of determination*** and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- R^2, when expressed as a percent, represents the percent of variation in the dependent (predicted) variable Y that can be explained by variation in the independent (explanatory) variable X using the regression (best-fit) line.
- $1 - R^2$, when expressed as a percentage, represents the percent of variation in Y that is NOT explained by variation in X using the regression line. This can be seen as the scattering of the observed data points about the regression line.

For example, the correlation between the parent's grumpiness and the parent's sleep quality is (r = -.90). Squaring this value gives:

$$R^2 = (\text{-}.90)^2 = .81$$

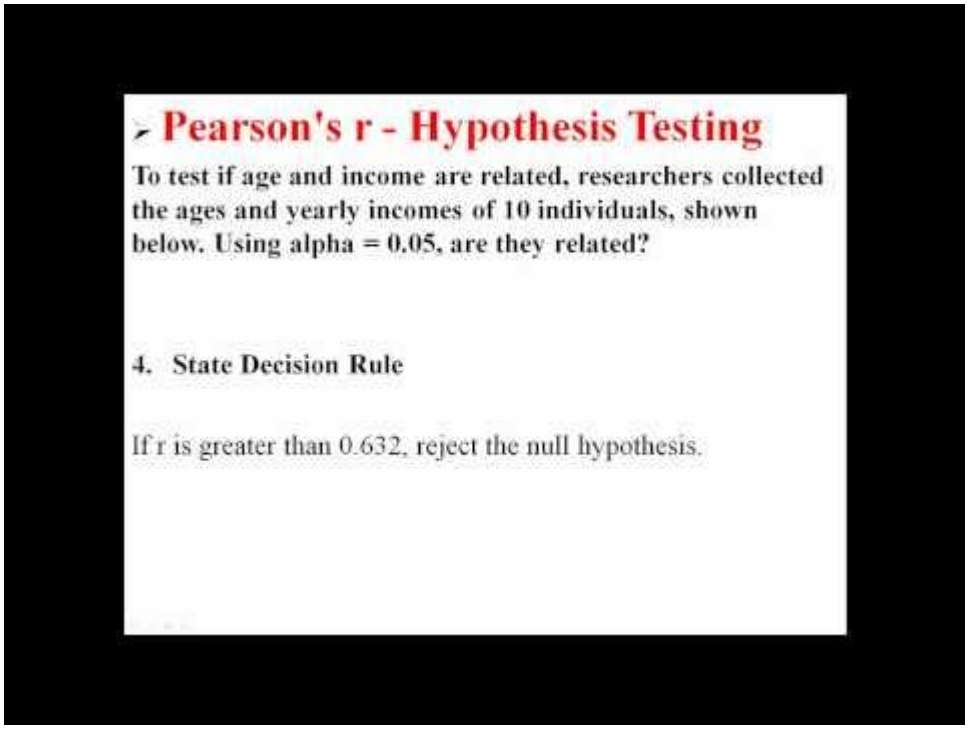
This means that approximately 81% of the variance in one variable is shared with the other variable. Because correlation does not establish causation, we should avoid saying that sleep quality "explains" or "causes" 81% of grumpiness unless the research design supports a causal interpretation.

Check Your Understanding

A correlation between social support and stress is $r = \text{-}.40$.

1. Is the relation positive or negative?
2. What is R^2?
3. Can we conclude that social support causes lower stress?

Please watch: Hypothesis Testing with Pearson's r (Statistics Lectures)



w ### Spearman Correlation

The Pearson correlation coefficient is pretty useful, but it does have shortcomings. One issue stands out: what it actually measures is the strength of the linear relation between two variables. In other words, what it gives you is a measure of the extent to which the data all tend to fall

on a single, perfectly straight line. Often, this is a pretty good approximation to what we mean when we say "relation", and so the Pearson correlation is a good thing to calculate. Sometimes though, it isn't.

One very common situation where the Pearson correlation isn't quite the right thing to use arises when an increase in one variable X really is reflected in an increase in another variable Y , but the nature of the relation isn't necessarily linear. An example of this might be the relation between effort and reward when studying for an exam. If you put zero effort (X) into learning a subject then you should expect a grade of 0% (Y).

However, a little bit of effort will cause a massive improvement. Just turning up to lectures means that you learn a fair bit, and if you just turn up to classes and scribble a few things down your grade might rise to 35%, all without a lot of effort. However, you just don't get the same effect at the other end of the scale. As everyone knows, it takes a lot more effort to get a grade of 90% than it takes to get a grade of 55%. What this means is that, if I've got data looking at study effort and grades, there's a pretty good chance that Pearson correlations will be misleading.

To illustrate, consider the data plotted below, showing the relation between hours worked and grade received for 10 students taking some class. The curious thing about this (highly fictitious) data set is that increasing your effort always increases your grade. It might be by a lot or it might be by a little, but increasing effort will never decrease your grade. If we run a standard Pearson correlation, it shows a strong relation between hours worked and grade received, with a correlation coefficient of $r = .91$.

However, this doesn't actually capture the observation that increasing hours worked always increases the grade. There's a sense here in which we want to be able to say that the correlation is perfect but for a somewhat different notion of what a "relation" is. What we're looking for is something that captures the fact that there is a perfect ***ordinal relation*** here. That is, if student 1 works more hours than student 2, then we can guarantee that student 1 will get a better grade. That's not what a correlation of $r = .91$ says at all.

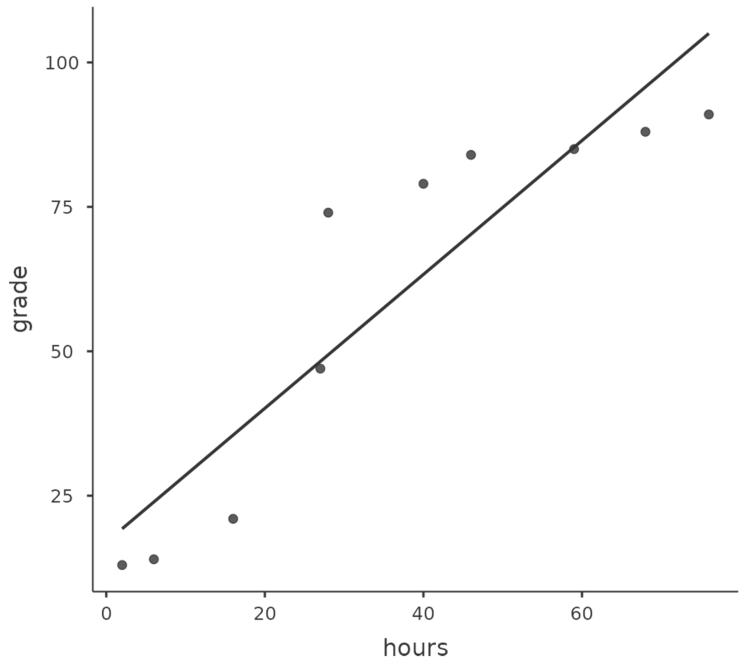


Figure 12.11: Plot showing the relationship between hours worked and grade received for a toy data set consisting of only 10 students (each dot corresponds to one student). The line through the middle shows the linear relationship between the two variables. This produces a strong Pearson correlation of . However, the interesting thing to note here is that there's actually a perfect monotonic relationship between the two variables. In this toy example, increasing the hours worked always increases the grade received, as illustrated by the solid line. This is reflected in a Spearman correlation of . With such a small data set, however, it's an open question as to which version better describes the actual relationship involved

Figure 12.9 *Scatterplot of Hours Worked and Grade Received*[10]

How should we address this? Actually, it's really easy. If we're looking for ordinal relations all we have to do is treat the data as if it were an ordinal scale! So, instead of measuring effort in terms of "hours worked", let's rank all 10 of our students in order of hours worked. That is, student 1 did the least work out of anyone (2 hours) so they get the lowest rank (rank = 1). Student 4 was the next, putting in only 6 hours of work over the whole semester, so they get the next lowest rank (rank = 2).

Notice that I'm using "rank =1" to mean "low rank". Sometimes in everyday language we talk about "rank = 1" to mean "top rank" rather than "bottom rank". So be careful, you can rank "from smallest value to largest value" (i.e., small equals rank 1) or you can rank "from largest value to smallest value" (i.e., large equals rank 1). In this case, I'm ranking from smallest to

largest, but as it's really easy to forget which way you set things up you have to put a bit of effort into remembering!

Okay, so let's have a look at our students when we rank them from worst to best in terms of effort and reward.

Table 12.2 *Students Ranked in Terms of Effort and Reward*[11]

	Rank (Hours Worked)	Rank (Grade Received)
student 1	1	1
student 2	10	10
student 3	6	6
student 4	2	2
student 5	3	3
student 6	5	5
student 7	4	4
student 8	8	8
student 9	7	7
student 10	9	9

Hmm. These are identical. The student who put in the most effort got the best grade, the student with the least effort got the worst grade, etc. As the table above shows, these two rankings are identical, so if we now correlate them we get a perfect relation, with a correlation of 1.0.

What we've just re-invented is Spearman's rank order correlation, usually denoted *(rho)* to distinguish it from the Pearson correlation, r. ***Spearman's correlation*** is the rank-based alternative to Pearson correlation. Use Spearman's correlation when the variables are ordinal, when the normality assumption is seriously violated, and/or when the relation is monotonic but not well described as linear. A ***monotonic relation*** means that as one variable increases, the other variable tends to increase or tends to decrease, but the pattern does not have to form a straight line.

A Spearman correlation is reported similarly to Pearson correlation and can be designated as (rho) or r_s.

> A Spearman correlation indicated that the two variables were significantly related, $= .42$, $p = .018$.

Please watch: Spearman Correlation (Statistics Lectures)

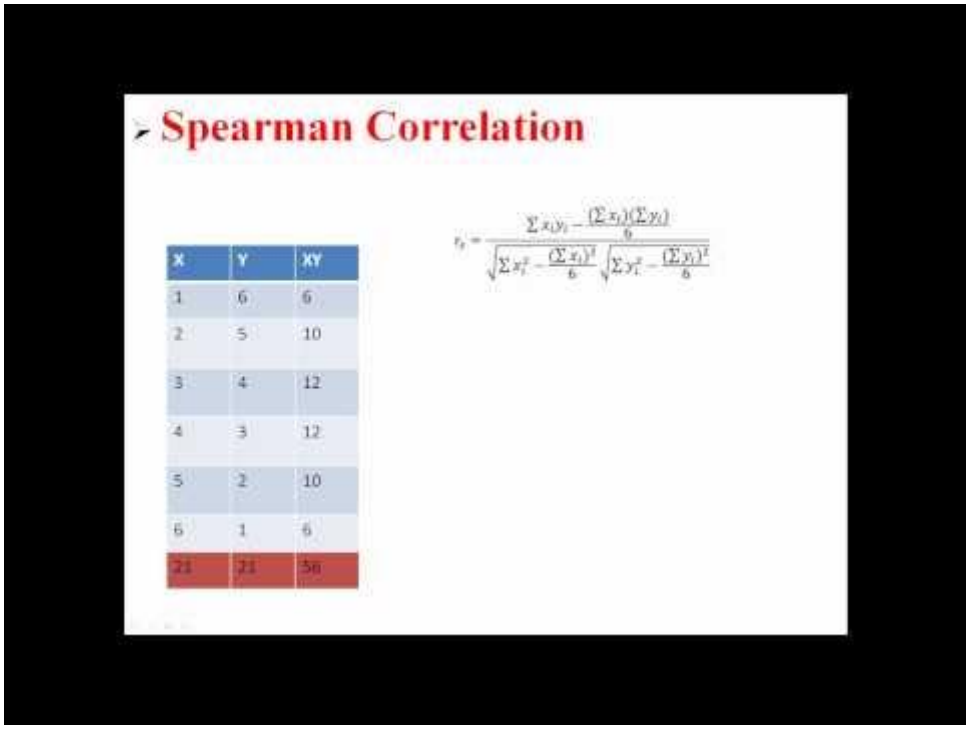


Figure 12.12: Spearman Correlation

12.2 Regression

Regression is used when we want to predict or explain variation in one continuous outcome variable using one or more predictor variables. Unlike t-tests or ANOVA (see Chapters 12 and 14), which focuses on categorical predictors, regression can include continuous predictors, categorical predictors, or both. In this course, we will focus on regression models with a continuous outcome variable.

Regression is directional in its setup. We identify one variable as the ***outcome variable*** and one or more variables as ***predictor variables***. This does not automatically mean the predictors cause the outcome. Causal conclusions depend on the research design, measurement quality, and possible alternative explanations.

There are several regression approaches you may encounter:

- ***Simple linear regression*** uses one predictor variable to predict one continuous outcome variable. When the predictor is continuous, simple linear regression is closely related to correlation.

- ***Multiple regression*** uses two or more predictors to predict one continuous outcome variable. Predictors may be continuous, categorical, or a combination of both.
- ***Hierarchical regression*** uses two or more blocks of predictors. This allows us to test whether adding a new block of predictors improves the model beyond the predictors already included.

Regression is also connected to many of the statistical tests we will learn. A regression with one dichotomous predictor produces results that are closely related to an independent t-test. A regression with one categorical predictor with three or more levels is closely related to ANOVA. A regression with one continuous predictor is closely related to correlation.

12.2.1 Understanding Regression

A ***linear regression model*** is based on a line. If you took geometry, you may remember the equation of a line as:

$y = mx + b$

where y = the outcome variable value, m = the slope for the predictor variable x, and b = the intercept or value of y when $x = 0$.

In regression, the same general idea applies. The model estimates a predicted outcome score from one or more predictors. The intercept is the predicted outcome value when all predictors are 0. The slope or coefficient tells us how much the predicted outcome changes when a predictor increases by one unit, holding the other predictors constant.

With one predictor, a regression equation can be written as:

$y = b_0 + b_1 x$

With two predictors, a regression equation can be written as:

$y = b_0 + b_1 x_1 + b_2 x_2$

The model does not perfectly predict every data point. The difference between the observed outcome and the predicted outcome is called a ***residual***.

Let's imagine we have a dataset of dragons. We want to predict each dragon's weight using two predictors: whether the dragon is spotted and the dragon's height. In this example, the outcome variable is weight. The predictors are spotted status and height.

The regression equation might be:

$y = 2.4 + 0.6 x_1 + 0.3 x_2$

The intercept, 2.4, is the predicted weight for a dragon that is not spotted and has a height of 0. This is needed for the equation, but it may not be substantively meaningful because a dragon with a height of 0 is not realistic. The coefficient for spotted (x_1), .6, means that

spotted dragons are predicted to weigh .6 tons more than non-spotted dragons with the same height. The coefficient for height (x_2), .3, means that for each one-unit increase in height, predicted weight increases by .3 tons, holding spotted status constant.

Figure 12.13: Two dragons next to a regression equation.

Figure 12.10 *Example Regression Equation Predicting Dragon Weight from Spots*[12]

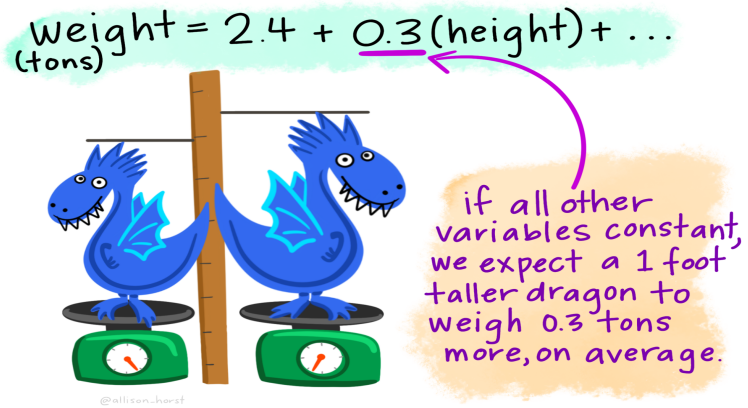


Figure 12.14: Two dragons next to a regression equation.

Figure 12.11 *Example Regression Equation Predicting Dragon Weight from Height*[13]

Regression finds the line or model that makes the residuals as small as possible. In other words, the model tries to make the predicted values as close as possible to the observed values.

The images below show the difference between a model that fits the data well and a model that fits poorly. When the residuals are smaller, the model is closer to the observed data points.

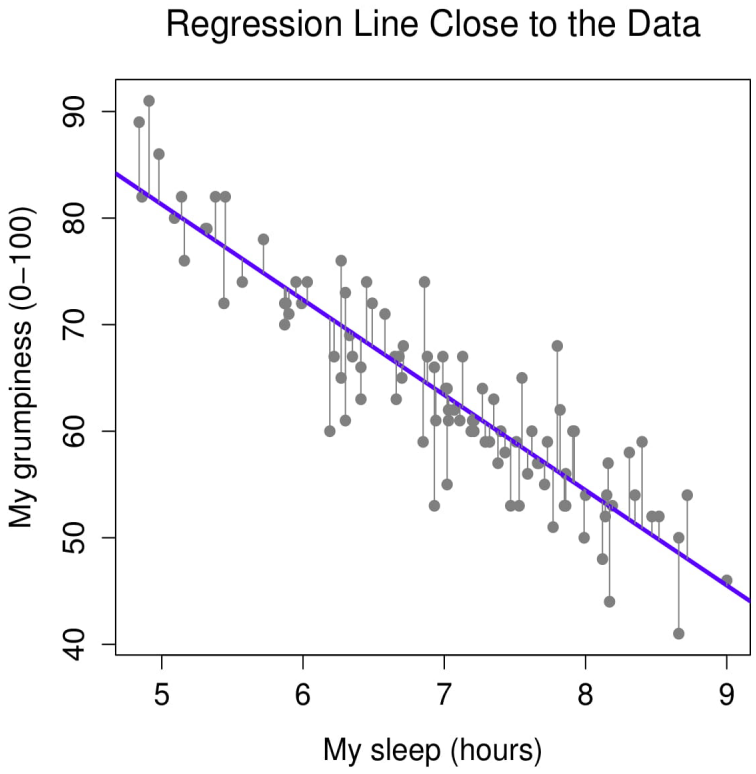


Figure 12.15: A regression line showing a good or close fit to the data with small residual values.

Figure 12.12 *Regression Line Predicting Grumpiness from Hours Slept with Close Fit to the Data*[14]

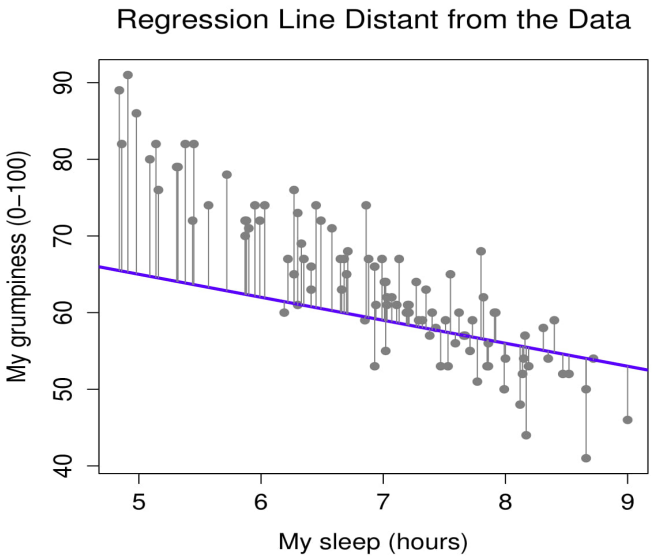


Figure 12.16: A regression line showing a poor or distant fit to the data with large residual values.

Figure 12.13 *Regression Line Predicting Grumpiness from Hours Slept with Distant Fit to the Data*[15]

Let's return to the dragon example. Suppose a dragon is not spotted and has a height of 5.1. Based on the regression model, we would predict the dragon to weigh 3.9 tons. If the dragon actually weighs 4.2 tons, the residual is .3 tons.

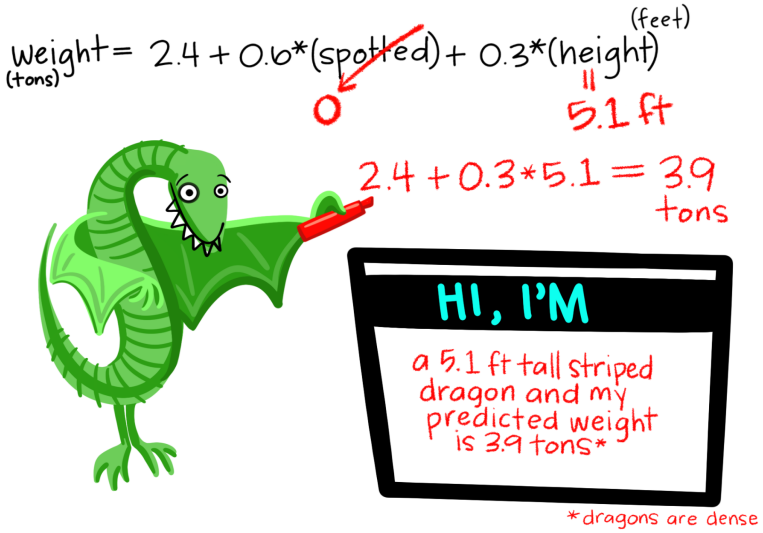


Figure 12.17: A dragon standing next to a regression equation.

Figure 12.14 *Example Regression Equation Predicting Dragon Weight from Spots and Height*[16]

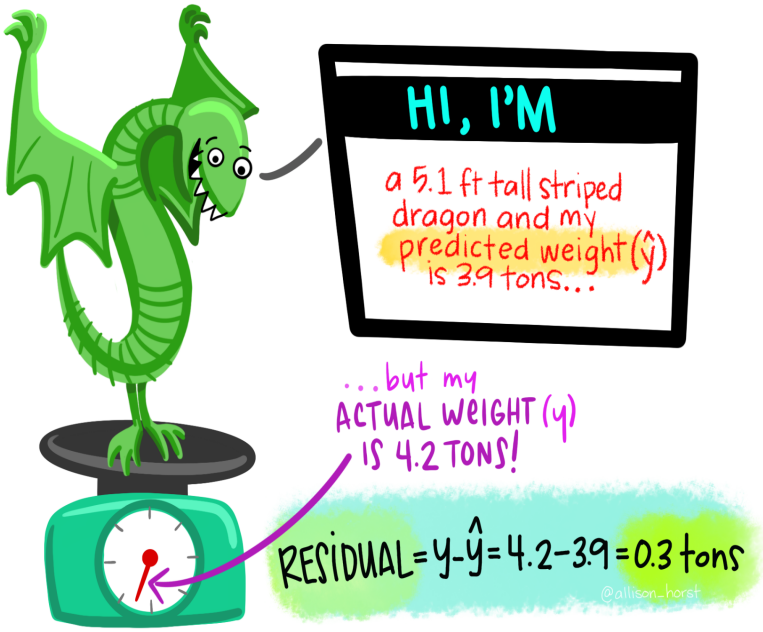


Figure 12.18: A dragon standing on a scale with their residual equation.

Figure 12.15 *Residual Example for Predicting Dragon Weight from Spots and Height*[17]

One of the assumptions we check in regression is whether the residuals are approximately normally distributed. This means we examine the distribution of prediction errors, not just the original outcome variable.

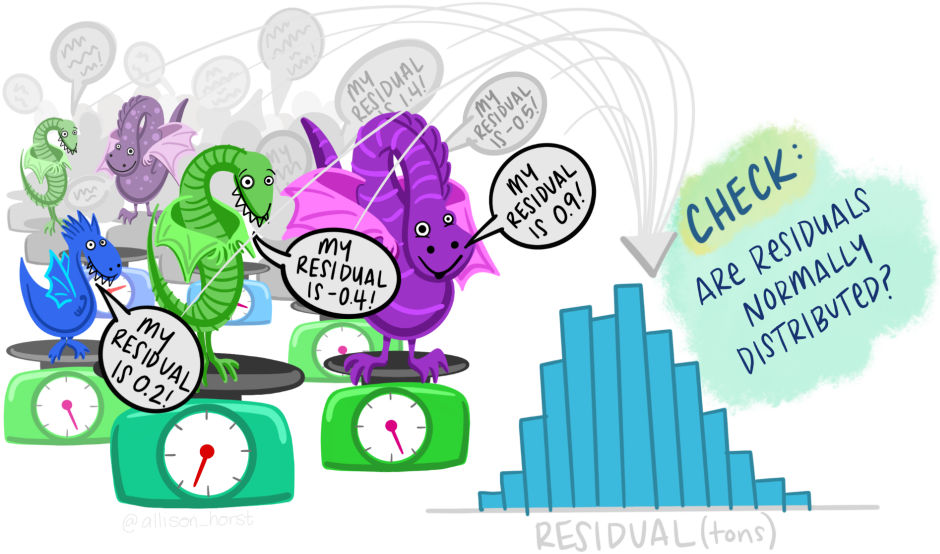


Figure 12.19: Several dragons standing on scales next to a normal distribution.

Figure 12.16 *Checking that Your Residuals are Normally Distributed is Important*[18]

> **Note**
>
> Regression includes more setup than many of the earlier tests because the model gives us several pieces of information: overall model fit, individual predictor coefficients, predicted values, residuals, and assumption checks. The goal is not to memorize every detail at once. Focus first on identifying the outcome, identifying the predictors, and interpreting the overall model and coefficients.

Check Your Understanding

A researcher uses hours studied and prior GPA to predict final exam score.

1. What is the outcome variable?
2. What are the predictor variables?

3. Why should the researcher avoid automatically saying that hours studies and prior GPA caused the exam scores?

12.2.2 Step 1: Look at the Data and State the Null and Alternative Hypotheses

We will return to the parenthood dataset. This dataset includes 100 days of data about Dan's sleep quality, the baby's sleep quality, Dan's grumpiness, and the day of the study from 1 to 100.

The main variables are:

- dan.grump: Dan's grumpiness
- dan.sleep: Dan's sleep quality
- baby.sleep: the baby's sleep quality
- day: the day of the study, from 1 to 100

For the first regression model, our research question is: Do Dan's sleep quality and the baby's sleep quality predict Dan's grumpiness?

12.2.2.1 Data Set-Up

For regression, the dataset needs one continuous (interval or ratio) outcome variable and one or more predictor variables. Predictors may be continuous or categorical, depending on the model. Each row should represent one participant, case, day, or other unit of analysis.

In this example, each row represents one day. The continuous outcome variable is dan.grump. The first two predictor variables are dan.sleep and baby.sleep. The variable day will be used later to demonstrate categorical predictors, but it is not included in the first multiple regression model.

Check Your Understanding

Look at the regression example.

1. What is the continuous outcome variable in the first regression model?
2. What are the predictor variables in the first regression model?
3. What does each row of the dataset represent?

12.2.2.2 Describe the Data

The descriptive statistics show that there are 100 cases and no missing data. The variables dan.grump, dan.sleep, and baby.sleep are the main psychological variables in the regression model. The variable day is different because it records the day of the study from 1 to 100. It has a uniform distribution by design rather than a normal distribution.

For regression, we should also examine scatterplots of the continuous predictors with the outcome. Scatterplots help us see whether the relations look roughly linear and whether there are unusual points that may strongly influence the model.

12.2.3 Step 2: Check Assumptions and Set the Critical Value(s)

Regression has several assumptions. Some assumptions are based on the research design and how the data were measured. Other assumptions are evaluated using model diagnostics.

12.2.3.1 Design and Data Assumptions

Before interpreting the regression output, we should confirm that the basic design and variable assumptions are reasonable:

1. The outcome variable is continuous.
2. The predictors are continuous or appropriately coded categorical variables.
3. The observations are independent. Each case should contribute one outcome value, and one case should not determine another case.

In this example, dan.grump is continuous, and the first two predictors, dan.sleep and baby.sleep, are also continuous. Each row represents a day, so we should think about whether the days can reasonably be treated as independent. Because these are repeated daily observations from the same person, this assumption may be imperfect. For this course example, we will proceed with the regression, but in a more advanced course you might learn models designed specifically for repeated daily data.

12.2.3.2 Outliers and Influential Cases

Regression can be strongly affected by unusual cases. One way to check for influential cases is Cook's distance. Cook's distance examines whether one entire row of data has a large influence on the regression model. A common rule of thumb is that Cook's distance values greater than 1 may indicate a highly influential case. In this example, the Cook's distance values are small, so this check does not raise concern.

Data Summary

Cook's Distance

			Range	
Mean	Median	SD	Min	Max
0.01	0.00	0.02	2.62e−5	0.11

Figure 12.20: A data output table with Cook's Distance values displayed.

Figure 12.17 *Cook's Distance is Used for Checking Influential Cases*[19]

If Cook's distance suggests an influential case, do not automatically delete it. First, investigate the case. Check whether it reflects a data-entry error, a measurement problem, or a legitimate but unusual observation. Any decision to remove a case should be justified and reported.

12.2.3.3 Normality of the Residuals

In regression, the key normality assumption is about the residuals. The residuals are the differences between the observed outcome values and the predicted outcome values.

We can evaluate residual normality using the Shapiro-Wilk test and the Q-Q plot of the residuals. In this example, the Shapiro-Wilk test is not statistically significant, and the Q-Q plot looks reasonable. Therefore, the residual normality assumption appears reasonable.

Large samples can make the Shapiro-Wilk test statistically significant even when the Q-Q plot looks acceptable. In that case, use multiple pieces of evidence rather than relying on one test alone.

Normality Test (Shapiro-Wilk)

Statistic	p
0.99	0.841

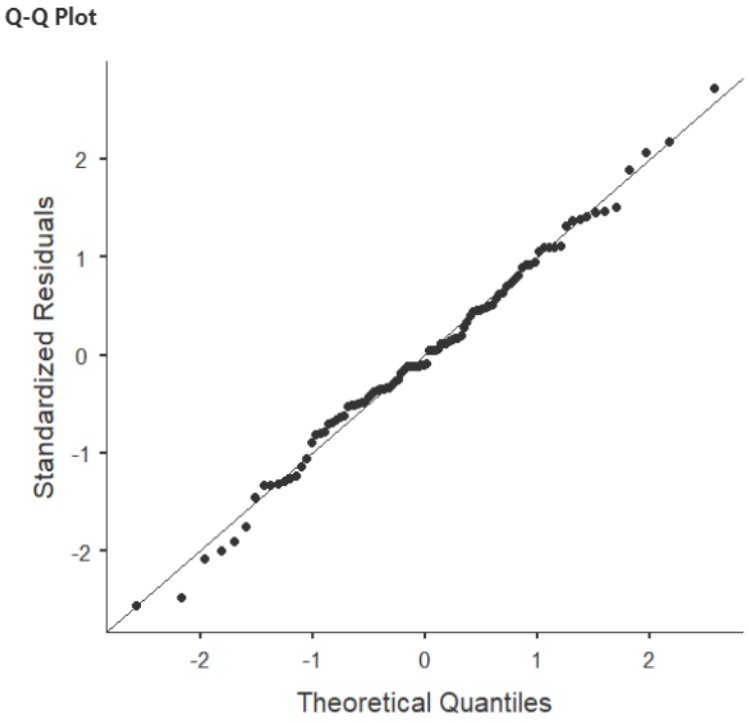


Figure 12.21: A data output table of Shapiro-Wilk's value for normality and a q-q plot.

Figure 12.18 *Shapiro-Wilk Test and Q-Q Plot*[20]

12.2.3.4 Linearity and Homoscedasticity

Linearity means that the relation between each continuous predictor and the outcome is roughly a straight line. If the relation is strongly curved, a linear regression model may not describe the data well.

Homoscedasticity means that the spread of the residuals is roughly similar across the range of predicted values. If the residuals fan out or become much narrower across the plot, the assumption may be violated.

We can examine these assumptions using the residual plot. The residuals should be scattered fairly randomly around 0, without a strong curve or funnel shape. In this example, the residual plot does not raise major concerns.

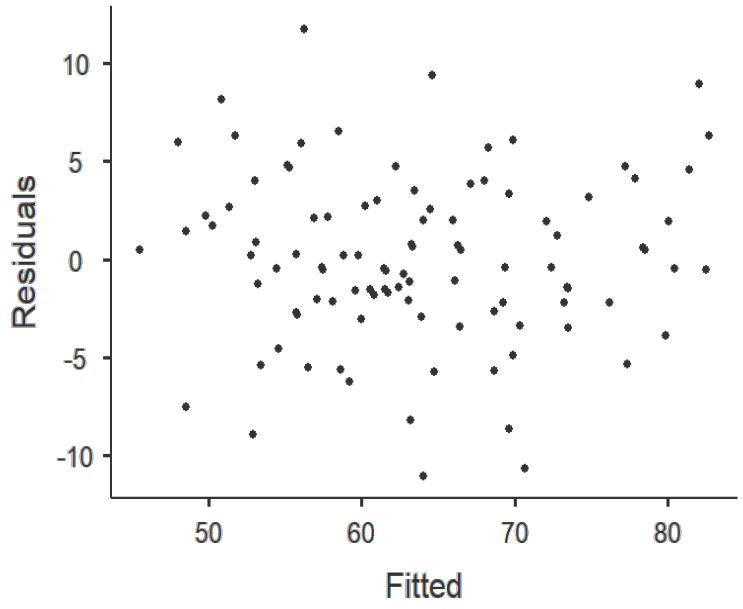


Figure 12.22: A residual scatterplot.

Figure 12.19 *Scatterplot of Residuals*[21]

12.2.3.5 Independence of Residuals

The independence assumption means that residuals should not be strongly related to one another. For example, in time-ordered data, one day's residual may be related to the next day's residual. The Durbin-Watson test can be used to check for autocorrelation in residuals.

Values near 2 suggest that residual autocorrelation is not a major concern. Values much closer to 0 or 4 may raise concern. In this example, the Durbin-Watson value does not raise major concern for the purposes of this example.

12.2.3.6 Multicollinearity

Multicollinearity occurs when predictors are very strongly related to each other. When predictors overlap too much, it becomes difficult to estimate the unique contribution of each predictor.

We can evaluate multicollinearity using tolerance and the variance inflation factor, or VIF. Tolerance values below .10 or VIF values above 10 are common warning signs. In this example, multicollinearity does not raise major concern.

12.2.4 Step 3: Calculate and Report the Descriptive Statistics, Test Statistic, and Effect Size

For this course, use linear regression when the outcome variable is continuous, the predictors are continuous or appropriately coded categorical variables, and the assumptions appear reasonable enough to proceed. If assumption checks raise serious concerns, note the concern and interpret the model cautiously. We will blend Steps 3 and 4 in reporting and interpreting the analyses.

12.2.5 Step 4: Interpret Your Findings in Relation to the Hypothesis (Translate Math to English)

The regression output gives us information about the overall model and the individual predictors.

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.90	0.82	0.81	215.24	2	97	< .001

Model Coefficients - dan.grump

Predictor	Estimate	SE	t	p	Stand. Estimate
Intercept	125.97	3.04	41.42	< .001	
dan.sleep	−8.95	0.55	−16.17	< .001	−0.90
baby.sleep	0.01	0.27	0.04	0.969	0.00

Figure 12.23: A data table output for a regression analysis.

Figure 12.20 *Linear Regression Output for Parent Sleep Quality Example*[22]

12.2.5.1 Overall Model Fit

The model fit table includes R, R^2, adjusted R^2, and the overall F test.

$\boldsymbol{R^2}$ is the proportion of variance in the outcome variable accounted for by the predictors in the model. In this example, Dan's sleep quality and the baby's sleep quality account for about 82% of the variance in Dan's grumpiness.

Adjusted $\boldsymbol{R^2}$ is similar to R^2, but it adjusts for the number of predictors in the model. Adding predictors will usually increase R^2, even if the added predictors are not very useful. Adjusted R^2 provides a more cautious estimate of model fit. This adjustment is an attempt to take the degrees of freedom into account. The big advantage of the adjusted R^2 value is that when you add more predictors to the model, the adjusted R^2 value will only increase if the new variables improve the model performance more than you would expect by chance.

The ***omnibus (overall) F-test*** tells us whether the set of predictors significantly predicts the outcome variable. In this example, the overall model is statistically significant, which means that Dan's sleep quality and the baby's sleep quality together predict Dan's grumpiness better than a model with no predictors.

12.2.5.2 Model Coefficients

The coefficients table gives us a separate test for each predictor. The ***unstandardized coefficient***, often written as b, tells us how much the predicted outcome changes for a one-unit increase in the predictor, holding the other predictors constant. The ***standardized coefficient***, often written as (beta), puts predictors on a common scale so their relative strength can be compared more easily.

In this example, dan.sleep significantly predicts dan.grump. As Dan's sleep quality increases, predicted grumpiness decreases, holding the baby's sleep quality constant. The baby's sleep quality does not significantly predict Dan's grumpiness after accounting for Dan's sleep quality.

The ***intercept*** is needed to write the prediction equation, but it is often not substantively meaningful. In this example, the intercept is the predicted grumpiness score when Dan's sleep quality and the baby's sleep quality are both 0. If 0 is outside the meaningful range of the predictors, the intercept should not be overinterpreted.

The unstandardized regression equation from this model is:

$y = 125.97 - 8.95x_1 + .01x_2$

If Dan's sleep quality was 5 and the baby's sleep quality was 8, the predicted grumpiness score would be:

$y = 125.97 - 8.95(5) + .01(8) = 81.30$

12.2.5.3 Write Up the Results in APA Style

An APA-style results section should remind the reader of the research question, report the model fit, report the relevant coefficients, and interpret the result.

> A multiple regression was conducted to examine whether Dan's sleep quality and the baby's sleep quality predicted Dan's grumpiness across 100 days. The overall model was statistically significant, $F(2, 97) = 215.24$, $p < .001$, adjusted $R^2 = .81$. Dan's sleep quality significantly predicted grumpiness, (b = -8.95, SE = .55, ☒ = -.90), $t(97)$ = -16.17, $p < .001$, such that days with better Dan sleep were associated with lower grumpiness. The baby's sleep quality did not significantly predict Dan's grumpiness after accounting for Dan's sleep quality, (b = .01, SE = .27, ☒ = .00), $t(97)$ = .04, p = .969.

In many research reports, assumption checks are described in the analysis plan or results section only when there is a concern.

12.2.6 Categorical Predictors

Regression can include categorical predictors. When a categorical predictor has two levels, it can be represented with a 0/1 code. The group coded 0 is the reference group, and the coefficient for the predictor represents the difference between the group coded 1 and the group coded 0, holding other predictors constant.

When a categorical predictor has more than two levels, the model needs multiple comparison variables.

The parenthood dataset does not include a categorical predictor that we need for the main regression example, so we will create one for demonstration purposes. We will transform the day variable into a new categorical variable with three groups:

- days 1-32,
- days 33-65,
- days 66-100.

This lets us demonstrate how regression handles categorical predictors with more than two levels.

12.2.6.1 Reference Groups

For categorical predictors, the reference group is the group that other groups are compared to. With reference level or dummy coding, the intercept is the predicted outcome for the reference group when all continuous predictors are 0. In this example, if group 1 is the reference group, the intercept is the predicted grumpiness score for days 1-32 when Dan's sleep quality and the baby's sleep quality are both 0.

The coefficient for day_3groups (2 - 1) compares group 2 to group 1, holding the other predictors constant. The coefficient for day_3groups (3 - 1) compares group 3 to group 1, holding the other predictors constant.

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.90	0.82	0.81	107.29	4	95	< .001

Model Coefficients - dan.grump

Predictor	Estimate	SE	t	p	Stand. Estimate
Intercept [a]	126.02	3.16	39.83	< .001	
dan.sleep	−8.87	0.56	−15.72	< .001	−0.90
baby.sleep	−0.02	0.27	−0.06	0.949	−0.00
day - 3_Groups:					
2 – 1	−1.12	1.08	−1.03	0.305	−0.11
3 – 1	−0.03	1.08	−0.03	0.978	−0.00

[a] Represents reference level

Figure 12.24: A data table output for a regression analysis with a categorical predictor.

Figure 12.21 *Linear Regression Output with a Categorical Predictor*[23]

In this example, neither comparison is statistically significant. This means that Dan's predicted grumpiness does not differ significantly between days 33-65 and days 1-32, or between days 66-100 and days 1-32, after accounting for Dan's sleep quality and the baby's sleep quality.

Check Your Understanding

A regression model includes a categorical predictor called condition coded 0 = control and 1 = treatment. The coefficient for condition is positive and statistically significant. What doe sthis mean?

12.2.7 Hierarchical Regression

Hierarchical regression is a form of multiple regression in which predictors are entered in blocks or steps. The blocks should be based on theory, prior research, temporal order, or the research question. Hierarchical regression is not just a way to try predictors in different orders until the results look best.

Hierarchical regression allows us to ask whether a new block of predictors improves prediction beyond the predictors already in the model. This is called ***incremental prediction***.

For this example, we will use two models:

- Model 1: baby.sleep
- Model 2: baby.sleep and dan.sleep

This lets us test whether Dan's sleep quality improves prediction of Dan's grumpiness above and beyond the baby's sleep quality. The output now includes model fit information for each model and a model comparison table. The model comparison table tells us whether the added block of predictors significantly improves the model.

Model Fit Measures

Model	Adjusted R^2	AIC	BIC	Overall Model Test			
				F	df1	df2	p
1	0.31	711.68	719.49	46.18	1	98	< .001
2	0.81	582.95	593.37	215.24	2	97	< .001

Model Comparisons

Comparison			ΔR^2	F	df1	df2	p
Model		Model					
1	-	2	0.50	261.52	1	97	< .001

Figure 12.25: Image of an SPSS table presenting statistical model fit measures and comparisons, showing two models with adjusted R^2, AIC, BIC, F-values, degrees of freedom, and p-values. Model 2 demonstrates a stronger fit with higher adjusted R^2 (0.81) and lower AIC/BIC values, while model comparison indicates a significant improvement ($\Delta R^2 = 0.50$, $p < .001$).

Figure 12.22 *Hierarchical Regression Output for Parent Sleep Quality Example*[24]

Model 1, which includes only baby.sleep, is statistically significant, $F(1, 98) = 46.18$, $p < .001$, adjusted $R^2 = .31$. Model 2, which includes both baby.sleep and dan.sleep, is also statistically significant, $F(2, 97) = 215.24$, $p < .001$, adjusted $R^2 = .81$.

The model comparison shows that adding dan.sleep significantly improves the model, $\Delta F(1, 97) = 261.52$, $p < .001$, $R^2 = .50$. This means that Dan's sleep quality explains additional variance in Dan's grumpiness beyond the variance explained by the baby's sleep quality.

12.2.7.1 Write Up Hierarchical Regression in APA Style

A hierarchical regression was conducted to examine whether Dan's sleep quality predicted Dan's grumpiness above and beyond the baby's sleep quality. In Model 1, the baby's sleep quality significantly predicted Dan's grumpiness, $F(1, 98) = 46.18$, $p < .001$, adjusted $R^2 = .31$. Higher baby sleep quality was associated with lower Dan grumpiness, b = -2.71, SE = .40, ☒ = -.57, $t(98)$ = -6.80, $p < .001$.

In Model 2, Dan's sleep quality was added to the model. Model 2 significantly predicted Dan's grumpiness, $F(2, 97) = 215.24$, $p < .001$, adjusted $R^2 = .81$. Adding Dan's sleep quality significantly improved model fit, $\Delta F(1, 97) = 261.52$, $p < .001$,

R^2 = .50. In Model 2, Dan's sleep quality significantly predicted grumpiness, b = -8.95, SE = .55, β = -.90, $t(97)$ = -16.17, $p < .001$. The baby's sleep quality did not significantly predict Dan's grumpiness after accounting for Dan's sleep quality, b = .01, SE = .27, β = .00, $t(97)$ = .04, p = .969.

12.2.8 Other Uses for Regression

Beyond simple, multiple, and hierarchical regression equations, the regression framework can also be used to examine additional variables of interest in unique ways. Two common variations are moderation and mediation models.

12.2.8.1 Moderation

Moderation involves examining when, for whom, or under what specific conditions two variables are related. For example, let's say researchers have found that there is a positive relation between height (X) and attractiveness (Y). However, they have identified that this relation depends on an additional factor – the biological sex of participants. For men, the relation between height and attractiveness is large and positive, while for women, the relation is near zero. In this case, we would say that the relation between height and attractiveness depends on the level of biological sex. In other words, biological sex is a moderator of this relation.

To conduct a moderation analysis, researchers can use the regression framework. In our example above, the outcome variable would be attractiveness (Y), and the predictor variables would be:

- Height (X_1)
- Biological sex (X_2)
- Height x biological sex (the interaction or moderator term)

The moderator is statistically represented by multiplying the two predictor variables together as a third predictor variable in the statistical equation. Moderators can be continuous or categorical variables. Common moderators in psychological research focus on demographic variables, such as biological sex, race/ethnicity, religion, and SES.

Moderation can also be examined using the PROCESS macro, though this is beyond the scope of this course.

12.2.8.2 Mediation

Mediation involves examining why two variables are related. For example, researchers have often found that there is a positive connection between playing violent video games (X) and aggressive behaviors (Y) in children. In wondering why these two variables are related, researchers may want to test how desensitization to violence mediates this relation. In other words, perhaps as children play violent video games, they become more desensitized to violence, which then predicts their aggressive behavior. In this case, we would say that the mediator or explanatory variable in this relation is desensitization to violence. If playing violent video games predicts aggressive behaviors by itself, this is called a ***direct effect.*** If playing violent video games predicts desensitization, which then predicts aggressive behaviors, this is called an ***indirect effect***.

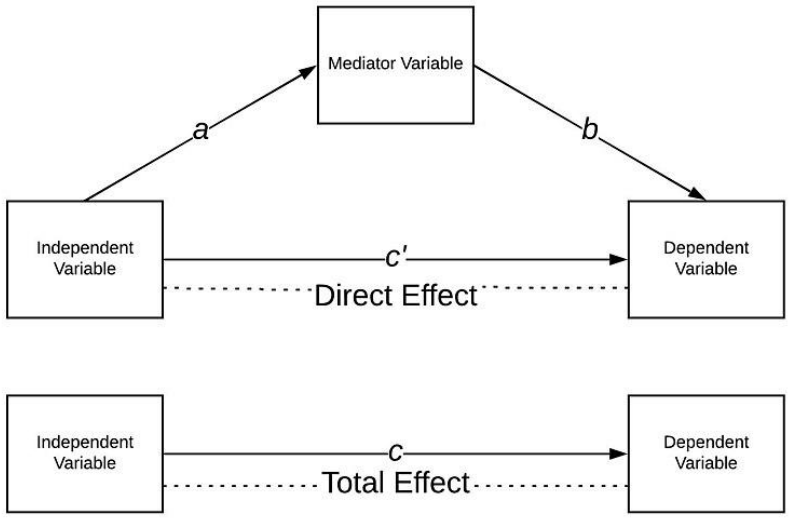


Figure 12.26: Diagram illustrating mediation analysis with labeled paths between variables. It shows independent variable affecting dependent variable directly (path c') and indirectly through mediator variable (paths a and b), with total effect (path c) represented separately.

Figure 12.23 *Mediated Prediction Example, With Indirect Effect Shown Through the Mediator Variable*[25]

To conduct a mediation analysis, researchers can use the regression framework to demonstrate a series of mediating steps outlined by Baron and Kenny (1986):

1. Test the relation between the predictor and outcome variables.
2. Test the relation between the predictor and mediator variables.
3. Test the relation between the mediator and outcome variables.
4. Test if the relation between the predictor and outcome variables diminishes when controlling for the mediator variable.
5. Ensure temporal ordering with the predictor variable measured first, the mediator measured second, and the outcome variable measured last.

While Baron and Kenny's steps are often still cited, most researchers use the PROCESS macro or structural equation modelling to analyze mediation models. These analyses are beyond the scope of this course.

Table 12.3 *Review of Common Regression Terms*[26]

Concept	What It Means in Regression
Outcome variable	The continuous variable being predicted
Predictor variable	A variable used to predict the outcome
Unstandardized coefficient, b	The expected change in the outcome for a one-unit increase in the predictor, holding other predictors constant
Standardized coefficient, 𝛽	A coefficient on a common scale that can help compare predictor strength
Intercept	The predicted outcome when all predictors are 0
Residual	The difference between the observed outcome and the predicted outcome
R^2	The proportion of variance in the outcome accounted for by the predictors
Adjusted R^2)	A version of (R^2) adjusted for the number of predictors
Hierarchical regression	Regression with predictors entered in blocks to test incremental prediction

12.3 Summary

Correlation and regression are ways to look at relations between variables. They are flexible, but that flexibility means we need to be careful. Always identify the outcome variable, identify the predictors, check assumptions, interpret the overall model, and then interpret the individual coefficients in relation to the research question.

12.4 Additional Resources

For tips and demonstrations on setting up and running statistical analyses in SPSS, please see Virginia Wickline: SPSS Statistics Helps (publishing dates vary).

12.5 Media Attributions

[1-5] Linnell, D. (n.d.). *Statistics with jamovi.* Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[6] Wickline, V. B. (2026). Pearson's correlation table for sleep and grumpiness example. Annotated image from Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). Introduction to statistics in the psychological sciences. Pressbooks. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[7-9] Linnell, D. (n.d.). *Statistics with jamovi.* Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[10-11] Navarro, D. J., & Foxcroft, D. R. (2025). *Learning statistics with jamovi: A tutorial for beginners in statistical analysis.* Open Book Publishers. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[12-13] Horst, A. (n.d.). Linear regression dragons are CC BY-NC-SA 4.0.

[14-15] Linnell, D. (n.d.). *Statistics with jamovi.* Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[16-18] Horst, A. (n.d.). Linear regression dragons are CC BY-NC-SA 4.0.

[19-24] Linnell, D. (n.d.). *Statistics with jamovi.* Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[25] CC BY-NC-SA 4.0 by 3275Sartell via Wikimedia Commons

[26] Linnell, D. (n.d.). *Statistics with jamovi.* Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

12.6 Text Attributions

Illowsky, B., Dean, S., Birmajer, D., Blount, B., Boyd, S., Einsohn, M., Foreman, N., Helreich, J., Kenyon, L., Lee, S., & Taub, J. (2023). Introductory Statistics 2e. Licensed under Creative Commons Attribution-NonCommercial-ShareAlike License v.4.0. Modified by current authors.

Linnell, D. (n.d.). Statistics with jamovi. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Navarro, D. J., & Foxcroft, D. R. (2025). Learning statistics with jamovi: A tutorial for beginners in statistical analysis. Open Book Publishers. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

12.7 References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

13 Chapter 13: Three Kinds of *t*-Tests

13.1 Introduction to *t*-Tests

The *t*-test looks at the difference in means between two things (e.g., groups, time, observations). There are three different types of *t*-tests.

1. The ***one-sample t-test*** tests how the sample mean compares to the population mean.
2. The ***independent t-test*** has two *independent* groups. The participants or things in group 1 are *not* the same as the participants or things in group 2. This is a between-subjects design in which different participants are in the two groups.
3. The ***dependent t-test*** has *dependent* or *paired* data. The dependent variable is measured at two different times or for two different conditions for all participants or things. This is a within-subjects design in which case the same participants are in both groups (or the participants are matched on some important third variable that can also have a big influence, like gender, genetics, or ethnicity).

13.2 Effect Size

[The most commonly used measure of effect size for a *t*-test is ***Cohen's d*** (Cohen, 1988). It's a very simple measure in principle, with quite a few wrinkles when you start digging into the details. Cohen himself defined it primarily in the context of an independent samples *t*-test, specifically the Student's *t*-test. In that context, a natural way of defining the effect size is to divide the difference between the means (M) by an estimate of the standard deviation (SD). In other words, we're looking to calculate something along the lines of this:]

$$d = \frac{(M1) \ - \ (M2)}{SD}$$

Cohen's d is then provided a label to help the reader interpret the size of the effect.

Table 13.1: **Table 13.1** A (Very) Rough Guide to Interpreting Cohen's d

d-value	Interpretation
0.2 (positive or negative)	Small effect
0.5 (positive or negative)	Moderate (or medium) effect
0.8 (positive or negative)	Large effect

You'd think that this would be pretty unambiguous, but it's not. This is largely because Cohen wasn't too specific on what he thought should be used as the measure of the standard deviation (in his defense, he was trying to make a broader point in his book, not nitpick about tiny details).

As discussed by McGrath and Meyer ([2006]), there are several different versions in common usage, and each author tends to adopt slightly different notation. For the sake of simplicity (as opposed to accuracy), I'll use d to refer to any statistic that you calculate from the sample, and use to refer to a theoretical population effect. Obviously, that does mean that there are several different things all called d.

13.3 The One-Sample *t*-Test

The one-sample t-test is used to test the difference between our sample's mean (M) for the dependent variable and the mean of the population (m).

There are three different types of alternative hypotheses we could have for the one sample t-test:

1. **Two-tailed (non-directional)**
    - [H_1: The sample mean has a different mean than the population mean.] ($M = \mu$)
    - [H_0: There is no difference in means between the sample and population.] ($M \neq \mu$)
2. **One-tailed (directional)**
    - [H_1: The sample has a greater mean than the population.] ($M > \mu$)
    - [H_0: The mean for the sample is less than or equal to the mean for the population.] ($M \leq \mu$)
3. **One-tailed (directional)**
    - H_1: The sample has a smaller mean than the population. ($M < \mu$)

- H_0: The mean for the sample is greater than or equal to the mean for the population. ($M \geq \mu$)

Please watch What is the One-Sample T-Test?

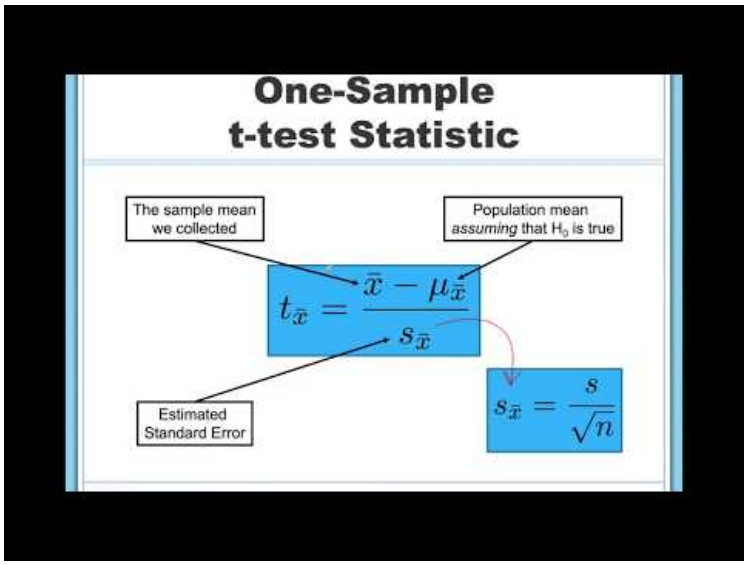


Figure 13.1: What is the One-Sample T-Test?

13.3.1 Step 1: Look at the Data and State the Null and Research Hypotheses

Before conducting a one-sample *t*-test, it is important to examine the data and verify that the assumptions of the test are met.

[If you learn one thing through this course, it is that *you should always look at your data!!!* Descriptive and inferential statistics can sometimes hide weird things with your data, so it's incredibly important you look at it first.]

13.3.1.1 Data Set-Up

Before conducting a one-sample *t*-test, it is important to ensure that the data are organized correctly. The dataset should contain one continuous (interval or ratio) dependent variable, and each row should represent a unique participant or unit of analysis. Properly organizing the data before conducting any statistical analysis helps minimize data-entry errors and ensures that the results are interpreted correctly.

13.3.1.2 Describe the Data

Once we have confirmed that our data are entered correctly, we should examine our data using descriptive statistics and graphs. First, our descriptive statistics are shown below. Looking at the mean, standard deviation, minimum, and maximum values allows us to become familiar with the distribution of our data and identify any potential data-entry errors or unusual values.

For this example, we will use a hypothetical dataset of 20 psychology students taking Dr. Zeppo's introductory statistics class. Dr. Zeppo wants to know if psychology students tend to get the same grade as everyone else ($M = 67.5$) or whether they get a higher or lower grade. As students in a psychology course, we're going to assume psychology students get higher grades.

For a one-sample *t*-test, we focus on one continuous (interval or ratio) outcome variable and one comparison value. In this example, the outcome variable is students' course grades. The comparison value is the overall course mean of 67.5.

Therefore, our hypotheses can be written as follows:

- H: Psychology students earn higher grades than the overall student population. ($M > \mu$)
- H_0: Psychology students do not earn higher grades than the overall student population. ($M \leq \mu$)

Relatedly, we want to be clear about our alpha value. We will use the conventional significance level of = .05. Therefore, for our results to be statistically significant, we want our calculated *p*-value to be less than .05.

13.3.2 Step 2: Check Assumptions and Set the Critical Value(s)

As a parametric test, the one-sample *t*-test has the same assumptions as other parametric tests:

1. The dependent variable is **normally distributed.**

   Normality. We're still assuming that the population distribution is normal, and there are standard tools that you can use to check to see if this assumption is met, and other tests you can do in its place if this assumption is violated (which goes beyond the scope of this course). At least look at the histogram and see if it appears approximately normal or not.

2. The dependent variable is **interval or ratio** (i.e., continuous).
3. Scores are **independent** of one another.

Independence. This assumption of the test is that the observations in your data set are not correlated with each other or related to each other in some funny way. An obvious (and stupid) example of something that violates this assumption is a data set where you "copy" the same observation over and over again in your data file so that you end up with a massive "sample size", which consists of only one genuine observation. More realistically, you have to ask yourself if it's really plausible to imagine that each observation is a completely random sample from the population that you're interested in. In practice this assumption is never met, but we try our best to design studies that minimize the problems of correlated data.

The second and third assumptions cannot be tested directly. Instead, they are evaluated based on our knowledge of the data and study design. For example, independence assumes that one participant's score does not influence another participant's score and that each observation represents a unique case.

However, we can and should evaluate the first assumption, normality. Unlike some statistical procedures with data software, the one-sample *t*-test may not include a built-in assumption check. Therefore, we can assess normality separately using descriptive statistics (e.g., are the mean, median, and mode all roughly equal?), graphical methods (e.g., does the histogram look approximately normal?), and formal tests of normality.

One thing to keep in mind is that statistical software often allows us to check assumptions while conducting an analysis. However, assumptions should always be evaluated before interpreting the results of a statistical test. Therefore, we will first examine whether the assumption of normality has been met.

For a formal test of normality, divide the value by its standard error to determine the z-score. If the absolute value of the z-score is close to zero than 1.96, then we assume it is normally distributed. We can also look to see if the skew and kurtosis values are close to zero (less than 2).

13.3.2.1 Setting the Critical Value(s)

Once we know the sample size (N), degrees of freedom (d.f.), and probability level (a), we use this information by looking at a probability table for the test we are using (in this case *t*-tests) to see what critical *t*-value(s) we will need to label a test as statistically significant. Please see Appendix A for links to probability tables. In this particular case, let's use a standard a = .05 (which we will do all semester, unless your professor requires otherwise). We are fairly confident we know how this difference will go: Psychology students will score higher than other students. Thus, we will use a one-tailed test, so we will have one critical value for our *t*-test. For a sample size of 20, the degrees of freedom are $N - 1$, or 20-1 = 19. The critical value is 1.729 (or 1.73 if we round to two decimal places).

APPENDIX B: T DISTRIBUTION TABLE (T TABLE)

	PROPORTION (α) IN ONE TAIL								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	PROPORTION (α) IN TWO TAILS COMBINED								
df	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.578
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.600
3	0.765	1.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	1.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850

Figure 13.2: **Figure 13.1** Critical value (1.729) for one-tailed, one-sample t-test with 19 degrees of freedom and a = .05[1]

13.3.3 Step 3: Calculate & Report the Descriptive Statistics, Test Statistic, and Effect Size

Now that we've checked the assumptions, we can perform the one-sample *t*-test.

13.3.3.1 Decide Whether to Use the Parametric or Nonparametric Test

If the normality assumption is reasonably met, use the one-sample *t*-test. If the normality assumption is seriously violated and no appropriate transformation addresses the issue, use the Wilcoxon signed-rank test (which goes beyond the scope of our course).

13.3.3.2 Decide Which Hypothesis You Should Be Using

When you specify hypotheses, decide whether the alternative hypothesis is directional (one-tailed) or non-directional (two-tailed) based on the research question.

If the alternative hypothesis is non-directional (two-tailed), interpret the two-tailed significance value (Sig. 2-tailed) reported in the SPSS output.

If the alternative hypothesis is directional (one-tailed), determine whether the sample mean is expected to be higher or lower than the comparison value before conducting the analysis. If SPSS only reports the two-tailed p-value, but the result is in the predicted direction, the one-tailed p-value is obtained by dividing the two-tailed p-value by two. If the result is not in the predicted direction, the one-tailed hypothesis is not supported.

Depending on your professor, you might then calculate your t-test by hand or using statistical software such as SPSS, SAS, R, Excel, or jamovi. Once your data are calculated, we will combine Step 3 with Step 4 below.

13.3.4 Step 4: Interpret Your Findings in Relation to the Hypothesis (Translate Math to English)

Once we are satisfied that we have met the assumptions for the one-sample t-test, we can interpret our results.

If our p-value is less than our a value of .05, our results are statistically significant. Like most of the statistics we'll come across, the larger the t-statistic (or F-statistic, or chi-square statistic…), the smaller the p-value will be.

If our results are statistically significant, we reject the null hypothesis and conclude that the sample mean differs from the comparison value in the direction predicted by the alternative hypothesis.

However, remember what we've learned in the previous chapters! We rejected the null hypothesis, but there's always the chance that we've made a Type 1 error.

13.3.4.1 A Note About Positive and Negative *t*-Values

Students often worry about whether a t statistic is positive or negative. For a one sample t-test, the sign shows whether the sample mean is above or below the comparison value. In this example, the t statistic is positive because the sample mean is higher than the comparison value.

If the sample mean had been lower than the comparison value, the mean difference and t statistic would have been negative. The sign helps identify the direction of the difference, but the p-value tells us whether that difference is statistically significant.

You will not get negative values for chi-square or F statistics, but t statistics can be positive or negative.

13.3.4.2 Cohen's *d* from One Sample

The simplest situation to consider is the one corresponding to a one-sample t-test. In this case, this is the one sample mean (X) and one (hypothesized) population mean (μ0) to compare it to. Not only that, there's really only one sensible way to estimate the population standard deviation. We just use our usual estimate ($\hat{\sigma}$) . Therefore, we end up with the following as the only way to calculate:

$$d = \frac{\bar{X} - \mu_0}{\hat{\sigma}}$$

13.3.4.3 Write Up the Results in American Psychological Association (APA) Style

An APA-style results section should remind the reader of the research question, summarize the relevant descriptive statistics, report the inferential test and effect size, and interpret the result.

Here is an example of how we can write up our results in APA style:

> Dr. Zeppo's psychology colleague hypothesized that psychology students would earn higher grades than the overall student population. Psychology students (M = 72.45, SD = 9.68, n = 20) earned significantly higher grades than the population mean (m= 67.50, $t(19)$ = 2.29, p = .017, d = 0.51, with a medium effect.

Note that this is not the only way we can write up the results in APA format. The key is that we include all four pieces of information as specified above.

Also note that the *M, SD, n, t, p*, and d are all italicized! This is an important part of APA style to remember.

13.4 Independent Samples *t*-Test

Although the one sample t-test has its uses, it's not the most typical example of a t-test. A much more common situation arises when you've got two different groups of observations. In psychology, this tends to correspond to two different groups of participants, where each group corresponds to a different condition in your study. For each person in the study, you measure some outcome variable of interest, and the research question that you're asking is whether or not the two groups have the same population mean. This is the situation that the independent samples t-test is designed for.

An independent samples t-test is used to test whether two independent groups differ on a continuous (interval or ratio) outcome variable. We use an independent samples t-test when we have one continuous outcome variable and one categorical grouping variable with exactly two groups. The groups are independent because different participants or cases are in each group. The goal is to determine whether two "independent samples" of data are drawn from populations with the same mean (the null hypothesis) or different means (the alternative hypothesis). When we say "independent" samples, what we really mean here is that there's no special relationship between observations in the two samples.

General Independent t-Test Hypotheses:

1. **Two-tailed (non-directional)**
    - H_1: There is a difference between the two groups on the outcome variable. ($M1 \neq M2$)
    - H_0: There is no difference between the two groups on the outcome variable. ($M1 = M2$)
2. **One-tailed**
    - H_1: Group 1 scores higher than Group 2 on the outcome variable. ($M1 > M2$)
    - H_0: Group 1 scores the same as or lower than Group 2 on the outcome variable. ($M1 \leq M2$)
3. **One-tailed**
    - H_1: Group 1 scores lower than Group 2 on the outcome variable. ($M1 < M2$)
    - H_0: Group 1 scores the same as or higher than Group 2 on the outcome variable. ($M1 \geq M2$)

We use an independent samples t-test whenever we compare two separate groups on one continuous (interval or ratio) outcome variable. For example, we might use an independent t-test in an experiment comparing a treatment group and a control group on an outcome of interest, like depression scores. We might also use an independent t-test to compare two existing groups, such as test scores of students taught by two different tutors.

Please watch What is the Independent-Samples T-Test?

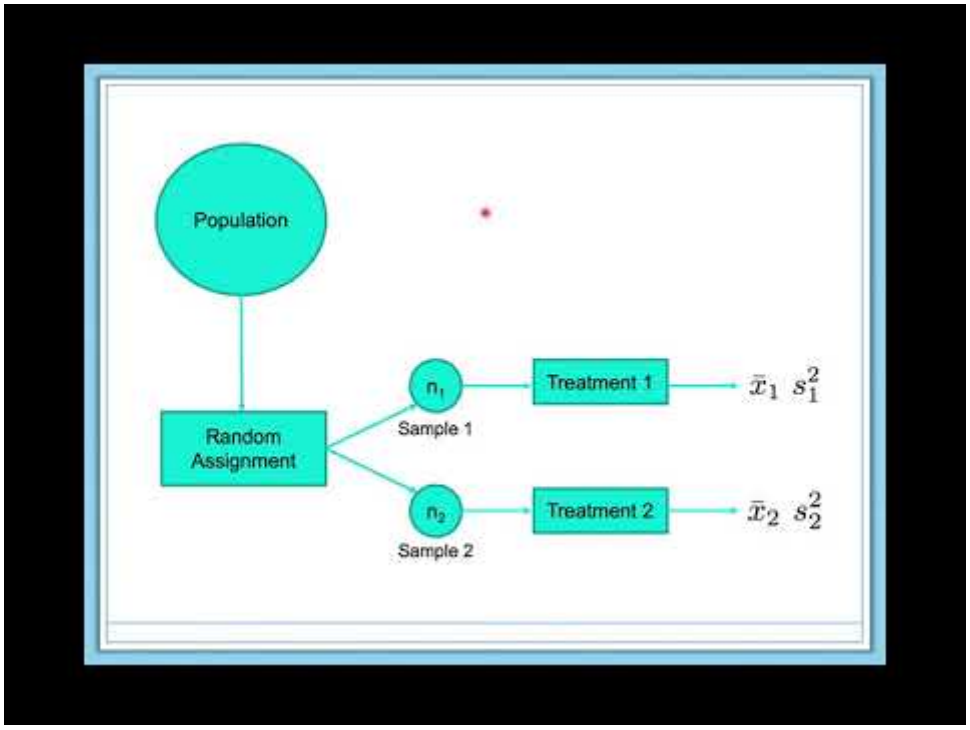


Figure 13.3: What is the Independent-Samples T-Test?

13.4.1 Step 1: Look at the Data and State the Null and Alternative Hypotheses

For this example, this dataset contains hypothetical data from 33 students taking Dr. Harpo's statistics lectures. Students were aided by one of two tutors: Anastasia ($n = 15$) or Bernadette ($n = 18$). Our research question is: Do students aided by Anastasia and Bernadette differ in their mean statistics test grades?

To conduct an independent samples t-test, the dataset needs one continuous (interval or ratio) outcome variable and one categorical (nominal or ordinal) grouping variable with exactly two groups. Each row should represent one participant or unit of analysis, and each participant should belong to only one group.

13.4.1.1 Describe the Data

Once we confirm that the data are entered correctly, we should describe the outcome variable using descriptive statistics. Looking at the sample size, mean, standard deviation, minimum, and maximum values allows us to become familiar with the distribution of the data and identify any potential data-entry errors or unusual values.

Before conducting the independent-samples t-test, we should also examine whether the distributions within each group are approximately normal and whether the groups have similar variability. These descriptive checks help prepare us for the formal assumption checks in the next step.

13.4.1.2 Specify the Hypotheses

Our research question is: Do students taught by Anastasia and Bernadette differ in their grades? This research question is non-directional (two-tailed) because it does not predict which tutor's students will have higher grades. Therefore, our hypotheses are:

- $\mathbf{H_1}$: There is a difference in grades between students taught by Anastasia and students taught by Bernadette. ($M1 \neq M2$)
- $\mathbf{H_0}$: There is no difference in grades between students taught by Anastasia and students taught by Bernadette. ($M1 = M2$)

If we had specific reason to believe that one tutor was more effective than the other, we could also set up a one-tailed hypothesis as we reviewed for a one-sample t-test.

We will use the conventional alpha level of $\alpha = .05$. Therefore, we will consider the result statistically significant if the p-value is less than .05.

13.4.2 Step 2: Check Assumptions and Set the Critical Value(s)

As a parametric test, the independent samples t-test has several assumptions:

- The outcome variable is similarly **approximately normally distributed** within each group.
  - Like the one-sample t-test, the independent-samples t-test assumes that the outcome variable is normally distributed within each group. We can evaluate this assumption using graphical methods (e.g., histograms and Q-Q plots) as well as formal tests of normality, such as the Shapiro-Wilk test.
- The two groups have roughly equal variances, which is called ***homogeneity of variance***.
  - Homogeneity of variance is assuming that the population standard deviation is the same in both groups. To determine whether the shapes are roughly similar for both groups, a Levene's test can be utilized.
- The outcome variable is measured on an **interval or ratio** (i.e., continuous) scale.
- Observations are **independent**. Each participant or case should contribute one score and belong to only one group.

- The independent samples *t*-test assumes that observations are independently sampled. This assumption has two aspects. First, observations within each group should be independent of one another. Second, there should be no overlap between groups, meaning each participant belongs to only one group and contributes only one score to the analysis. This assumption cannot be tested statistically and instead depends on proper study design and data collection.

We cannot test the third and fourth assumptions using the output alone; those assumptions are based on how the data were measured and collected.

However, we can and should evaluate the first two assumptions (normality and homogeneity of variance). Although most statistical software allows us to obtain these assumption checks while conducting the independent samples *t*-test, assumptions should always be evaluated before interpreting the results. Therefore, we will first examine whether these assumptions have been met by reviewing the histograms for each group and running a ***Levene's test***, which will allow us to determine if the distribution for each sample are similar shapes.

13.4.2.1 Setting the Critical Value(s)

Once we know the sample size (N), degrees of freedom (d.f.), and probability level (a), we use this information by looking at a probability table for the test we are using (in this case *t*-tests) to see what critical *t*-value(s) we will need to label a test as statistically significant. Please see Appendix A for links to probability tables. In this particular case, let's use a standard a = .05 (which we will do all semester, unless your professor requires otherwise). We are not confident we know how this difference will go: Either tutor's students could score better. Thus, we will use a two-tailed test, so we will have two critical values for our *t*-test. For a sample size of 33, the degrees of freedom are $N - 2$, or 33-2 = 31. Since our sample is getting larger and more robust, we will not find a critical value for each and every degree of freedom. Therefore, we will round down to the closest number, in this case, 30 instead of 31. The critical value is 2.042 (or 2.04 if we round to two decimal places). Since we are doing a two-tailed test, we will have two critical values: +2.042 and -2.042. That way, if either tutor's students are scoring better, we will have captured that difference at either end of the distribution.

APPENDIX B: T DISTRIBUTION TABLE (T TABLE)

	PROPORTION (α) IN ONE TAIL								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	PROPORTION (α) IN TWO TAILS COMBINED								
df	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.578
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.600
3	0.765	1.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	1.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.689
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.660
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.290

Adapted from "Table t" by Jmarchn/Wikimedia Commons, CC BY SA 4.0.

Figure 13.4: **Figure 13.2** Critical value (2.042) for two-tailed, independent-samples t-test with 31 degrees of freedom and a = .05[2]

13.4.3 Step 3: Calculate & Report the Descriptive Statistics, Test Statistic, and Effect Size

13.4.3.1 Decide Whether to Use Student's *t*-test, Welch's *t*-test, or Mann-Whitney

[For an independent samples *t*-test, we must determine whether the Student's *t*-test, Welch's *t*-test, or Mann-Whitney *U*-test is most appropriate. Based on our assumption checks, we can decide which test to report.]

Table 13.2: **Table 13.2** Assumption Pattern and Which t-Test to Report[3]

Assumption Pattern	**Test to Report**
Normality is reasonably met and homogeneity of variance is reasonably met	Student's *t*-test
Normality is reasonably met but homogeneity of variance is not met	Welch's *t*-test or correction with Levene's test
Normality is seriously violated and no appropriate transformation addresses the issue	Mann-Whitney U-test

Some researchers prefer Welch's *t*-test as the default because it performs well when variances are unequal and usually gives very similar results to Student's *t*-test when variances are equal. In this example, both the normality and homogeneity-of-variance assumptions were met. Therefore, the Student's independent-samples *t*-test is the appropriate test to report.

13.4.3.2 Decide Which Hypothesis You Should Be Using

When you specify hypotheses, decide whether the alternative hypothesis is directional (one-tailed) or non-directional (two-tailed) based on the research question.

If the alternative hypothesis is non-directional, choose Mean 1 ≠ Mean2.

If the alternative hypothesis is directional, choose Mean 1 > Mean 2 or Mean 1 < Mean2 based on the direction predicted in advance by the research question. Do not choose the direction [after] looking at which group has the higher sample mean.

*The hypothesis direction must be chosen *before* looking at the results. Do not switch from a two-tailed test to a one-tailed test because the sample means appear to differ in one direction.*

Another way to decide whether the samples have met the homogeneity of variance test is to utilize the ***Levene's test***, which compares the variance of each sample by calculating an F-value. If the variances are the same (which is what we want to see because it is less messy), the *F*-value will be non-significant (assuming equal variances), and we proceed as usual. Tally ho! Full steam ahead! If the Levene's test is significant, equal variances cannot be

assumed, and a modification is needed to account for these differing shapes. If you are using a statistics program to calculate your data, you will drop down and use the lower set of values (equal variances not assumed), rather than reporting the upper line of values (equal variances assumed).

Depending on your professor, you might then calculate your t-test by hand or using statistical software such as SPSS, SAS, R, Excel, or jamovi. Once your data are calculated, we will combine Step 3 with Step 4 below.

13.4.4 Step 4: Interpret Your Findings in Relation to the Hypothesis (Translate Math to English)

Once we are satisfied that the assumptions for the independent samples t-test are reasonably met, we can interpret the results.

Again, if the p-value is less than our alpha value of .05, the result is statistically significant. We reject the null hypothesis that equal population means exist. In this case, let's assume the sample provides evidence that students taught by Anastasia and Bernadette differ in their mean grades.

Because Anastasia's students had the higher sample mean, the difference is in favor of Anastasia's class. However, the statistical test tells us whether the group means differ more than we would expect by chance; it does not, by itself, [prove] that the tutor caused the difference unless the study design supports a causal conclusion, so be cautious (and fair!) in how you understand and explain this outcome.

13.4.4.1 A Note About Positive and Negative *t* Value

Students often worry about whether a t-statistic is positive or negative. For an independent samples t-test, the sign depends on which group is treated as Group 1 and which group is treated as Group 2. In this example, the mean difference is arbitrarily calculated as Anastasia's class minus Bernadette's class: $(76.79 - 70.21 = 6.58)$. Because Anastasia's class has the higher mean, the t-statistic is positive.

If the group order were reversed, and Bernadette's class were used as the starting point, the mean difference and t-statistic would be negative. The sign helps identify the direction of the difference based on the group order, but the p-value tells us whether that difference is statistically significant.

You will not get negative values for F-statistics or chi-square statistics, but t-statistics can be positive or negative.

13.4.4.2 Cohen's *d* from Independent Samples

The majority of discussions of Cohen's d focus on a situation that is analogous to Student's independent samples t-test, and it's in this context that the story becomes messier, since there are several different versions of d that you might want to use in this situation. To understand why there are multiple versions of d, it helps to take the time to write down a formula that corresponds to the true population effect size . It's pretty straightforward:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

where, as usual, μ and μ are the population means corresponding to group 1 and group 2 respectively, and is the standard deviation (the same for both populations). The obvious way to estimate is to do exactly the same thing that we did in the t-test itself, i.e., use the sample means as the top line and a pooled standard deviation estimate for the bottom line:

$$d = \frac{\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2}{\hat{\sigma}_p}$$

where is the exact same pooled standard deviation measure that appears in the t-test. This is the most commonly used version of Cohen's d when applied to the outcome of a Student's t-test. It is sometimes referred to as Hedges' g statistic.

However, there are other possibilities that I'll briefly describe (but are really beyond the scope of this course). First, you may have reason to want to use only one of the two groups as the basis for calculating the standard deviation. This approach (often called Glass' Δ, pronounced delta) only makes most sense when you have good reason to treat one of the two groups as a purer reflection of "natural variation" than the other. This can happen if, for instance, one of the two groups is a control group. Secondly, in the usual calculation of the pooled standard deviation we divide by $N - 2$ to correct for the bias in the sample variance. In one version of Cohen's d this correction is omitted, and instead we divide by N. This version makes sense primarily when you're trying to calculate the effect size in the sample rather than estimating an effect size in the population. Finally, there is a version called Hedges' g, based on Hedges & Olkin (1985), who point out there is a small bias in the usual (pooled) estimation for Cohen's d.

13.4.4.3 Write Up the Results in APA Style

An APA-style results section should remind the reader of the research question, summarize the relevant descriptive statistics, report the inferential test and effect size, and interpret the result.

We can write up our results in APA something like this:

> Dr. Harpo wondered if students statistics grades would differ based on their tutor. The Levene's test was non-significant, indicating equal variances for the samples. An independent samples t-test indicated there was a difference in students taught by Anastasia ($M = 76.79$, $SD = 5.61$, $n = 14$) and students taught by Bernadette ($M = 70.21$, $SD = 5.71$, $n = 19$), $t(31) = 3.29$, $p = .002$, $d = 1.16$. This finding is significant and large; the hypothesis is supported. Anastasia's students had higher grades than Bernadette's students.

This is not the only correct way to write the result. The key is to include the correct information (i.e., group descriptive statistics, test statistic, degrees of freedom, p-value, effect size, and interpretation) and to make it clear and easy to read.

13.5 Dependent Samples *t*-test

A dependent samples t-test (also called a paired samples t-test) is used to test whether two related measurements differ on a continuous outcome variable. We use a dependent t-test when the same participants are measured twice or when observations are matched in pairs based on an important third variable.

There are three different types of alternative hypotheses we could have for a dependent t-test:

1. **Two-tailed (non-directional)**
    - H_1: There is a difference between the two related measurements. ($M1 \neq M2$)
    - H_0: There is no difference between the two related measurements. ($M1 = M2$)
2. **One-tailed (directional)**
    - H_1: Scores are higher on Measurement 1 than on Measurement 2. ($M1 > M2$)
    - H_0: Scores are the same or lower on Measurement 1 than on Measurement 2. ($M1 \leq M2$)
3. **One-tailed (directional)**
    - H_1: Scores are lower on Measurement 1 than on Measurement 2. ($M1 < M2$)
    - H_0: Scores are the same or higher on Measurement 1 than on Measurement 2. ($M1 \geq M2$)

Please watch What is the Dependent-Samples T-Test?

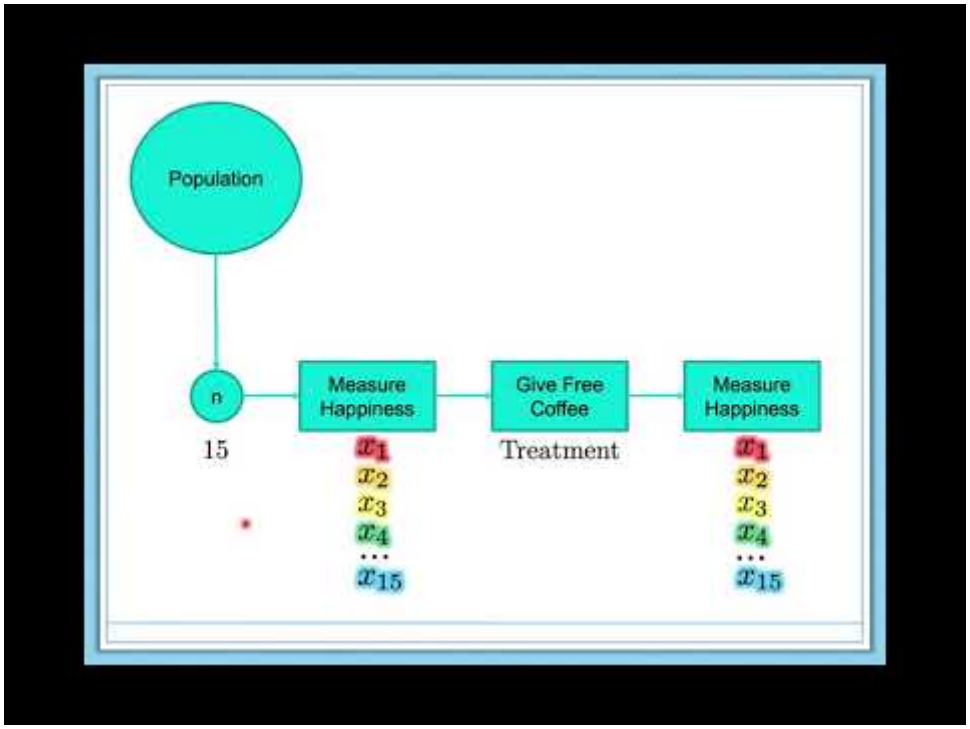


Figure 13.5: What is the Dependent-Samples T-Test?

13.5.1 Step 1: Look at the Data and State the Null and Research Hypotheses

This dataset contains hypothetical data from 20 students in Dr. Chico's class. Each student took two tests: one early in the semester and one later in the semester. Dr. Chico thinks the first test is a "wake-up call" for students. She predicts that students will work harder after the first test and score higher on the second test.

13.5.1.1 Data Set-Up

To conduct a dependent samples t-test, the dataset usually needs two columns: one column for the first measurement and one column for the second measurement. Each row should represent one participant or matched pair. The two scores in the same row are paired because they belong to the same participant or matched case.

13.5.1.2 Describe the Data

For a dependent samples t-test, we should pay attention to the pattern of possible change within participants. The descriptive statistics might show that the average score was higher on the second test, but we still need to test whether the difference is large enough to be statistically significant.

13.5.1.3 Specify the Hypotheses

Dr. Chico predicts that students will score higher on the second test than on the first test. This is a directional research question. Therefore, our hypotheses are:

- H_1: Students score higher on the second test than on the first test. ($M1 < M2$)
- H_0: Students score the same or lower on the second test than on the first test. ($M1 \geq M2$)

We will use the conventional alpha value of $\alpha = .05$. Therefore, we will consider the result statistically significant if the p-value is less than .05.

13.5.2 Step 2: Check Assumptions and Set the Critical Value(s)

As a parametric test, the dependent samples t-test has several assumptions:

1. The difference scores are **approximately normally distributed**. A difference score is the difference between each participant's two measurements.
2. The outcome variable is **interval or ratio** (i.e., continuous).
3. Pairs are independent of other pairs. The two **measurements within a row are related**, but each participant or matched pair should be independent of the other participants or matched pairs.

We cannot test the second and third assumptions using the output alone; those assumptions are based on how the data were measured and collected. However, we can and should evaluate the first assumption.

One thing to keep in mind in all statistical software is that we often check assumptions simultaneously while performing the statistical test. However, we should always check assumptions first before looking at and interpreting our results. Therefore, we discuss checking assumptions here first to help ingrain the importance of always checking assumptions for interpreting results.

13.5.2.1 Testing Normality

For a dependent samples *t*-test, the normality assumption applies to the difference scores, not to each measurement separately. In this example, the difference score is the difference between each student's first test score and second test score. Review the histogram of the difference scores to determine if it is approximately normal.

*Note: A dependent samples *t*-test does not require a homogeneity-of-variance test because the two measurements are paired rather than independent groups. The key assumption is that the difference scores are approximately normally distributed. When there are three or more related measurements, we use repeated-measures ANOVA and evaluate a related assumption called sphericity, which we will discuss in Chapter 14.*

13.5.2.2 Setting the Critical Value(s)

Once we know the sample size (N), degrees of freedom (d.f.), and probability level (a), we use this information by looking at a probability table for the test we are using (in this case *t*-tests) to see what critical *t*-value(s) we will need to label a test as statistically significant. Please see Appendix A for links to probability tables. In this particular case, let's use a standard a = .05 (which we will do all semester, unless your professor requires otherwise). We are fairly confident we know how this difference will go: Students will score higher on the second test than the first test. Thus, we will use a one-tailed test, so we will have one critical value for our *t*-test. For a sample size of 20, the degrees of freedom are $N - 1$, or 20-1 = 19. The critical value from the table is 1.729 (or 1.73 if we round to two decimal places). However, we expected students to do better on the second test, the final critical value will be -1.729 (or -1.73 if we round to two decimal places), because we expect if we subtract Test1 – Test2, the difference will show as negative (favoring the second test).

APPENDIX B: T DISTRIBUTION TABLE (T TABLE)

	PROPORTION (α) IN ONE TAIL								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	PROPORTION (α) IN TWO TAILS COMBINED								
df	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.578
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.600
3	0.765	1.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	1.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850

Figure 13.6: **Figure 13.3** Critical value (-1.729) for one-tailed, dependent-samples t-test with 19 degrees of freedom and a = .05[4]

13.5.3 Step 3: Calculate & Report the Descriptive Statistics, Test Statistic, and Effect Size

13.5.3.1 Decide Whether to Use the Parametric or Nonparametric Test

If the normality assumption for the difference scores is reasonably met, use the dependent (paired-samples) t-test. If the normality assumption is seriously violated and no appropriate transformation addresses the issue, use the Wilcoxon signed-rank test (which goes beyond the scope of this course).

13.5.3.2 Decide Which Hypothesis You Should Be Using

When you specify hypotheses, decide whether the alternative hypothesis is directional (one-tailed) or non-directional (two-tailed) based on the research question.

If the alternative hypothesis is non-directional, select two-tailed. If the alternative hypothesis is directional, select $>$ or $<$ based on the direction predicted by the research question and the order of the paired variables. Do not choose the direction after looking at which measurement has the higher sample mean.

The hypothesis direction must be chosen before looking at the results.

Depending on your professor, you might then calculate your t-test by hand or using statistical software such as SPSS, SAS, R, Excel, or jamovi. Once your data are calculated, we will combine Step 3 with Step 4 below.

13.5.4 Step 4: Interpret Your Findings in Relation to the Hypothesis (Translate Math to English)

Once we are satisfied that the assumptions for the dependent samples t-test are reasonably met, we can interpret the results.

In this example, let's assume the p-value is less than .05, so the result is statistically significant. We will reject the null hypothesis. The sample provides evidence that students scored higher on the second test than on the first test.

13.5.4.1 A Note About Positive and Negative *t* Values

Students often worry about whether a t-statistic is positive or negative. For a dependent samples t-test, the sign depends on the order of the paired variables. In this example, grade_test1 is listed before grade_test2, so the mean difference is calculated as Test 1 minus Test 2: $(56.98 - 58.38 = -1.40)$. Because the first test mean is lower than the second test mean, the t-statistic is negative.

If the variables were entered in the reverse order, the mean difference and t-statistic would be positive. The sign helps identify the direction of the difference based on the variable order, but the p-value tells us whether that difference is statistically significant.

You will not get negative values for F-statistics or chi-square statistics, but t statistics can be positive or negative.

13.5.4.2 Cohen's d from Paired Samples

Finally, what should we do for a paired-samples t-test? In this case, the answer depends on what you are trying to measure. A paired-samples t-test measures effect size relative to the distribution of the difference scores, and the measure of d is calculated as:

$$d = \frac{D}{\hat{\sigma}_D}$$

where D is the mean of the difference scores and $\hat{\sigma}_D$ is the estimated standard deviation of the difference scores.

This is the version of Cohen's d that is typically reported for a dependent samples t-test. However, it is important to recognize that this effect size is based on the difference scores rather than the original variables. Depending on the research question, this may or may not be the most meaningful interpretation.

In some situations, researchers are more interested in measuring the effect size relative to the original variables instead of the difference scores. In those cases, alternative versions of Cohen's d similar to those used for Student's or Welch's t-tests may be more appropriate. The choice of effect size should therefore depend on the research question and how the magnitude of the effect is intended to be interpreted.

13.5.4.3 Write Up the Results in APA Style

An APA-style results section should remind the reader of the research question, summarize the relevant descriptive statistics, report the inferential test and effect size, and interpret the result.

We can write up our results in APA something like this:

> We hypothesized that students would perform better on the second exam than the first exam. The hypothesis was supported with a large effect. A dependent samples t-test indicated that students scored significantly higher on the second test (M = 58.38, SD = 6.40) than on the first test (M = 56.98, SD = 6.64), $t(19)$ = -6.45, p < .001, d = 1.44.

This is not the only correct way to write the results. The key is to include the correct information (i.e., the descriptive statistics for both measurements, test statistic, degrees of freedom, p-value, effect size, and interpretation) and to make it clear and easy to read.

13.6 Wrapping it Up

The t-test is an incredibly useful tool for comparing mean differences between two groups or measurements. Different forms of the t-test are used depending on the research design: the one-sample t-test compares a sample mean to a known population mean, the independent-samples t-test compares two independent groups, and the paired-samples (dependent) t-test compares two related measurements from the same participants (or pairs of participants matched on an important third variable). In addition to determining whether a difference is statistically significant, t-tests also allow researchers to calculate an effect size, such as Cohen's d, which describes the magnitude of the difference. Together, statistical significance and effect size provide a more complete understanding of the results.

13.7 Additional Resources

For tips and demonstrations on setting up and running statistical analyses in SPSS, please see Virginia Wickline: SPSS Statistics Helps (publishing dates vary).

13.8 Media Attributions

[1] Wickline, V. B. (2026). Critical value (1.729) for one-tailed, one-sample t-test with 19 degrees of freedom and a = .05. Annotated image from Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). Introduction to statistics in the psychological sciences.

Pressbooks. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[2] Wickline, V. B. (2026). Critical value (2.042) for two-tailed, independent-samples t-test with 31 degrees of freedom and a = .05. Annotated image from Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). Introduction to statistics in the psychological sciences. Pressbooks. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[3] Linnell, D. (n.d.). *Statistics with jamovi*. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[4] Wickline, V. B. (2026). Critical value (-1.729) for one-tailed, dependent-samples t-test with 19 degrees of freedom and a = .05. Annotated image from Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). Introduction to statistics in the psychological sciences. Pressbooks. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

13.9 Text Attributions

Linnell, D. (n.d.). Statistics with jamovi. https://danalinnell.github.io/statistics-with-jamovi/? Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Navarro, D. J., & Foxcroft, D. R. (2025). Learning statistics with jamovi: A tutorial for beginners in statistical analysis. Open Book Publishers. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

13.10 References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum. https://doi.org/10.4324/9780203771587

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press. https://doi.org/10.1016/C2009-0-03396-0

[McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods*, *11*, 386–401.] https://doi.org/10.1037/1082-989x.11.4.386

Storage, D. (2019, July 9). What is the dependent-samples t-test? https://www.youtube.com/watch?v=nYDgFIpmUWs

Storage, D. (2019, July 2). *What is the independent-samples t-test?* https://www.youtube.com/watch?v=EK72YyYDt6s

Storage, D. (2019, July 1). *[What is the one-sample t-test?]* https://www.youtube.com/watch?v=KlGMRu4MF8M

14 Chapter 14: Analysis of Variance (ANOVA) – Between and Within Groups

14.1 What is ANOVA?

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to their upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

Analysis of variance (ANOVA) serves the same purpose as the *t*-tests we learned earlier in this module. ANOVA tests for differences in group means. Although ANOVA is used to compare means, it does so by analyzing variability. ANOVA compares how much scores vary *between* groups or conditions to how much scores vary *within* groups or conditions. If the between-group variability is large relative to the within-group variability, the ANOVA is more likely to be statistically significant. Thus, as the name suggests, it looks at ***variances*** (meaning, the squared average difference between each score and the mean), but it looks at more than just the variance for each group (IV level). This chapter will describe the general design of ANOVA, with a focus first on calculating the ***one-way ANOVA for independent samples (also called one-way ANOVA between, between-subjects ANOVA***). An extension of the independent samples *t*-test, a one-way ANOVA between is used when three or more different groups are compared on a single independent (or grouping) variable. We will also look at one-way ANOVA within, which has several purposes but is most often used to measure the same individuals over time or across conditions/treatments.

Which type of research design should you try to use, between or within? The answer is actually somewhat practical rather than statistical. First, some independent variables cannot be repeated. You cannot teach someone to swim and then put them in a condition in which they are expected to not know how to swim again. In those cases, you must use a between groups design. Second, with each additional condition (group), we have to get at least 30 more participants to have good power for our design. This can get time-consuming and expensive quickly. Since repeated measures (within) designs are more cost-effective, and they take into account each participants' individual tendencies and quicks (variability), they are preferred if they are feasible. The best of both worlds is to get the same amount of participants that we'd

expect in a between groups design but measure them repeatedly. This increases our sample size without actually increasing our number of participants!

Please watch: What are Analyses of Variance? One-Way and Factorial ANOVAs

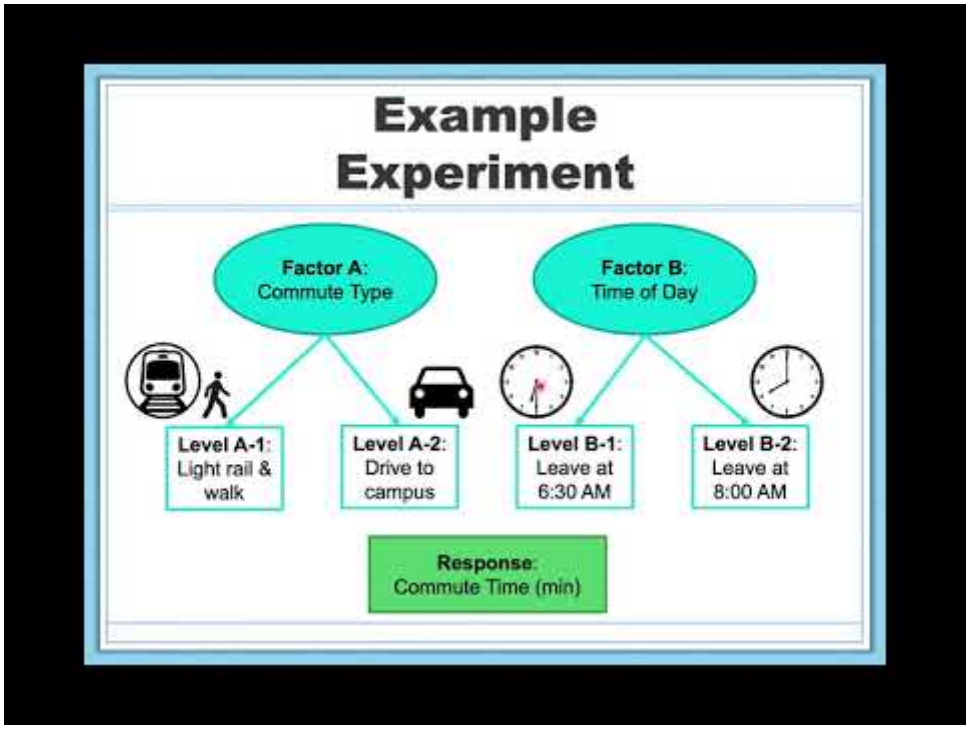


Figure 14.1: What are Analyses of Variance? One-way and Factorial ANOVAs

ANOVA is more flexible than a t-test: Unlike t-tests, which are limited to two groups (independent samples) or two time points (dependent samples), ANOVA can handle any number of groups. Thus, the purpose and interpretation of ANOVA will be the same as it was for t-tests, as will much of the hypothesis-testing procedure. However, at first glance ANOVA will look much different from a mathematical perspective, although as we will see, the basic logic behind the test statistic for ANOVA is actually the same as t-tests. The F-distribution is derived from the t-distribution. The values of the F-distribution are squares of the corresponding values of the t-distribution. One-Way ANOVA expands the t-test for comparing more than two groups. It is preferable to use ANOVA when there are more than two groups instead of performing pairwise t-tests because performing multiple tests introduces the likelihood of making a Type 1 error.

14.2 Type I Error

You may be wondering why we do not just conduct a t-test to test our hypotheses about three or more groups. After all, we are still just looking at group mean differences, right? The answer is that the t-statistic formula can only handle up to two groups, one minus the other. In order to use t-tests to compare three or more means, we would have to run a series of ***pairwise comparisons***. For only three groups, we would have three t-tests: group 1 vs group 2, group 1 vs group 3, and group 2 vs group 3. This may not sound like a lot, especially with the advances in technology that have made running an analysis very fast, but it quickly scales up. With just one additional group, bringing our total to four, we would have six comparisons: group 1 vs group 2, group 1 vs group 3, group 1 vs group 4, group 2 vs group 3, group 2 vs group 4, and group 3 vs group 4. This makes for a logistical and computation nightmare for five or more groups.

So, why is that a bad thing? Statistical software could run that in a jiffy, easy peasy. The real issue is the probability of committing a Type I error. Remember Type I and Type II errors? As a brief refresh, a Type I error is a false positive; the chance of committing a Type I error is equal to our significance level, **a**. This is true if we are only running a single analysis (such as a t-test with only two groups) on a single dataset. However, when we start running multiple analyses on the same dataset, we have the same Type I error rate for each analysis. The Type I error rate is cumulative; it increases with each new analysis. The more calculations we run, the more our Type I error rate expands. This raises the probability that we are capitalizing on random chance and rejecting a null hypothesis when we should not. ANOVA, by comparing all groups simultaneously with a single analysis, averts this issue and keeps our error rate at the **a** we set.

14.3 Observing and Interpreting Variability

We have seen time and again that scores, whether they are individual data or group means, will differ naturally. Sometimes this is due to random chance, and other times it is due to actual differences. Our job as scientists, researchers, and data analysts is to determine if the observed differences are systematic and meaningful (via a hypothesis test) and, if so, what might be causing those differences. Although we are usually interested in the mean or average score, it is the variability in the scores that is key.

Take a look at Figure 14.1, which shows scores for many people on a test of skill used as part of a job application. The x-axis has each individual person, in no particular order, and the y-axis contains the score each person received on the test. As we can see, the job applicants differed quite a bit in their performance, and understanding why that is the case would be extremely useful information. However, there is no interpretable pattern in the data, especially because we only have information on the test, not on any other variable (remember that the x-axis here only shows individual people and is not ordered or interpretable).

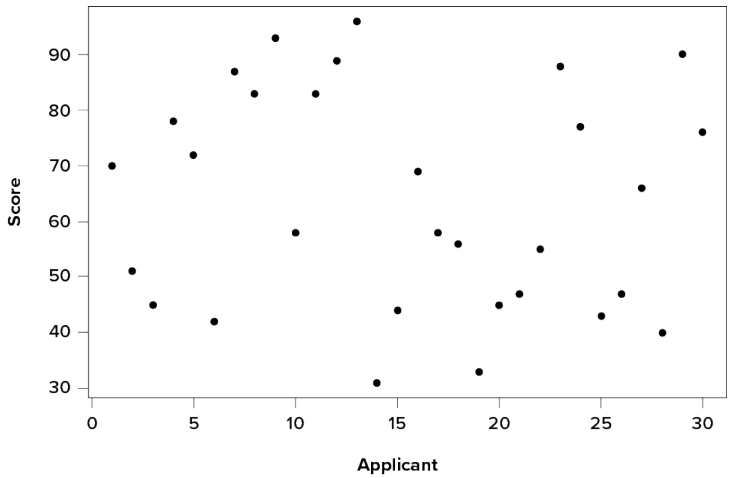


Figure 14.2: **Figure 14.1** Scores on a job test[1]

Our goal is to explain the variability that we are seeing in the dataset. Let's assume that as part of the job application procedure we also collected data on the highest degree each applicant earned. With knowledge of what the job requires, we could sort our applicants into three groups: applicants who have a college degree related to the job, applicants who have a college degree that is not related to the job, and applicants who did not earn a college degree. This is a common way that job applicants are sorted, and we can use ANOVA to test if these groups are actually different. Figure 14.2 presents the same job applicant scores, but now they are differentiated by color and shape by group membership (i.e., which group they belong in). Now that we can differentiate between applicants this way, a pattern starts to emerge: applicants with a relevant degree (coded red) tend to be near the top, applicants with no college degree (coded black) tend to be near the bottom, and applicants with an unrelated degree (coded blue) tend to fall into the middle. However, even within these groups, there is still some variability, as shown in Figure 14.2.

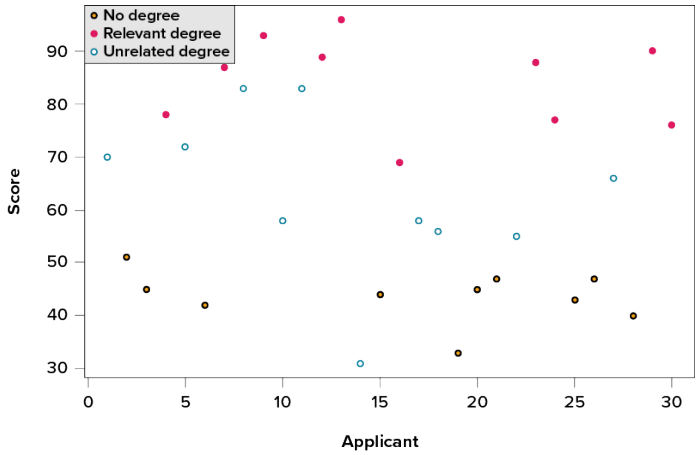


Figure 14.3: **Figure 14.2** Applicant scores on a job test, coded by degree earned[2]

This pattern is even easier to see when the applicants are sorted and organized into their respective groups, as shown in Figure 14.3

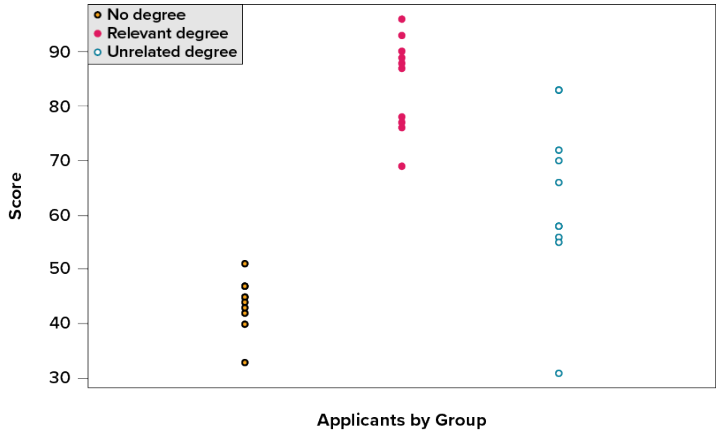


Figure 14.4: **Figure 14.3** Applicant scores by group, with 'no degree' on the left, 'relevant degree' in the middle, and 'unrelated degree' on the right [3]

Now that our data are organized into an easily interpretable format, we can clearly see that our applicants' scores differ largely along group lines. Those applicants who do not have a college degree received the lowest scores, those who had a degree relevant to the job received the highest scores, and those who did have a degree but one that is not related to the job tended to fall somewhere in the middle. Thus, we have ***systematic variability*** *between* our groups of applicants.

We can also clearly see that *within* each group, our applicants' scores differed from one another (the spread of the dots from top to bottom in each column). Those applicants without a

degree tended to score very similarly, since the scores are clustered close together. Our group of applicants with relevant degrees varied a little bit more than that, and our group of applicants with unrelated degrees varied quite a bit. It may be that there are other factors that cause the observed score differences within each group, or they could just be due to random chance. Because we do not have any other explanatory data in our dataset, the variability we observe within our groups is considered ***random error***, with any deviations between a person and that person's group mean caused only by chance. Thus, we have ***unsystematic variability*** *within* our groups.

The process and analyses used in ANOVA will take these two sources of variability (systematic variability between groups and random error within groups, or how much groups differ from each other and how much people differ within each group) and compare them to one another to determine if the groups have any explanatory value in our outcome variable. By doing this, we will test for statistically significant differences between the group means, just like we did for *t*-tests. For those asked to do hand calculations, we will go step by step to break down the math to see how ANOVA actually works. Recognize, however, that your professor might ask you to understand ANOVA conceptually but conduct the analyses on a statistical program like SPSS, R, or jamovi, rather than by hand.

14.4 ANOVA is a Ratio of Variances

The critical ingredient for a one-factor (one IV), between-subjects (multiple groups) ANOVA, is that there is one independent variable, with at least two levels (at least two different groups in that one IV). You might be thinking, "When you have one IV with two levels, you can run a *t*-test." And you would be correct! You could also run an ANOVA. Interestingly, they give you almost the exact same results. You will get a *p*-value from both tests that is identical (they are really doing the same thing under the hood). The *t*-test gives a *t*-value as the important sample statistic. The ANOVA gives us the ***omnibus*** (that is, overall) *F*-value (for Fisher, the inventor of the test) as the important sample statistic. It turns out that t^2 equals F when there are only two groups in the design. They are the same test.

Remember that t is the mean difference divided by the standard error of the sample. The idea behind F is the same basic idea that goes into making t. Here is the general idea behind the formula, it is again a ratio (division) of the independent variable—the effect we are measuring (in the numerator)—and the variation associated with everything else that can make individuals different (in the denominator).

$$\mathrm{F} = \frac{\text{measure of effect}}{\text{measure of error}}$$

This idea is the same as with the *t*-test. The difference with the *F*-test is that we use variances (how different, on average, each score is from the sample mean) to describe both the measure of the effect and the measure of error. So, F is a ratio of two variances.

When the variance associated with the effect is the same size as the variance associated with sampling error, we will get two of the same numbers, this will result in an F-value of 1 (a number divided by itself equals 1). When the variance due to the effect is larger than the variance associated with sampling error, then F will be greater than 1. When the variance associated with the effect is smaller than the variance associated with sampling error, F will be less than one.

Let's rewrite in plainer English. We are talking about two concepts that we would like to measure from our data. 1) A measure of what we want to explain using our IV levels, and 2) a measure of error, or stuff about our data we cannot explain by our IV levels. So, the formula looks like this:

$$\text{F} = \frac{\text{Can Explain by IV}}{\text{Can't Explain by IV}}$$

When what we can explain based on what group participants are in is as much as what we cannot explain, $F = 1$. This is not that great of a situation for us to be in. It means we have a lot of uncertainty. When what we can explain is much more than what we cannot explain, we are doing a good job, and F will be greater than 1. When what we can explain is less than what we cannot, we really cannot explain very much, so F will be less than 1. That is the reasoning behind calculating F.

If you saw an F in the wild, and it was .6, then you would automatically know the researchers could not explain much of the differences in their data. If you saw an F of 5, then you would know the researchers could explain 5 times more than they could not, so that is pretty good! The point of this section is to give you an intuition about the meaning of an F-value, even before you learn how to compute and interpret it.

14.5 Sources of Variability

ANOVA is all about looking at the different ***sources of variability*** (i.e., the reasons that scores differ from one another) in a dataset. The way we calculate these sources of variability takes the form of sum of squares. Before we get into the calculations themselves, we must first lay out some important terminology and notation.

In ANOVA, we are working with two variables, a grouping or explanatory variable and a continuous outcome variable. The ***grouping variable*** is our ***predictor variable*** (it predicts or explains the values in the outcome variable) or, in experimental terms, our ***independent variable***. It is made up of k groups, with k being any whole number 2 or greater. That is, ANOVA requires two or more groups to work, and it is usually conducted with three or more. In ANOVA, we refer to groups as ***levels***, so the number of levels is just the number of groups, which again is k. In the above example, our grouping variable was education, which had 3 levels, so $k = 3$. When we report any descriptive value (e.g., mean, sample size, standard deviation)

for a specific group, we will use a subscript 1…k (one through k) to denote which group it refers to. For example, if we have three groups and want to report the standard deviation s for each group, we would report them as s_1, s_2, and s_3.

Our second variable is our ***outcome variable***. This is the variable on which people differ, and we are trying to explain or account for those differences based on group membership. In experimental terms, this is our ***dependent variable***. In the example above, our outcome was the score each person earned on the test. Our outcome variable will still use X for scores as before. When describing the outcome variable using means, we will use subscripts to refer to specific, ***individual group means***. So, if we have $k = 3$ groups, our means will be M_1, M_2, and M_3. We will also have a single mean representing the average of all participants across all groups. This is known as the ***grand mean***, and we use the notation M_G. These different means—the individual group means and the overall grand mean—will be how we calculate our sums of squares.

Finally, we now have to differentiate between several different sample sizes. Our data will now have sample sizes for each group, and they could be the same or different. We will denote these with a lower case n and a subscript, just like with our other descriptive statistics: n_1, n_2, and n_3. We also have the overall sample size in our dataset, and we will denote this with a capital N. The total sample size is just the group sample sizes added together.

14.6 Between-Groups Sum of Squares

One source of variability we identified in Figure 14.3 was differences in test score variability between the degree groups. That is, the degree groups clearly had different average test scores. The variability arising from these differences is known as ***between-groups variability***, and between-groups sum of squares is used to calculate between-groups variability.

Our calculations for ***sums of squares*** in ANOVA will take on the same form as it did for regular calculations of variance. Each observation, in this case the group means, is compared to the overall mean, in this case the ***grand mean***, to calculate a deviation (difference) score. These deviation scores are squared so that they do not cancel each other out and sum to zero. The squared deviations are then summed (added up). There is, however, one small difference. Because each group mean represents a group composed of multiple people, before we sum the deviation scores we must multiply them by the number of people within that group. Incorporating this, we find our equation for between-groups sum of squares (SS_B) to be:

$$SS_B = \sum n_j(M_j - M_G)^2$$

The subscript j refers to the "j^{th}" group where $j = 1…k$ to keep track of which group mean and sample size we are working with. As you can see, the only difference between this equation and

the familiar sum of squares for variance is that we are adding in the sample size. Everything else logically fits together in the same way.

14.7 Within-Groups Sum of Squares

The other source of variability in the figures—***within-groups variability***—comes from differences that occur within each group. That is, each individual deviates a little bit from their respective group mean, just like the group means differed from the grand mean. We therefore label this source the within-groups variance. Because we are trying to account for variance based on group-level means, any deviation from the group means indicates an inaccuracy or error. Thus, our within-groups variability represents our error in ANOVA.

The formula for this sum of squares is again going to take on the same form and logic. What we are looking for is the distance between each individual person and the mean of the group to which they belong. We calculate this deviation score, square it so that they can be added together, then sum all of them into one overall value for the within-groups sum of squares (SS_W):

$$SS_W = \sum (X_{ij} - M_j)^2$$

In this instance, because we are calculating this deviation score for each individual person, there is no need to multiply by how many people we have. The subscript j again represents a group and the subscript i refers to a specific person. So, X_{ij} is read as "the i^{th} person of the j^{th} group." It is important to remember that the deviation score for each person is only calculated relative to their group mean; do not calculate these scores relative to the other group means.

Total Sum of Squares

Total sum of squares (SS_T) can also be computed as a check for our calculations of between-groups and within-groups sums of squares. The calculation for this score is exactly the same as it would be if we were calculating the overall variance in the dataset (because that is what we are interested in explaining) without worrying about or even knowing about the groups into which our scores fall:

$$SS_T = \sum (X_i - M_G)^2$$

We can see that our total sum of squares is just each individual score minus the grand mean. As with our within-groups sum of squares, we are calculating a deviation score for each individual person, so we do not need to multiply anything by the sample size; that is only done for a between-groups sum of squares.

An important feature of the sums of squares in ANOVA is that they all fit together. We could work through the algebra to demonstrate that if we added together the formulas for SS_{B} and SS_{W}, we would end up with the formula for SS_{T}. That is:

$$SS_T = SS_B + SS_W$$

This will prove to be very convenient, because if we know the values of any two of our sums of squares, it is very quick and easy to find the value of the third. It is also a good way to check calculations: if you calculate each *SS* by hand, you can make sure that they all fit together as shown above, and if not, you know that you made a math mistake somewhere.

We can see from the above formulas that calculating an ANOVA by hand from raw data can take a very, very long time. For this reason, unless your professor requires it, you will not be required to calculate the *SS* values by hand and will let the statistical computer program do it for you. However, you should still take the time to understand how they fit together and what each one represents to ensure you understand the reasons behind the analysis itself.

14.8 ANOVA Table

All of our sources of variability fit together in meaningful, interpretable ways as we saw above, and the easiest way to show these relationships is to organize them in a table. The ANOVA table (Table 14.1) shows how we calculate the *df*, *MS*, and *F* values. The first column of the ANOVA table, labeled "Source," indicates which of our sources of variability we are using: between groups (B), within groups (W), or total (T). The second column, labeled "*SS,*" contains our values for the sum of squared deviations, also known as the sum of squares that we reviewed above.

Table 14.1: A Sample ANOVA Table. The sample ANOVA table shows the theoretical formulas for variation between (sums of squares between, divided by k-1 degrees of freedom, which gives mean squares between) and within (sums of squares within, divided by N-k degrees of freedom, which gives mean squares within). Mean squares between is divided by mean squares within to calculate the F-value.

Source	***SS***	***df***	***MS***	***F***
Between	SS_{B}	$k-1$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_W}$
Within	SS_{W}	$N-k$	$\frac{SS_W}{df_W}$	
Total	SS_{T}	$N-1$		

Table 14.1 A Sample ANOVA Table[4]

As noted previously, calculating these by hand takes awhile and is not the focus of many of our instructors, so the formulas are not presented in Table 14.1. However, remember that SS_{T} is the sum of SS_{B} and SS_{W}, in case you are only given two *SS* values and need to calculate the third.

The next column, labeled "*df,*" is our degrees of freedom. As with the sums of squares, there is a different *df* for each group, and the formulas are presented in the table. Total degrees of freedom is calculated by subtracting 1 from the overall sample size (N). (Remember, the capital N in the *df* calculations refers to the overall sample size, not a specific group sample size.) Notice that df_{T}, just like for total sums of squares, is the Between (df_{B}) and Within (df_{W}) rows added together. If you take $N - k + k - 1$, then the "$- k$" and "$+ k$" portions will cancel out, and you are left with $N - 1$. This is a convenient way to quickly check your calculations.

The third column, labeled "*MS*," shows our mean squared deviation for each source of variance. A ***mean square*** is just another way to say variability and is calculated by dividing the sum of squares by its corresponding degrees of freedom. Notice that we show this in the ANOVA table for the Between row and the Within row, but not for the Total row. There are two reasons for this. First, our Total mean square would just be the variance in the full dataset (put together the formulas to see this for yourself), so it would not be new information. Second, the mean square values for Between and Within would not add up to equal the Total mean square because they are divided by different denominators. This is in contrast to the first two columns, where the Total row was both the conceptual total (i.e., the overall variance and degrees of freedom) and the literal total of the other two rows.

The final column in the ANOVA table, labeled "*F*," is our test statistic for ANOVA. The *F*-statistic, just like a *t*- or *z*-statistic, is compared to a critical value to see whether we can reject for fail to reject a null hypothesis. Thus, although the calculations look different for ANOVA, we are still doing the same thing that we did in all of our *t*-tests. We are simply using a new type of data to test our hypotheses. We will see what these hypotheses look like shortly, but first, we must take a moment to address why we are doing our calculations this way.

Now that we have our hypotheses for ANOVA, an example is in order. We will continue to use the data from Figure 14.1, Figure 14.2, and Figure 14.3 for continuity.

14.9 Example One-Way ANOVA (Between): Scores on Job-Application Tests (Including Effect Size)

Our data come from three groups of 10 people each, all of whom applied for a single job opening: those with no college degree, those with a college degree that is not related to the job opening, and those with a college degree from a relevant field. We want to know if we can use this group membership to account for our observed variability and, by doing so, test if there is a difference between our three group means. We will start, as always, with our hypotheses.

14.9.1 Step 1: Look at the Data and State the Null and Research Hypotheses

14.9.1.1 Data Set-Up

To conduct a one-way ANOVA, the dataset needs one continuous outcome variable and one categorical grouping variable with three or more independent groups. Each row should represent one participant or unit of analysis, and each participant should belong to only one group.

14.9.1.2 Describe the Data

Once we confirm that the data are set up correctly, we should describe the outcome variable overall and within each group. For a one-way ANOVA, the group means, standard deviations, sample sizes, and visual distributions are especially important because the test compares the groups on the outcome variable.

14.9.1.3 Hypotheses in ANOVA

So far, we have seen what ANOVA is used for, why we use it, and how we use it. Now, we can turn to the formal hypotheses we will be testing. As with before, we have a null and an alternative hypothesis to lay out. Our null hypothesis is still the idea of "no difference" in our data. Because we have multiple group means, we simply list them out as equal to each other:

H_0: There is no difference in the group means. (H_0: m 1 = m 2 = m3).

We list as many m parameters as groups we have. In the example above, we have three groups to test, so we have three parameters in our null hypothesis. If we had more groups, say, four, we would simply add another m to the list and give it the appropriate subscript, giving us:

H_0: There is no difference in the group means. (H_0: m 1 = m 2 = m3 = m4).

Notice that we do not say that the means are all equal to zero, we only say that they are equal to one another; it does not matter what the actual value is, so long as it holds for all groups equally.

Our alternative hypothesis for ANOVA is a little bit different. Let us take a look at it and then dive deeper into what it means:

H_A: At least one mean is different from the others.

The first difference is obvious: Although some statisticians will use a mathematical operant here to show the scores are not equal ($\mu_1 \neq \mu_2 \neq \mu_3$), there is no exact mathematical statement of the alternative hypothesis in ANOVA. This is due to the second difference: We are not saying *which* group is going to be different, only that *at least one* will be. Because we do not hypothesize about which mean will be different, there is no way to write it precisely mathematically (although some will separate the means with "not equal" signs as an approximation).

Similarly, we do not have directional hypotheses (greater than or less than) like we did in *t*-tests or correlations. Due to this, our alternative hypothesis is always exactly the same: At least one mean is different from the others.

In an earlier chapter, we saw that if we reject the null hypothesis, we can adopt or support the alternative, and this made it easy to understand what the differences looked like. In ANOVA, we will still support the alternative hypothesis as the best explanation of our data if we reject the null hypothesis. However, when we look at the alternative hypothesis, we can see that it does not give us much information. We will know that a difference exists somewhere, but we will not know *where* that difference is. Is only Group 1 different, but Groups 2 and 3 are the same? Is only Group 2 different? Are all three of them different? Based on our alternative hypothesis, there is no way to be sure just from the *F*-test. Shortly, we will see how to find out specific differences. For now, just remember that we are testing for *any* difference in group means, and it does not matter where that difference occurs.

Our hypotheses are concerned with the means of groups based on education level, so:

H_0: There is no difference between the means of the education groups. $H_0 : \mu_1 = \mu_2 = \mu_3$

H_A: At least one mean is different.

Again, we phrase our null hypothesis in terms of what we are actually testing, and we use a number of population parameters equal to our number of groups. Our alternative hypothesis is always exactly the same.

14.9.2 Step 2: Check Assumptions and Set the Critical Value(s)

Assumptions of a One-Way ANOVA Test (Between)

As we already stated, the purpose of a one-way ANOVA test is to determine the existence of a statistically significant difference among several group means. The test actually uses variances to help determine whether the means are equal or not. In order to perform a one-way ANOVA test, there are five basic assumptions to be fulfilled:

1. Each population from which a sample is taken is assumed to be normal. This assumption is known as the ***normality*** assumption. If you are using a statistical program, this can be found in the ANOVA output, the normality test and Q-Q plot are based on the model residuals. You can also examine the skew/kurtosis and histogram using Descriptives.
2. All samples are randomly selected and ***independent*** from each other (no one repeats in the groups).
3. The populations are assumed to have equal standard deviations (or variances), otherwise known as ***homogeneity of variance*** (the distribution of scores for one group does not vary in shape or size from the other groups).

4. The sorting variable (also known as the predictor or independent variable) is a categorical variable (nominal or ordinal scale).
5. The outcome (also known as the dependent variable) is a continuous variable (interval or ratio scale).

14.9.2.1 Testing Normality

We evaluate normality using the four methods introduced earlier: the Shapiro-Wilk test, Q-Q plot, skew and kurtosis values, and visual inspection of the distribution. For a one-way ANOVA, we are interested in whether the outcome is approximately normal within groups. We should also calculate z-scores for skew and kurtosis by dividing each value by its standard error; absolute values below approximately 1.96 do not suggest a substantial departure from normality. Finally, we should examine grouped histograms or box plots. Overall, the normality assumption appears reasonable for this example.

14.9.2.2 Testing Homogeneity of Variance

We evaluate homogeneity of variance using Levene's test and by comparing the variability of the outcome across groups. If the homogeneity of variance assumption was violated, we would move to a non-parametric version of the test like Welch's ANOVA or the Kruskal-Wallis test (but that will go beyond the scope of this course).

14.9.2.3 Setting the Critical Value(s)

Our test statistic for ANOVA, as we saw above, is F. Because we are using a new test statistic, we will get a new table: the F distribution table. A segment of the F-distribution table is included below. A link to a complete F-distribution table can be found in the Appendix.

The F table we utilize only displays critical values for $\mathbf{a} = .05$. This is because other significance levels are uncommon for beginning statisticians, and so it is not worth it to use up the space to present them. If you search online for other F-tests, you will come across ones with more **a** levels. There are now two degrees of freedom we must use to find our critical value: numerator and denominator. These correspond to the numerator and denominator of our test statistic, which, if you look at the ANOVA table presented earlier (Table 14.1), are our Between and Within rows, respectively. The df_{B} is the "*df:* Numerator (Between)" because it is the degrees of freedom value used to calculate the Mean Square Between, which in turn is the numerator of our F statistic. Likewise, the df_{W} is the "*df:* Denominator (Within)" because it is the degrees of freedom value used to calculate the Mean Square Within, which is our denominator for F.

The formula for df_{B} is $k - 1$; remember that k is the number of groups we are assessing. In this example, $k = 3$ so our $df_{\mathrm{B}} = 2$. This tells us that we will use the second column, the one

labeled 2, to find our critical value. To find the proper row, we simply calculate the df_W, which was $N - k$. The original prompt told us that we have "three groups of 10 people each," so our total sample size is 30. This makes our value for $df_W = 27$. If we follow the second column down to the row for 27, we find that our critical value is 3.35. We use this critical value the same way as we did before: it is our criterion against which we will compare our obtained test statistic to determine statistical significance.

Table 14.2: **Table 14.2** Sample Critical Values for F (F table)[5]

***df:* Denominator (Within)**	***df:* Numerator 1**	**2**	**3**	**4**	**5**	**6**
20	4.35	3.49	3.10	2.87	2.71	2.60
21	4.32	3.47	3.07	2.84	2.68	2.57
22	4.30	3.44	3.05	2.82	2.66	2.55
23	4.28	3.42	3.03	2.80	2.64	2.53
24	4.26	3.40	3.01	2.78	2.62	2.51
25	4.24	3.38	2.99	2.76	2.60	2.49
26	4.22	3.37	2.98	2.74	2.59	2.47
27	4.21	**3.35**	2.96	2.73	2.57	2.46

14.9.3 Step 3: Calculate & Report the Descriptive Statistics, Test Statistic, and Effect Size

Now that we have our hypotheses and the criteria we will use to test them, we can calculate our test statistic.

To do this, we will fill in the ANOVA table, working our way from left to right and filling in each cell to get our final answer, or we will utilize a statistical program to calculate these values. We will assume that we are given the *SS* values as shown below:

Table 14.3: **Table 14.3** Sample ANOVA table with Sums of Squares Between and Within[6]

Source	***SS***	***df***	***MS***	***F***
Between	8246			
Within	3020			
Total				

These may seem like random numbers, but remember that they are based on the distances between the groups themselves and within each group. Figure 14.4 shows the plot of the data with the group means and grand mean included. If we wanted to, we could use this information, combined with our earlier information that each group has 10 people, to calculate the between-groups sum of squares by hand. However, doing so would take some time, and without the

specific values of the data points, we would not be able to calculate our within-groups sum of squares, so we will trust that these values are the correct ones.

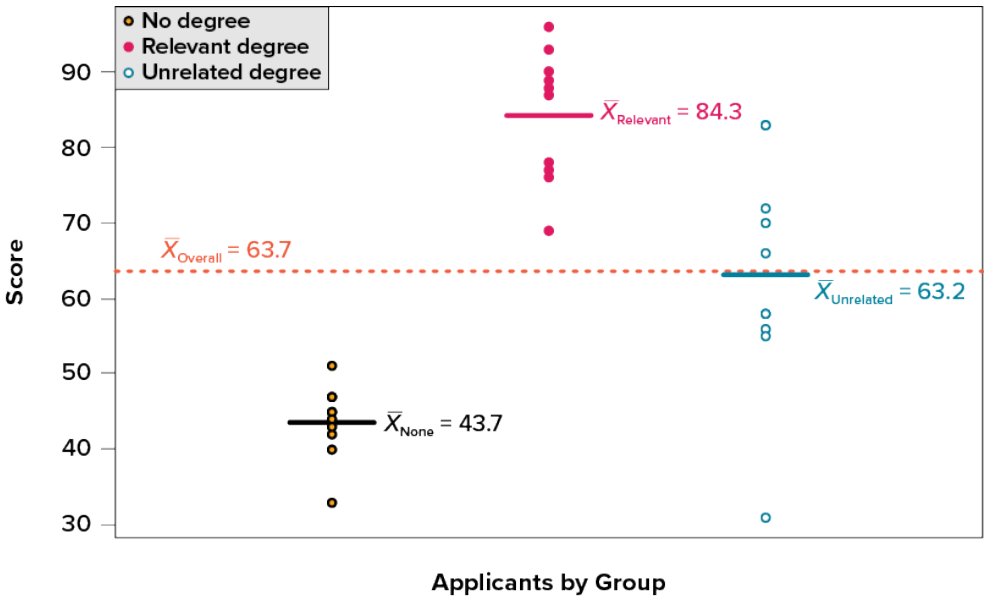


Figure 14.5: **Figure 14.4** Job Test Scores Group Means[7]

We were given the sums of squares values for our first two rows, so we can use those to calculate the total sum of squares.

Table 14.4: **Table 14.4** Sample ANOVA Table with Sums of Squares Between, Within, and Total[8]

Source	***SS***	***df***	***MS***	***F***
Between	8246			
Within	3020			
Total	11266			

We also calculated our degrees of freedom earlier, so we can fill in those values. Additionally, we know that the total degrees of freedom is $N - 1$, which is 29. This value of 29 is also the sum of the other two degrees of freedom, so everything checks out.

Table 14.5: **Table 14.5** Sample ANOVA Table with Sums of Squares Between, Within, and Total, Plus Degrees of Freedom [9]

Source	***SS***	***df***	***MS***	***F***
Between	8246	2		
Within	3020	27		
Total	11266	29		

Now we have everything we need to calculate our mean squares. Our *MS* values for each row are just the *SS* divided by the *df* for that row, giving us:

Table 14.6: **Table 14.6** Sample ANOVA table with sums of squares between, within, and total, plus degrees of freedom and mean squares values[10]

Source	***SS***	***df***	***MS***	***F***
Between	8246	2	4123	
Within	3020	27	111.85	
Total	11266	29		

Remember that we do not calculate a Total Mean Square, so we leave that cell blank. Finally, we have the information we need to calculate our test statistic. F is our MS_{B} divided by MS_{W}.

Table 14.7: **Table 14.7** Sample ANOVA table with sums of squares between, within, and total, plus degrees of freedom, mean squares values, and F-ratio[11]

Source	***SS***	***df***	***MS***	***F***
Between	8246	2	4123	36.86
Within	3020	27	111.85	
Total	11266	29		

So, working our way through the table, given only two *SS* values and the sample size and group size from before, we calculate our test statistic to be $F_{\text{obt}} = 36.86$, which we will compare to the critical value in Step 2.

14.9.4 Effect Size: Variance Explained

Recall that the purpose of ANOVA is to take observed variability and see if we can explain those differences based on group membership. To that end, our effect size will be just that: the variance explained. You can think of variance explained as the proportion or percent of the

differences we are able to account for based on our groups. We know that the overall observed differences are quantified as the total sum of squares, and that our observed effect of group membership is the between-groups sum of squares. Our effect size, therefore, is the ratio of these two sums of squares. Specifically:

$$\eta^2 = \frac{SS_B}{SS_T}$$

Figure 14.6: The formula for Eta-squared: Eta-squared, written as the Greek letter with a superscript 2, equals the Sum of Squares Between (SS subscript B) all over the Sum of Squares Total (SS subscript T). Eta-squared is a measure of effect size, representing the proportion of total variance explained by a factor in ANOVA.

The effect size ^{2} is called "eta-squared" and represents variance explained. For our example, our values give an effect size of:

$$\eta^2 = \frac{8246}{11266} = .73$$

Figure 14.7: The formula eta squared equals the fraction 8246 divided by 11266, which simplifies to 0.73.

So, we are able to explain 73% of the variance in job-test scores based on education. This is, in fact, a huge effect size, and most of the time we will not explain nearly that much variance. Our guidelines for the size of our effects (Cohen, 1988) are:

Table 14.8: **Table 14.8** ANOVA effect size (eta-squared) values and their interpretations in English[12]

η^2	**Size**
.01	Small
.06	Medium
.14	Large

So, we found that not only do we have a statistically significant result, but that our observed effect was very large! However, and this is super important: *We still do not know specifically which groups are different from each other.* It could be that they are all different, or that only those job seekers who have a relevant degree are different from the others, or that only those

who have no degree are different from the others. To find out which is true, we need to do a special analysis called a post hoc test.

14.9.5 Step 4: Interpret Your Findings in Relation to the Hypothesis (Translate Math to English)

Remember, if we are looking for differences, we want a big F-in value. Oops! We mean a big F-value (where did your mind go with this?!?). If the calculated F-score is bigger (more extreme) than the critical F-score, then you reject the null hypothesis. Another way of saying this is that if the calculated F-score is to the right (in the shaded critical area), then the null hypothesis should be rejected. This means that there is a small probability (less than 5%, $p < .05$) that all of the means are similar (suggesting that at least one mean is different from one other mean). In contrast, if the calculated F-score is smaller than the critical value (to the left of the shaded area), then the null hypothesis is retained; there is a large probability (larger than 5%. $p > .05$) that the group means are similar.

$Critical < |Calculated|$ =***Reject null : Means are different.*** $p < .05$

$Critical > |Calculated|$ =***Retain null : Means are similar.*** $p \geq .05$

Our obtained test statistic was calculated to be $F = 36.86$ and our critical value was found to be $F = 3.35$. Our obtained statistic is larger than our critical value, so we can reject the null hypothesis.

> The results of the one-way ANOVA between indicated that statistically significant differences in job skills test scores existed for applicants in each of the three education groups: no degree ($M = 43.7$), unrelated degree ($M = 63.2$), relevant degree ($M = 84.3$). Additionally, the effect size was large, $F(2, 27) = 36.86$, $p < .05$, $h^2 = .73$. We reject the null hypothesis; the alternate hypothesis is supported. Post hoc tests (see the next section) were performed to determine where the differences were.

Notice that when we report F, we include both degrees of freedom. We always report the numerator and then the denominator, separated by a comma and space. If we are calculating with a computer program, we should also provide the standard deviation for each group with its mean. We must also note that, because we were only testing for any difference, we cannot yet conclude *which groups* are different from the others. To do so, we need to perform what is called a post hoc test.

14.10 Post Hoc Tests

A ***post hoc test*** is used only *after* we find a statistically significant result when we need to determine *where* our differences truly came from. The term *post hoc* comes from the Latin for "after the event." Many different post hoc tests have been developed, and most of them will

give us similar answers. We will only focus here on the most commonly used ones. We will also only discuss the concepts behind each and will not worry about calculations.

14.10.1 Bonferroni Test

A Bonferroni test is perhaps the simplest post hoc analysis. A ***Bonferroni test*** is a series of *t*-tests performed on each pair of groups. As we discussed earlier, the number of groups quickly increases the number of comparisons, which inflates Type I error rates. To avoid this, a Bonferroni test divides our significance level **a** by the number of comparisons we are making so that when they are all run, they sum back up to our original Type I error rate. Once we have our new significance level, we simply run independent samples *t*-tests to look for differences between our pairs of groups. This adjustment is sometimes called a Bonferroni Correction, and it is easy to do by hand if we want to compare obtained *p*-values to our new corrected **a** level, but it is more difficult to do when using critical values like we do for our analyses, so we will leave our discussion of it to that.

14.10.2 Tukey's Honestly Significant Difference

Tukey's Honestly Significant Difference (HSD) is a popular post hoc analysis that, like Bonferroni, makes adjustments based on the number of comparisons; however, it makes adjustments to the test statistic when running the comparisons of two groups. These comparisons give us an estimate of the difference between the groups and a confidence interval for the estimate. We use this confidence interval in the same way we use a confidence interval for a regular independent samples *t*-test: if it contains 0.00, the groups are not different, but if it does not contain 0.00 then the groups are different. However, the sample sizes of all groups need to be equivalent to run the Tukey's test appropriately. It is not a good choice for a post hoc test when there are unequal sample sizes.

Below are the differences between the group means and the Tukey's HSD confidence intervals for the differences:

Table 14.9: **Table 14.9** Pairwise comparison values of the test-takers, the difference between the scores, and the Tukey's HSD confidence intervals[13]

Comparison	Difference	Tukey's HSD CI
None vs. relevant	40.60	(28.87, 52.33)
None vs. unrelated	19.50	(7.77, 31.23)
Relevant vs. unrelated	21.10	(9.37, 32.83)

As we can see, none of these intervals contain 0.00, so we can conclude that all three groups are different from one another. Thus, we would add to our interpretation from before:

The results of the one-way ANOVA between indicated that statistically significant differences in job skills test scores existed for applicants in each of the three education groups: no degree ($M = 43.7$), unrelated degree ($M = 63.2$), relevant degree ($M = 84.3$). Additionally, the effect size was large, $F(2, 27) = 36.86$, $p < .05$, $\blacksquare^2 = .73$. We reject the null hypothesis; the alternate hypothesis is supported. A Tukey HSD post hoc test was performed to determine the differences further. People with no degree scored significantly lower on the test than those with an unrelated degree. People with a relevant degree scored significantly higher than those with either no degree or an unrelated degree.

14.10.3 Scheffé Test

Another common post hoc test is the Scheffé test. Like Tukey's HSD, the ***Scheffé test*** adjusts the test statistic for how many comparisons are made, but it does so in a slightly different way. The result is a test that is "conservative," which means that it is less likely to commit a Type I error, but this comes at the cost of less power to detect effects. We can see this by looking at the confidence intervals that the Scheffé test gives us:

Table 14.10: **Table 14.10** Pairwise comparison values of the test-takers, the difference between the scores, and the Tukey's HSD confidence intervals[14]

Comparison	Difference	Scheffé CI
None vs. relevant	40.60	(28.35, 52.85)
None vs. unrelated	19.50	(7.25, 31.75)
Relevant vs. unrelated	21.10	(8.85, 33.35)

As we can see, these are slightly wider than the intervals we got from Tukey's HSD. This means that, all other things being equal, they are more likely to contain zero. In our case, however, the results are the same, and we again conclude that all three groups differ significantly from one another.

14.10.4 Fisher's Least Significant Difference (LSD) Test

The first post hoc test created was the Fisher's Least Significant Difference (LSD) test. ***Fisher's LSD test*** is more flexible and has more power, but the trade-off is that it is less "conservative," so it is more likely to commit a Type I error (saying a difference exists in the population when in reality it does not) than other types of post hoc tests. It is also best when there are three and only three groups. Thus, professors may allow students to use Fisher's LSD to practice interpreting hypothesis tests, but it is not used as often in published research.

14.10.5 Which Post Hoc Test Do We Choose?

There are many more post hoc tests than just these four, and they all approach the task in different ways, with some being more conservative and others being more powerful. In general, though, they will give highly similar answers. Check first with your professor to see if they have a preference. Some professors will prefer Fisher's LSD for beginners, because it is more flexible and gives you the most practice at interpreting differences between groups. Some prefer Scheffé or Tukey because it is more like what producers and publishers of research do.

What is important here is to be able to interpret a post hoc analysis. If you are given post hoc analysis confidence intervals, like the ones seen above, read them the same way you would read confidence intervals for *t*-tests. If they contain zero, there is no difference; if they do not contain zero, there is a difference.

If you run the post hoc test in a computer program like SPSS, you can also look at the mean difference between the two groups in question (one mean subtracted from the other) and whether or not this pair is statistically significant (p [<] .05).

14.11 Example One-Way ANOVA (Between), with Post-Hoc Testing: Mood Improvement in a Clinical Drug Trial

For this second example, we will work with a hypothetical dataset from a clinical trial testing a new antidepressant drug called *Joyzepam*. Participants were randomly assigned to one of three drug conditions: a placebo, an existing antidepressant/anxiety medication called *Anxifree*, or the new medication *Joyzepam*. The outcome variable is mood.gain, which measures improvement in mood after three months. Scores range from -5 to +5, with higher scores indicating greater improvement. The grouping variable for this chapter is drug, which has three groups: placebo, anxifree, and joyzepam.

Step 1: Look at the Data and State the Null and Research Hypotheses

14.11.1 Data Set-Up

To conduct a one-way ANOVA (between), the dataset needs one continuous outcome variable and one categorical grouping variable with three or more independent groups. Each row should represent one participant or unit of analysis, and each participant should belong to only one group. In this example, mood.gain is the continuous outcome variable and drug is the categorical grouping variable. Because drug has three groups, a one-way ANOVA is more appropriate than an independent *t*-test.

	drug	mood.gain
1	placebo	.50
2	placebo	.30
3	placebo	.10
4	joyzepam	.60
5	joyzepam	.40
6	joyzepam	.20
7	anxifree	1.40
8	anxifree	1.70
9	anxifree	1.30
10	placebo	.60
11	placebo	.90
12	placebo	.30
13	joyzepam	1.10
14	joyzepam	.80
15	joyzepam	1.20
16	anxifree	1.80
17	anxifree	1.30
18	anxifree	1.40

Figure 14.8: **Figure 14.5** Participant Drug Type and Mood Gains [15]

14.11.2 Describe the Data

Once we confirm that the data are set up correctly in our stats program, we should describe the outcome variable overall and within each group. For a one-way ANOVA, the group means, standard deviations, sample sizes, and visual distributions are especially important because the test compares the groups on the outcome variable.

In this example, there are 18 participants, with 6 participants in each drug condition. This is a ***balanced design*** because the groups have the same sample size. Balanced designs are helpful because ANOVA tends to behave better when group sizes are equal or similar.

We should also examine the distribution of mood.gain across the three drug conditions. A grouped box plot or histogram (or a box plot or histogram for each group separately, using select cases) is especially useful because it shows the center, spread, overlap, and possible outliers for each group. Visually, joyzepam appears to have higher mood gain than the other two conditions, but we need the ANOVA to test whether the overall group difference is statistically significant.

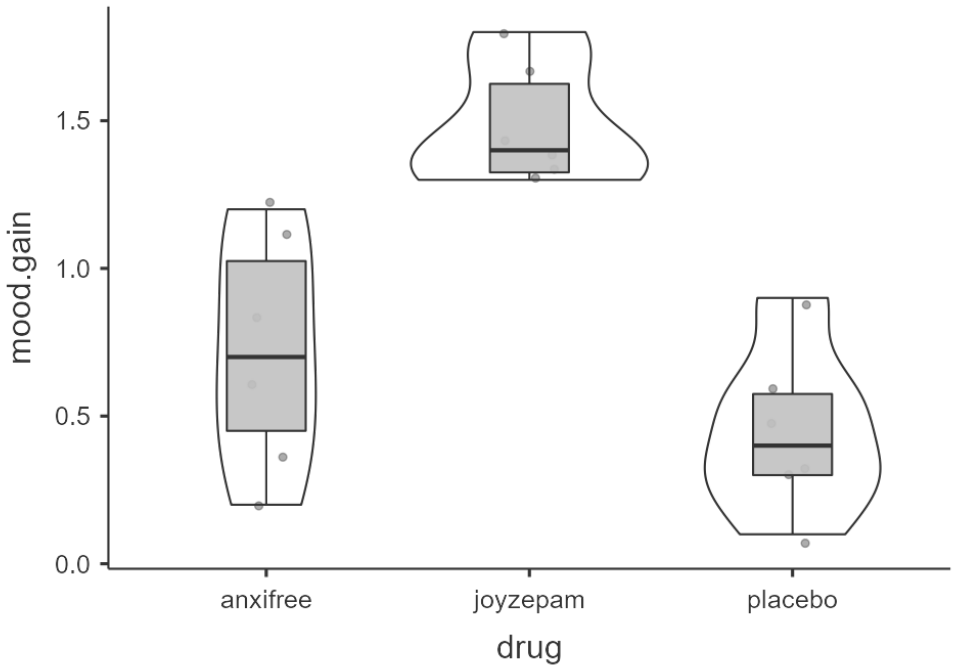


Figure 14.9: **Figure 14.6** Grouped box plot of mood gain scores by drug type[16]

14.11.2.1 *Specify* the Hypotheses

Our research question is: Do participants in the three drug conditions differ in mood gain? This research question is non-directional because the one-way ANOVA tests whether there is a difference somewhere among the three groups. Therefore, our hypotheses are:

H_0: There is no difference in mood gain between the three drug conditions. $H_0 : \mu_1 = \mu_2 = \mu_3$

H_A: There is a difference in mood gain between at least two of the three drug conditions.

We will use the conventional probability, a = .05. Therefore, we will consider the result statistically significant if the calculated p-value is less than .05.

14.11.3 Step 2: Check Assumptions and Set the Critical Value(s)

As a parametric test, the one-way ANOVA has several assumptions:

1. The outcome variable is **approximately normally distributed within each group**. In ANOVA output, the normality test and Q-Q plot are based on the model residuals. You can also examine the skew/kurtosis and histogram using Descriptives.
2. The groups have roughly equal variances, which is called **homogeneity of variance**.
3. The outcome variable is **interval or ratio** (i.e., continuous).
4. Observations are **independent**. Each participant or case should contribute one score and belong to only one group.

We cannot test the third and fourth assumptions using the output alone; those assumptions are based on how the data were measured and collected. However, we can and should evaluate the first two assumptions.

14.11.3.1 ANOVA Is Somewhat Robust to Assumption Violations

ANOVA is often described as ***robust***, which means it can still perform reasonably well when assumptions are not perfectly met. However, robustness has limits. ANOVA is most robust when group sizes are equal or similar and group variances are equal or similar.

Assumption violations become more concerning when group sizes are very unequal, variances are very unequal, sample sizes are small, or the distributions are strongly non-normal. When assumptions are seriously violated, we may need to use an alternative test, such as Welch's ANOVA or the Kruskal-Wallis test.

14.11.3.2 Testing Normality

We evaluate normality using the four methods introduced earlier: the Shapiro-Wilk test, Q-Q plot, skew and kurtosis values, and visual inspection of the distribution. For a one-way ANOVA, we are interested in whether the outcome is approximately normal within groups. In jamovi's ANOVA output, the normality test and Q-Q plot evaluate the model residuals.

The Shapiro-Wilk test was not statistically significant ($W = .96$, $p = .605$), which means the test does not provide evidence that the residuals differ from normality. The points in the Q-Q plot are also fairly close to the diagonal line. We should also calculate z-scores for skew and kurtosis by dividing each value by its standard error; absolute values below approximately 1.96 do not suggest a substantial departure from normality. Finally, we should examine grouped histograms, box plots, or violin plots of mood.gain by drug. Overall, the normality assumption appears reasonable for this example.

Assumption Checks

Normality Test (Shapiro-Wilk)

	W	p
mood.gain	0.96	0.605

Note. A low p-value suggests a violation of the assumption of normality

Homogeneity of Variances Test (Levene's)

	F	df1	df2	p
mood.gain	1.45	2	15	0.266

[3]

Figure 14.10: **Figure 14.7** Normality test and Levene's test of homogeneity of variance output indicate no violations of assumptions [17]

Plots

mood.gain

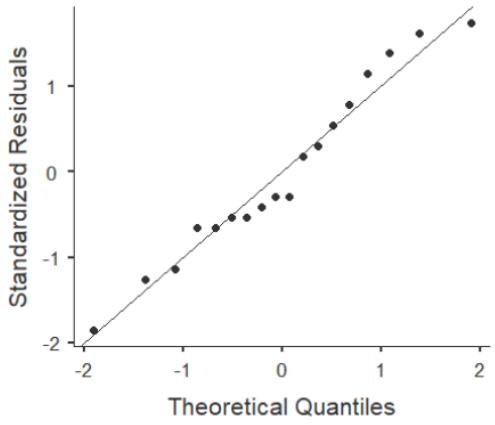


Figure 14.11: **Figure 14.8** Q-Q plot showing mood.gain residuals are approximately normally distributed[18]

14.11.3.3 Testing Homogeneity of Variance

We evaluate homogeneity of variance using Levene's test and by comparing the variability of the outcome across groups. Levene's test was not statistically significant, $F(2, 15) = 1.45$, $p = .266$. This means the test does not provide evidence that the group variances differ substantially, so the homogeneity-of-variance assumption appears reasonable.

That said, the sample size is small ($N = 18$), so assumption checks should be interpreted cautiously. When sample sizes are small, it is especially helpful to examine the group standard deviations and grouped plots rather than relying on Levene's test alone.

14.11.3.4 Setting the Critical Value

Once we know the sample size (N), degrees of freedom between ($df_{b,}$ k – 1), degrees of freedom within ($df_{w,}$ N - k), and probability level (a), we use this information by looking at a probability table for the test we are using (in this case F) to see what critical F-value we will need to label a test as statistically significant. Please see Appendix A for links to probability tables. In this particular case, let's use a standard a = .05 (which we will do all semester, unless your professor requires otherwise). For a sample size of 18, the degrees of freedom between are k - 1, or 3-1 = 2. The degrees of freedom within are N – k, or 18 – 3 = 15. The critical value is 3.68. Keep in mind if you are using a statistical program like jamovi or SPSS, it will be utilizing the critical value for the calculations, but it will not show it to you.

APPENDIX C: CRITICAL VALUES FOR F (F TABLE)

Note: The critical values in this table are for $p = .05$.

df: Denominator (Within)	*df:* Numerator (Between) 1	2	3	4	5	6	7	8	9	10	11	12	14	16	20
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12

Figure 14.12: **Figure 14.9** Critical value (F) for 2, 15 degrees of freedom is 3.68[19]

14.11.4 Step 3: Calculate & Report the Descriptive Statistics, Test Statistic, and Effect Size

14.11.4.1 Decide Whether to Use ANOVA, Welch's ANOVA, or Kruskal-Wallis

The appropriate test depends on whether the assumptions are reasonably met. One would typically use the assumption checks to decide which test to report. Welch's ANOVA is the alternative when group variances are unequal but the outcome is still approximately normal. The Kruskal-Wallis test is the nonparametric alternative when the normality assumption is seriously violated.

Table 14.11: **Table 14.11** Deciding between One-way ANOVA, Welch's ANOVA, and Kruskal-Wallis test[20]

Assumption Pattern	Test to Report
Normality is reasonably met and homogeneity of variance is reasonably met	One-way ANOVA
Normality is reasonably met but homogeneity of variance is not met	Welch's ANOVA
Normality is seriously violated and no appropriate transformation addresses the issue	Kruskal-Wallis test

Welch's ANOVA and Kruskal-Wallis go beyond the scope of this course. For our example, we will be using the one-way ANOVA (between). Depending on your professor, you might calculate the ANOVA by hand or using statistical software such as SPSS, SAS, R, Excel, or jamovi. Once your data are calculated, we will combine Step 3 with Step 4 below.

14.11.4.2 A Note About Directional Hypotheses in ANOVA

Remember: The overall ANOVA is not directional in the same way a *t*-test can be directional. The *F* statistic tests whether there is a difference somewhere among the groups, but it does not test whether one specific group is higher or lower than another specific group.

Oneway

[DataSet2] C:\Ginger\Research\ALG OER grant_PSYC 3141\ALG OER_CH14_ANOVA between.sav

Descriptives

mood.gain

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
placebo	6	.4500	.28107	.11475	.1550	.7450	.10	.90
joyzepam	6	.7167	.39200	.16003	.3053	1.1280	.20	1.20
anxifree	6	1.4833	.21370	.08724	1.2591	1.7076	1.30	1.80
Total	18	.8833	.53385	.12583	.6179	1.1488	.10	1.80

Figure 14.13: **Figure 14.10** Image of an SPSS Descriptives Table Showing Mean Gains in Mood by Drug Type[21]

ANOVA

mood.gain

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.453	2	1.727	18.611	<.001
Within Groups	1.392	15	.093		
Total	4.845	17			

ANOVA Effect Sizes[a]

		Point Estimate	95% Confidence Interval Lower	Upper
mood.gain	Eta-squared	.713	.333	.815
	Epsilon-squared	.674	.244	.791
	Omega-squared Fixed-effect	.662	.234	.781
	Omega-squared Random-effect	.495	.132	.641

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

Figure 14.14: **Figure 14.11** Image of an SPSS ANOVA and Effect Sizes Table for Gains in Mood by Drug Type [22]

14.11.5 Step 4: Interpret Your Findings in Relation to the Hypothesis (Translate Math to English)

Once we are satisfied that the assumptions for the one-way ANOVA are reasonably met, we can interpret the results.

The p-value is less than .05, so the result is statistically significant. We reject the null hypothesis of no group differences. The sample provides evidence that mood gain differs across the three drug conditions.

The effect size we will report for one-way ANOVA is eta-squared (h^2). Remember that values closer to 0 indicate a smaller effect, and values closer to 1 indicate a larger effect.

Because the one-way ANOVA is an omnibus test, this result tells us that at least two drug conditions differ from each other. It does not tell us which specific drug conditions differ. We need follow-up analyses to answer that question.

14.11.5.1 Write Up the Results in APA Style

An APA-style results section should describe the research question, summarize the relevant descriptive statistics, report the inferential test and effect size, and interpret the result.

> We hypothesized that there would be differences in mood gain across the drug treatment conditions. A one-way ANOVA (between) indicated that mood gain differed significantly across the three drug conditions, with a large effect, $F(2, 15) = 18.61$, $p < .001$, $\mathrm{h}^2 = .71$.

The two numbers in parentheses are the degrees of freedom for the test. For most students, the key interpretation is whether the p-value is less than alpha and what the result means for the research question. This write-up is not complete by itself because the ANOVA was statistically significant. We still need follow-up analyses to determine which drug conditions differ from each other.

14.11.5.2 Visualize the Results

For a one-way ANOVA, the graph should help readers compare the distribution of the continuous outcome across the groups. A grouped box plot is usually a good choice because it shows the median, spread, overlap between groups, and possible outliers. A bar plot with error bars can also be used to communicate group means, especially when the goal is to summarize the mean differences. Before interpreting follow-up tests, we should return to the descriptive statistics and graph. The follow-up tests tell us which differences are statistically significant, but the means and visualizations help us understand the direction and size of those differences.

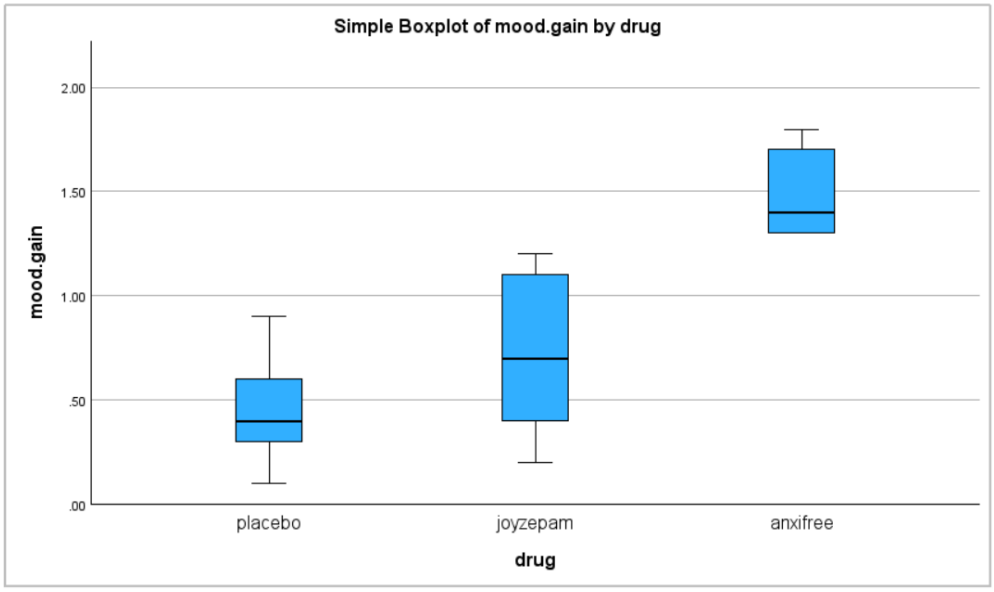


Figure 14.15: **Figure 14.12** Grouped boxplot of mood gains by drug type[23]

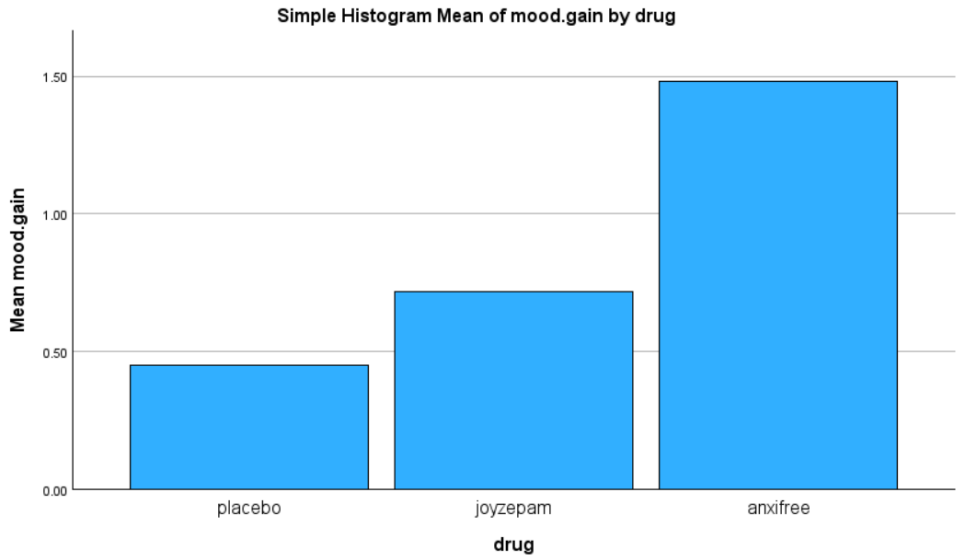


Figure 14.16: **Figure 14.13** Bar chart showing differences in mood gains by drug type[24]

14.11.5.3 Post Hoc Comparisons

Post hoc comparisons are used after a significant omnibus ANOVA when we want to compare the groups pair by pair. Because there are three groups, there are three pairwise comparisons:

1. placebo versus anxifree
2. placebo versus joyzepam
3. anxifree versus joyzepam

These comparisons are similar to conducting multiple independent t-tests, but they include corrections to control the familywise error rate. The ***familywise error rate*** is the probability of making at least one Type I error across a family of related tests. Tukey's HSD is a common post hoc procedure for one-way ANOVA when sample sizes are equal. It adjusts the pairwise comparisons so that the overall Type I error rate is controlled across the set of comparisons.

Post Hoc Tests

Multiple Comparisons

Dependent Variable: mood.gain

Tukey HSD

(I) drug	(J) drug	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
placebo	joyzepam	-.26667	.17586	.312	-.7235	.1901
	anxifree	-1.03333*	.17586	<.001	-1.4901	-.5765
joyzepam	placebo	.26667	.17586	.312	-.1901	.7235
	anxifree	-.76667*	.17586	.002	-1.2235	-.3099
anxifree	placebo	1.03333*	.17586	<.001	.5765	1.4901
	joyzepam	.76667*	.17586	.002	.3099	1.2235

*. The mean difference is significant at the 0.05 level.

Figure 14.17: **Figure 14.14** Tukey's highly significant difference (HSD) post hoc test for mood gains by drug type[25]

14.11.5.4 Write Up Post Hoc Results in APA Style

An APA-style results section should describe the research question, summarize the relevant descriptive statistics, report the inferential test and effect size, and interpret the result. When

an ANOVA is statistically significant, the write-up should also report the relevant follow-up comparisons.

> We hypothesized that there would be differences in mood gain across the drug treatment conditions. A one-way ANOVA (between) indicated that mood gain differed significantly across the three drug conditions, with a large effect, $F(2, 15) = 18.61$, $p < .001$, $h^2 = .71$. Tukey post hoc comparisons indicated that participants in the joyzepam condition ($M = 1.48$, $SD = 0.21$) had significantly greater mood gain than participants in the anxifree condition ($M = 0.72$, $SD = 0.39$), $p = .002$, and the placebo condition ($M = 0.45$, $SD = 0.28$), $p < .001$. Mood gain did not differ significantly between the anxifree and placebo conditions, $p = .312$.

This is not the only correct way to write the result. The key is to report the overall ANOVA, describe the group means, identify which pairwise comparisons were statistically significant, and avoid treating a pairwise p-value as if it belongs to only one group.

14.12 One-Way ANOVA (Within)(Repeated Measures or Across Time)

The ***one-way ANOVA (within)*** is a statistical test for significant differences between three or more groups where the same participants are in each group (or each participant is closely matched with participants in other experimental groups on some important third variable). For this reason, there should always be an equal number of scores (data points) in each experimental group. This type of design and analysis can also be called a related ANOVA, repeated measures ANOVA, or a within subjects ANOVA. The logic behind a one-way ANOVA within is very similar to that of a one-way ANOVA between. Remember that earlier we showed that in between subjects ANOVA, total variability is partitioned into between groups variability (SS_b) and within groups variability (SS_w), then divided by the respective degrees of freedom to give MS_b and MS_w (see Table 13.1), whereupon the F-ratio is calculated as:

$$F = \frac{MS_b}{MS_w}$$

In a one-way ANOVA (within), the F-ratio is calculated in a similar way, but whereas in an independent ANOVA the within-group variability (SS_w) is used as the basis for the denominator, in a repeated measures ANOVA the SS_w is partitioned into two parts. As we are using the same subjects in each group, we can remove the variability due to the individual differences between subjects (referred to as $SS_{subjects}$) from the within groups variability.

We will not go into too much technical detail about how this is done, but essentially each subject becomes a level of a factor called subjects. The variability in this within subjects factor is then calculated in the same way as any between subjects factor.

And then we can subtract $SS_{subjects}$\$from SS_{w}\$ to provide a smaller SS_{error} term:

$$\text{Between Subjects ANOVA: } SS_{error} = SS_w$$

$$\text{Repeated Measures ANOVA: } SS_{error} = SS_w - SS_{subjects}$$

This change in SS_{error} term often leads to a more powerful statistical test, but this does depend on whether the reduction in the SS_{error} more than compensates for the reduction in degrees of freedom for the error term (as degrees of freedom go from $(n-k)$ to $(n-1)(k-1)$ (remembering that there are more subjects in the independent ANOVA design).

Let's set the mathy-math stuff aside for a second and look at this again conceptually. A one-way ANOVA (within), also called a repeated measures ANOVA, is used to test whether three or more related measurement groups differ on one continuous (interval or ratio) outcome variable. We use a one-way ANOVA (within) when the same participants are measured across three or more time points, conditions, or tasks. A repeated measures ANOVA is similar to a dependent t-test, but it is used when there are three or more related measurements instead of two.

There are two common ways we might use a repeated measures ANOVA. First, the same participants might be measured on the same outcome at three or more time points. In that case, the repeated measures factor is time.

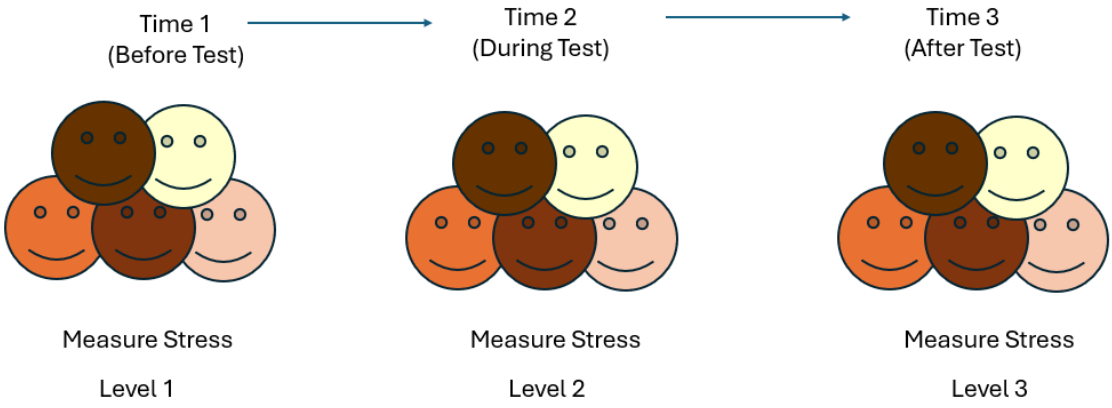


Figure 14.18: **Figure 14.15** A repeated measures ANOVA utilizing the same individuals measuring stress across three time periods (before, during, after a test)[26]

Second, the same participants might complete three or more conditions, tasks, or treatments. In that case, the repeated measures factor is condition, task, or treatment.

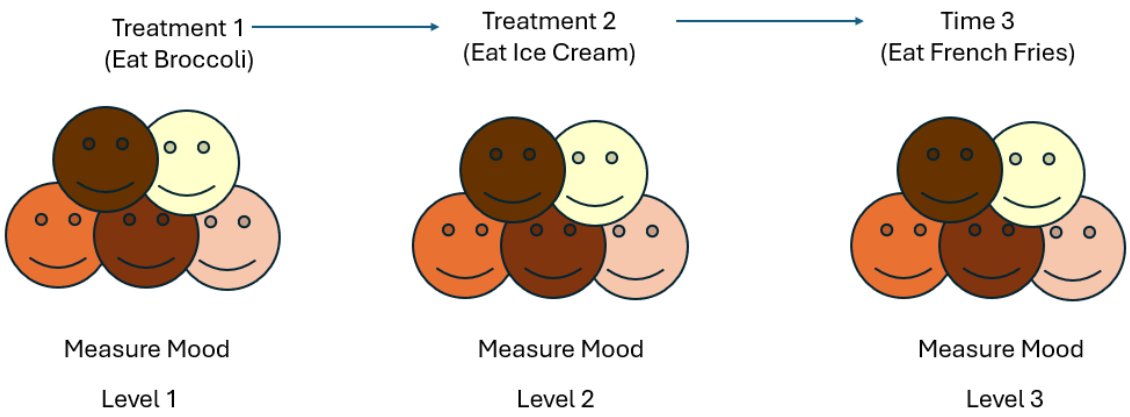


Figure 14.19: **Figure 14.16** A repeated measures ANOVA utilizing the same individuals measuring mood across three treatment conditions (eating broccoli, ice cream, and french fries)[27]

14.12.1 Step 1: Look at the Data and State the Null and Research Hypotheses

For this example, we will work with a hypothetical data from six patients with Broca's aphasia, a language difficulty that can occur after a stroke. Each patient completed three word-recognition tasks:

1. **Speech:** repeating single words read aloud by the researcher
2. **Conceptual:** matching pictures with their correct names
3. **Syntax:** reordering syntactically incorrect sentences

Each task included 10 attempts, and the score for each task is the number of attempts completed successfully. Because each patient completed all three tasks, the three task scores are related measurements. Our research question is: Do Broca's aphasia patients differ in word-recognition scores across the three tasks?

14.12.1.1 Data Set-Up

To conduct a one-way ANOVA within, the dataset needs multiple columns: one column for the first measurement, one column for the second measurement, etc. We need as many columns as we have tests or scores for each person. In this example, we need three columns because we have three scores. If you are using statistical software, you do not need a column called "participants" because the software will automatically assign a number to each person in the

data set. Each row will therefore represent one participant. The three scores in the same row are paired because they belong to the same participant or matched case.

speech	conceptual	syntax
8.00	7.00	6.00
7.00	8.00	6.00
9.00	5.00	3.00
5.00	4.00	5.00
6.00	6.00	2.00
8.00	7.00	4.00

Figure 14.20: **Figure 14.17** Word recognition scores across three types of tasks (speech, conceptual, syntax)[28]

14.12.1.2 Describe the Data

Once we confirm that the data are set up correctly, we should describe each related measurement. In this example, there are only six patients, so the dataset is very small. The descriptive statistics show that patients appear to score highest on the speech task and lowest on the syntax task. However, we need the repeated measures ANOVA to test whether the differences across tasks are statistically significant.

For repeated measures data, the pattern of scores within participants is important. A table of means and standard deviations is useful, but a graph showing the repeated measurements can help us see whether most participants show a similar pattern across tasks.

Descriptives

Descriptives

	Speech	Conceptual	Syntax
N	6	6	6
Missing	0	0	0
Mean	7.17	6.17	4.33
Median	7.50	6.50	4.50
Standard deviation	1.47	1.47	1.63
Minimum	5	4	2
Maximum	9	8	6

14.12.1.3

Figure 14.18 Descriptive data for the Broca's aphasia patient scores[29]

14.12.1.4 Specify the Hypotheses

Our research question is: Do Broca's aphasia patients differ in word-recognition scores across the three tasks? This research question is non-directional because the repeated measures ANOVA tests whether there is a difference somewhere across the related measurements. Therefore, our hypotheses are:

H_0: There is no difference in word-recognition scores across the three tasks. $H_0: \mu_1 = \mu_2 = \mu_3$

H_A: There is a difference in word-recognition scores between at least two of the tasks.

We will use the conventional probability, a = .05. Therefore, we will consider the result statistically significant if the p-value is less than .05.

14.12.2 Step 2: Check Assumptions and Set the Critical Value(s)

As a parametric test, the repeated measures ANOVA has several assumptions:

1. The outcome is **approximately normally distributed within the repeated measurements**, or the model residuals are approximately normal.

2. The variances of the differences between pairs of repeated measurements are roughly equal. This is called **sphericity**.
3. The outcome variable is **interval or ratio** (i.e., continuous).
4. Cases are **independent of other cases**. The repeated measurements within a row are related, but each participant or matched case should be independent of the other participants or matched cases.

We cannot test the third and fourth assumptions using the output alone; those assumptions are based on how the data were measured and collected. However, we can and should evaluate normality and sphericity.

14.12.2.1 Testing Normality

For repeated measures ANOVA, we are interested in whether the repeated-measures model is reasonably consistent with the normality assumption. You can perform a Q-Q plot. If the points fall reasonably close to the diagonal line, then the normality assumption appears reasonable.

When possible, you should also examine the distributions of the repeated measurements using descriptive statistics and graphs. This may include checking skew and kurtosis values and visually inspecting histograms or box plots for each repeated measurement. These checks do not replace the residual Q-Q plot, but they can help you better understand the data.

14.12.2.2 Testing Sphericity

Sphericity is the repeated measures ANOVA assumption that is most different from the tests we have used so far. Sphericity means that the variances of the differences between pairs of repeated measurements are approximately equal. For example, with three tasks, there are three sets of difference scores:

1. speech minus conceptual
2. speech minus syntax
3. conceptual minus syntax

The sphericity assumption asks whether the variability of those difference scores is roughly equal across the three pairs. This assumption only applies when there are three or more related measurements, which is why we did not discuss it for the dependent *t*-test.

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
test	.849	.657	2	.720	.868	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: test

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Figure 14.21: **Figure 14.19** Mauchly's test of sphericity for Broca's aphasias patients meets the sphericity assumption[30]

Mauchly's test of sphericity tests the null hypothesis that the sphericity assumption is met. If Mauchly's test is not statistically significant, the test does not provide evidence that sphericity has been violated. In our current example, Mauchly's test is not statistically significant, so the sphericity assumption appears reasonable.

If Mauchly's test is statistically significant, then the sphericity assumption is not met. In that case, we should use a sphericity correction, such as Greenhouse-Geisser or Huynh-Feldt, when interpreting the repeated measures ANOVA. For this course, use the following rule:

Table 14.12: **Table 14.12** Deciding between corrections for the ANOVA test depending upon Mauchly's test[31]

Sphericity Result	**What to Report**
Mauchly's test is not statistically significant	Report the uncorrected repeated measures ANOVA result
Mauchly's test is statistically significant and Greenhouse-Geisser ($< .75$)	Report the Greenhouse-Geisser corrected result
Mauchly's test is statistically significant and Greenhouse-Geisser ($> .75$)	Report the Huynh-Feldt corrected result

14.12.2.3 Setting the Critical Value

For one-way ANOVA (within), there are still two degrees of freedom we must use to find our critical value: numerator and denominator. As before, these correspond to the numerator and denominator of our test statistic. The df_B is the "*df:* Numerator (Between)" because it is the degrees of freedom value used to calculate the Mean Square Between, which in turn is the

numerator of our F statistic. The formula for df_B is $k - 1$; remember that k is the number of groups we are assessing. In this example, $k = 3$ so our $df_B = 2$. This tells us that we will use the second column, the one labeled 2, to find our critical value.

However, the denominator is different for ANOVA within than it was for ANOVA between. The denominator is now the df_{error}, which is calculated as $(n - 1)(k - 1)$. The original prompt told us that we have six patients ($n = 6$), so $(6 – 1 = 5)$. We have three testing conditions ($k = 3$), so $(3 – 1 = 2)$. If we multiply, 5 x 2 = 10. This makes our value for $df_{error} = 10$. If we follow the second column down to the row for 10, we find that our critical value is 4.10. We use this critical value the same way as we did before: it is our criterion against which we will compare our obtained test statistic to determine statistical significance. Keep in mind if you are using a statistical program like jamovi or SPSS, it will be utilizing the critical value for the calculations, but it will not show it to you.

APPENDIX C: CRITICAL VALUES FOR F (F TABLE)

Note: The critical values in this table are for $p = .05$.

df: Denominator (Within)	*df:* Numerator (Between) 1	2	3	4	5	6	7	8	9	10	11	12	14	16	20
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77

Figure 14.22: **Figure 14.20** Critical value (F) for 2, 10 degrees of freedom is 4.10[32]

14.12.3 Step 3: Calculate & Report the Descriptive Statistics, Test Statistic, and Effect Size

14.12.3.1 Decide Whether to Use Repeated Measures ANOVA or Friedman's Test

The appropriate test depends on whether the assumptions are reasonably met. In this course, use the assumption checks to decide which test to report.

Table 14.13: **Table 14.13** Deciding Between Repeated Measures Tests Based on Normality and Sphericity Assumptions[33]

Assumption Pattern	Test to Report
Normality appears reasonable and sphericity is met	Repeated measures ANOVA
Normality appears reasonable but sphericity is not met	Repeated measures ANOVA with the appropriate sphericity correction
Normality is seriously violated and no appropriate transformation addresses the issue	Friedman test

The Friedman test is the nonparametric alternative to the repeated measures ANOVA. It is used when the normality assumption is seriously violated for a design with three or more related measurements. The Friedman test goes beyond the scope of our course.

Depending on your professor, you might calculate the ANOVA by hand or using statistical software such as SPSS, SAS, R, Excel, or jamovi. Once your data are calculated, we will combine Step 3 with Step 4 below.

Descriptive Statistics

	Mean	Std. Deviation	N
speech	7.1667	1.47196	6
conceptual	6.1667	1.47196	6
syntax	4.3333	1.63299	6

Figure 14.23: **Figure 14.21** Descriptive statistics for Broca's aphasia patients language scores[34]

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
test	Sphericity Assumed	24.778	2	12.389	6.925	.013	.581	13.851	.819
	Greenhouse-Geisser	24.778	1.737	14.265	6.925	.018	.581	12.029	.770
	Huynh-Feldt	24.778	2.000	12.389	6.925	.013	.581	13.851	.819
	Lower-bound	24.778	1.000	24.778	6.925	.046	.581	6.925	.563
Error(test)	Sphericity Assumed	17.889	10	1.789					
	Greenhouse-Geisser	17.889	8.685	2.060					
	Huynh-Feldt	17.889	10.000	1.789					
	Lower-bound	17.889	5.000	3.578					

a. Computed using alpha = .05

Figure 14.24: **Figure 14.22** Repeated measures ANOVA results comparing mean scores for language tests by test type[35]

14.12.4 Step 4: Interpret Your Findings in Relation to the Hypothesis (Translate Math to English)

Once we are satisfied that the assumptions for the repeated measures ANOVA are reasonably met, we can interpret the results.

> We hypothesized that there would be a difference in task scores for Broca's aphasia patients. The overall effect of task is statistically significant, with a large effect, $F(2, 10) = 6.93$, $p = .013$, $h^2 = .58$. We reject the null hypothesis of no task differences. The sample provides evidence that word-recognition scores differ across the three tasks.

Because the repeated measures ANOVA is an omnibus test, this result tells us that at least two task scores differ from each other. It does not tell us which specific tasks differ. We need follow-up comparisons to answer that question.

Pairwise Comparisons

Measure: MEASURE_1

(I) test	(J) test	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
speech	conceptual	1.000	.683	.203	-.756	2.756
	syntax	2.833*	.910	.026	.495	5.172
conceptual	speech	-1.000	.683	.203	-2.756	.756
	syntax	1.833*	.703	.048	.026	3.641
syntax	speech	-2.833*	.910	.026	-5.172	-.495
	conceptual	-1.833*	.703	.048	-3.641	-.026

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Figure 14.25: **Figure 14.23** Post hoc comparisons using Least Significant Differences (LSD) by language test type[36]

14.12.5 Which Post Hoc Test Do We Choose?

Because the repeated measures ANOVA is an omnibus test, this result tells us that at least two task scores differ from each other. It does not tell us which specific tasks differ. We need follow-up comparisons to answer that question. For a repeated measures ANOVA, select either the Least Significant Difference (LSD) post hoc (if you want a more flexible option) or the Bonferroni post hoc (if you want a more rigorous or picky option). For this example, we have chosen the LSD option.

Because the overall repeated measures ANOVA is statistically significant, we can interpret the LSD post hoc comparisons. The post hoc table shows that scores differed significantly between the speech and syntax tasks, $p = .026$, and the syntax and conceptual tasks, $p = .048$. The conceptual task did not differ significantly from the speech task. The estimated marginal means table helps us interpret the direction of the difference. Participants recognized more words in the speech task and conceptual task than they did in the syntax task.

14.12.5.1 Write Up the Results in American Psychological Association (APA) Style

An APA-style results section should describe the research question, summarize the relevant descriptive statistics, report the inferential test and effect size, and interpret the result. When a

repeated measures ANOVA is statistically significant, the write-up should also report appropriate follow-up comparisons from the post hoc analyses.

> We hypothesized that there would be a difference in task scores for Broca's aphasia patients. The overall effect of task is statistically significant, with a large effect, $F(2, 10) = 6.93$, $p = .013$, $h^2 = .58$. We reject the null hypothesis of no task differences. The sample provides evidence that word-recognition scores differ across the three tasks.
>
> LSD post hoc comparisons indicated that participants recognized significantly fewer words in the syntax task ($M = 4.33$, $SD = 1.63$) than the speech task ($M = 7.17$, $SD = 1.47$), $p = .03$, or the conceptual task conceptual task ($M = 6.17$, $SD = 1.47$), $p = .05$. The speech and conceptual tasks did not differ significantly from each other.

This is not the only correct way to write the result. The key is to report the overall repeated measures ANOVA, the relevant descriptive statistics, the effect size, the follow-up comparisons, and the interpretation.

14.12.5.2 Visualize the Results

For a repeated measures ANOVA, the graph should help readers understand how scores differed across time or condition for the related measurements. Because the same participants completed all three tasks, the most informative graph often shows the repeated nature of the data.

A line graph (plot) can be useful when the repeated measurements have a meaningful order, such as time points. For unordered tasks or conditions like our current example, a bar graph of estimated marginal means is more appropriate to help readers compare the task scores. In the current example, the bar graph demonstrates that word-recognition scores were highest for the speech task and lowest for the syntax task.

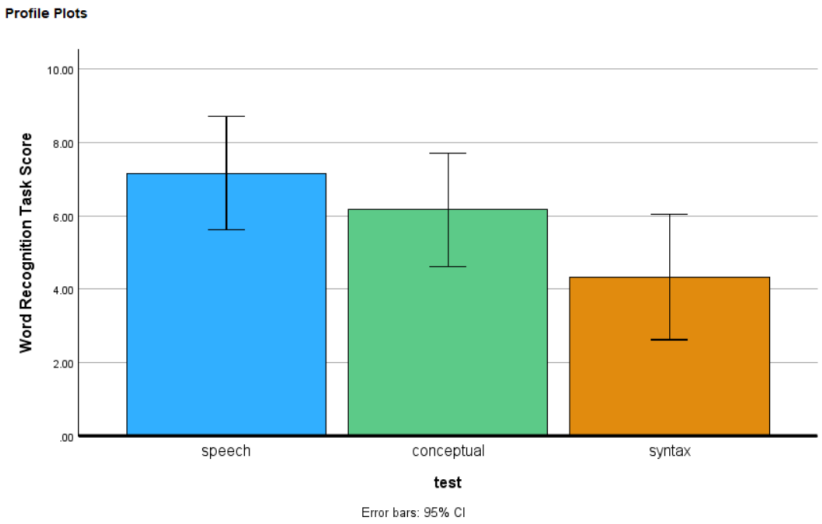


Figure 14.26: **Figure 14.24** Bar chart showing word recognition scores by test type with 95% confidence intervals[37]

14.13 Other ANOVA Designs

We have only just scratched the surface on ANOVA in this chapter. There are many other variations available for the one-way ANOVA presented here. There are also other types of ANOVAs that you are likely to encounter. One such example is called a factorial ANOVA. A ***factorial ANOVA*** uses multiple grouping (independent) variables, not just one, to look for group mean differences. Just as there is no limit to the number of groups in a one-way ANOVA, there is no limit to the number of grouping variables in a factorial ANOVA. However, it becomes very difficult to find and interpret significant results with loads of factors, so usually they are limited to two or three grouping variables with only a small number of groups in each. Factorial designs will be discussed more in a future chapter.

14.14 Summary

ANOVA is an incredibly useful tool for looking at mean differences in scores when there are three or more groups. Different forms of ANOVA are used when the individuals are distinct (independent, ANOVA between) and when the people are connected or repeat (ANOVA within). Other kinds of ANOVA can be used when there is more than one predictor variable.

14.15 Additional Resources

For tips and demonstrations on setting up and running statistical analyses in SPSS, please see Virginia Wickline: SPSS Statistics Helps (publishing dates vary).

14.16 Additional Video Helps

- ANOVA: Crash Course Statistics #33
- ANOVA Part 2: Dealing with Intersectional Groups: Crash Course Statistics #34

14.17 Text Attributions

Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). *Introduction to statistics in the psychological sciences*. Pressbooks. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Illowsky, B., Dean, S., Birmajer, D., Blount, B., Boyd, S., Einsohn, M., Foreman, N., Helmreich, J., Keynon, L., Lee, S., & Taub, J. (2023, updated 2026). *Introductory Statistics 2e*. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Linnell, D. (n.d.). Statistics with jamovi. https://danalinnell.github.io/statistics-with-jamovi/? Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Navarro, D. J., & Foxcroft, D. R. (2025). *Learning statistics with jamovi: A tutorial for beginners in statistical analysis*. Open Book Publishers. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Oja, M. (2021). *Taft College PSYC 2200: Elementary statistics for behavioral and social sciences*. Licensed under CC BY-SA 4.0, except where otherwise noted. Modified by current authors.

14.18 Media Attributions

[1] ("Job Test Scores" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[2] ("Job Test Scores by Degree" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[3] ("Job Test Scores by Group" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[4-6] Tables 14.1-14.3. (CC BY-SA 4.0 by Cote et al., 2021)

[7] ("Job Test Scores Group Means" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

[8-14] Tables 14.4-14.10. (CC BY-SA 4.0 by Cote et al., 2021)

[15] Wickline, V.B. (2026). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[16-18] Figures 14.6-14.8 Linnell, D. (n.d.). *Statistics with jamovi.* Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[19] Wickline, V.B. (2026). Annotated image from Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). Introduction to statistics in the psychological sciences. Pressbooks. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[20] (CC BY-SA 4.0 by Navarro & Foxcroft, 2025)

[21-28] Figures 14.10-14.17. Wickline, V.B. (2026). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[29] Linnell, D. (n.d.). *Statistics with jamovi.* Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[30] Wickline, V.B. (2026). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[31] Linnell, D. (n.d.). *Statistics with jamovi.* Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[32] Wickline, V.B. (2026). Annotated image from Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). Introduction to statistics in the psychological sciences. Pressbooks. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[33] Linnell, D. (n.d.). *Statistics with jamovi.* Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[34-37] Figures 14.21 – 14.24. Wickline, V.B. (2026). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

14.19 References (including Video Attributions)

Cohen J. (1988). *Statistical power analysis for the behavioral sciences.* Routledge Academic.

Geschwind, N. (1972). Language and the brain. *Scientific American*, *226(4)*, 76–83. https://doi.org/10.1038/scientificamerican0472-76

Storage, D. (2019, July 10). *What are Analyses of Variance? One-way and factorial ANOVAs.* https://www.youtube.com/watch?v=q6E0LoMFiUs

Part V

Module 5: Research Designs

15 Chapter 15: Non-Experimental Research

What do the following classic studies have in common?

- Stanley Milgram found that about two thirds of his research participants were willing to administer (what they thought to be) dangerous shocks to another person just because they were told to by an authority figure (Milgram, 1963).
- Elizabeth Loftus and Jacqueline Pickrell showed that it is relatively easy to "implant" false memories in people by repeatedly asking them about childhood events that did not actually happen to them (Loftus & Pickrell, 1995).
- John Cacioppo and Richard Petty evaluated the validity of their Need for Cognition Scale—a measure of the extent to which people like and value thinking—by comparing the scores of university professors with those of factory workers (Cacioppo & Petty, 1982).
- David Rosenhan found that confederates who went to psychiatric hospitals claiming to have heard voices saying things like "empty" and "thud" were diagnosed with schizophrenia by the hospital staff and kept there even though they behaved normally in all other ways (Rosenhan, 1973).

The answer for purposes of this chapter is that they are not experiments. In this chapter, we look more closely at non-experimental research. We begin with a general definition of non-experimental research, along with a discussion of when and why non-experimental research is more appropriate than experimental research. We then look separately at two important types of non-experimental research: correlational research and observational research.

15.1 What Is Non-Experimental Research?

Non-experimental research is research that lacks the manipulation of an independent variable. Rather than manipulating an independent variable, researchers conducting non-experimental research simply measure variables as they naturally occur (in the lab or real world).

Most researchers in psychology consider the distinction between experimental and non-experimental research to be an extremely important one. This is because although experimental research can provide strong evidence that changes in an independent variable cause differences in a dependent variable, non-experimental research generally cannot. However, this inability to

draw causal conclusions does not mean that non-experimental research is less important than experimental research. It is simply used in cases where experimental research is not able to be carried out.

15.2 When to Use Non-Experimental Research

Experimental research is appropriate when the researcher has a specific research question or hypothesis about a causal relationship between two variables—and it is possible, feasible, and ethical to manipulate the independent variable. It stands to reason, therefore, that non-experimental research is appropriate—even necessary—when these conditions are not met. There are many times in which non-experimental research is preferred, including when:

- The research question or hypothesis relates to a single variable rather than a statistical relationship between two variables (e.g., how accurate are people's first impressions?)
- The research question pertains to a non-causal statistical relationship between variables (e.g., is there a correlation between verbal intelligence and mathematical intelligence?)
- The research question is about a causal relationship, but the independent variable cannot be manipulated or participants cannot be randomly assigned to conditions or orders of conditions for practical or ethical reasons (e.g., does damage to a person's hippocampus impair the formation of long-term memory traces?)
- The research question is broad and exploratory, or is about what it is like to have a particular experience (e.g., what is it like to be a working mother diagnosed with depression?)

The choice between the experimental and non-experimental approaches is generally dictated by the nature of the research question. The three goals of science are to describe, to predict, and to explain. If the goal is to explain and the research question pertains to causal relationships, then the experimental approach is typically preferred. If the goal is to describe or to predict, a non-experimental approach is appropriate. But the two approaches can also be used to address the same research question in complementary ways. For example, in Milgram's original (non-experimental) obedience study, he was primarily interested in one variable—the extent to which participants obeyed the researcher when he told them to shock the confederate—and he observed all participants performing the same task under the same conditions. However, Milgram subsequently conducted experiments to explore the factors that affect obedience. He manipulated several independent variables, such as the distance between the experimenter and the participant, the participant and the confederate, and the location of the study (Milgram, 1974).

15.3 Types of Non-Experimental Research

Non-experimental research falls into two broad categories: correlational research and observational research.

The most common type of non-experimental research conducted in psychology is correlational research. Correlational research is considered non-experimental because it focuses on the statistical relationship between two variables but does not include the manipulation of an independent variable. More specifically, in ***correlational research***, the researcher measures two variables with little or no attempt to control extraneous variables and then assesses the relationship between them. As an example, a researcher interested in the relationship between self-esteem and school achievement could collect data on students' self-esteem and their GPAs to see if the two variables are statistically related.

Observational research is non-experimental because it focuses on making observations of behavior in a natural or laboratory setting without manipulating anything. Milgram's original obedience study was non-experimental in this way. He was primarily interested in the extent to which participants obeyed the researcher when he told them to shock the confederate and he observed all participants performing the same task under the same conditions. The study by Loftus and Pickrell is also a good example of observational research. The variable was whether participants "remembered" having experienced mildly traumatic childhood events (e.g., getting lost in a shopping mall) that they had not actually experienced but that the researchers asked them about repeatedly. In this particular study, nearly a third of the participants "remembered" at least one event. (As with Milgram's original study, this study inspired several later experiments on the factors that affect false memories).

15.3.1 Cross-Sectional, Longitudinal, and Cross-Sequential Studies

When psychologists wish to study change over time (for example, when developmental psychologists wish to study aging) they usually take one of three non-experimental approaches: cross-sectional, longitudinal, or cross-sequential. ***Cross-sectional studies*** involve comparing two or more pre-existing groups of people (e.g., children at different stages of development). What makes this approach non-experimental is that there is no manipulation of an independent variable and no random assignment of participants to groups. Using this design, developmental psychologists compare groups of people of different ages (e.g., young adults spanning from 18–25 years of age versus older adults spanning 60–75 years of age) on various dependent variables (e.g., memory, depression, life satisfaction). Of course, the primary limitation of using this design to study the effects of aging is that differences between the groups other than age may account for differences in the dependent variable. For instance, differences between the groups may reflect the generation that people come from (a ***cohort effect***) rather than a direct effect of age. For this reason, ***longitudinal studies***, in which one group of people is followed over time as they age, offer a superior means of studying the effects of aging. However, longitudinal studies are by definition more time-consuming and so require a much greater

investment on the part of the researcher and the participants. A third approach, known as ***cross-sequential studies***, combines elements of both cross-sectional and longitudinal studies. Rather than measuring differences between people in different age groups or following the same people over a long period of time, researchers adopting this approach choose a smaller period of time during which they follow people in different age groups. For example, they might measure changes over a 10-year period among participants who at the start of the study fall into the following age groups: 20 years old, 30 years old, 40 years old, 50 years old, and 60 years old. This design is advantageous because the researcher reaps the immediate benefits of being able to compare the age groups after the first assessment. Further, by following the different age groups over time they can subsequently determine whether the original differences they found across the age groups are due to true age effects or cohort effects.

The research approaches described here are all quantitative, referring to the fact that the data consist of numbers that are analyzed using statistical techniques. However, many observational research studies are more qualitative in nature. Qualitative data are usually nonnumerical and are typically analyzed using interpretive techniques. However, researchers can systematically code qualitative material into categories or frequencies and analyze those coded data statistically, particularly in mixed-methods research. Rosenhan's observational study of the experience of people in psychiatric wards was primarily qualitative. The data were the notes taken by the "pseudopatients"—the people pretending to have heard voices—along with their hospital records. Rosenhan's analysis consisted mainly of a written description of the experiences of the pseudopatients, supported by several concrete examples. To illustrate the hospital staff's tendency to "depersonalize" their patients, he noted, "Upon being admitted, I and other pseudopatients took the initial physical examinations in a semi-public room, where staff members went about their own business as if we were not there" (Rosenhan, 1973, p. 256). Qualitative data have their own set of analysis methods, and the appropriate method depends on the research question. For example, thematic analysis identifies and interprets recurring patterns of meaning across the data. Conversation analysis examines how people produce and organize talk in interactions, including features such as turn-taking, pauses, and word choice.

15.4 Internal Validity

Internal validity is the extent to which the design of a study supports the conclusion that changes in the independent variable caused any observed differences in the dependent variable. Figure 15.1 shows how experimental, quasi-experimental, and non-experimental (correlational) research vary in terms of internal validity. Experimental research tends to be highest in internal validity because the use of manipulation (of the independent variable) and control (of extraneous variables) help to rule out alternative explanations for the observed relationships. If the average score on the dependent variable in an experiment differs across conditions, it is quite likely that the independent variable is responsible for that difference. Non-experimental (correlational) research is lowest in internal validity because these designs fail to use manipulation or control. Quasi-experimental research falls in the middle because it contains some, but not all, of

the features of a true experiment. For instance, a quasi-experiment may assign participants to groups based on preexisting characteristics or naturally occurring circumstances rather than through random assignment. Imagine, for example, that a researcher finds two similar schools, starts an anti-bullying program in one, and then finds fewer bullying incidents in that "treatment school" than in the "control school." While a comparison is being made with a control condition, the inability to randomly assign children to schools could still mean that students in the treatment school differed from students in the control school in some other way that could explain the difference in bullying (e.g., there may be a selection effect).

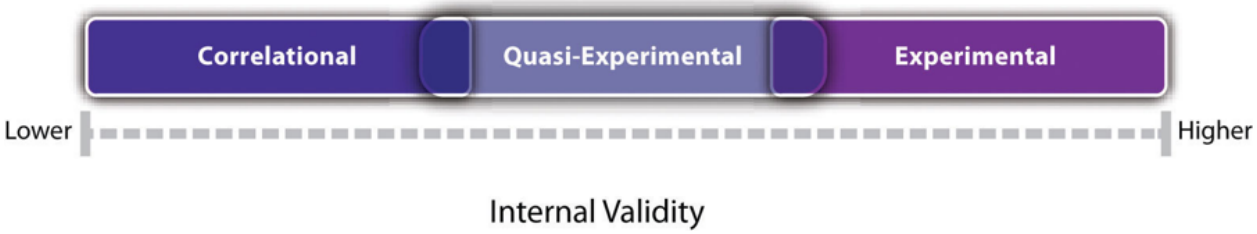


Figure 15.1: Horizontal continuum of internal validity from lower to higher. Correlational studies occupy the lower end, quasi-experimental studies the middle, and experimental studies the higher end, with overlapping ranges.

Figure 15.1 *Internal Validity of Correlational, Quasi-Experimental, and Experimental Studies*[1]

Notice also in Figure 15.1 that there is some overlap in the internal validity of experiments, quasi-experiments, and correlational (non-experimental) studies. For example, a poorly designed experiment that includes many confounding variables can be lower in internal validity than a well-designed quasi-experiment with no obvious confounding variables. Internal validity is also only one of several validities that one might consider.

15.5 Media Attributions

[1] Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

15.6 Text Attributions

Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-

15.7 References

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals, 25*(12), 720–725. https://doi.org/10.3928/0048-5713-19951201-07

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*(4), 371–378. https://doi.org/10.1037/h0040525

Milgram, S. (1974). *Obedience to authority: An experimental view*. Harper & Row.

Rosenhan, D. L. (1973). On being sane in insane places. *Science, 179*(4070), 250–258. https://doi.org/10.1126/science.179.4070.250

16 Chapter 16: Qualitative Research

16.1 What Is Qualitative Research?

This course is primarily about ***quantitative research***, in part because most studies conducted in psychology are quantitative in nature. Quantitative researchers typically start with a focused research question or hypothesis, collect a small amount of numerical data from a large number of individuals, describe the resulting data using statistical techniques, and draw general conclusions about some large population. Although this method is by far the most common approach to conducting empirical research in psychology, there is an important alternative called ***qualitative research***. Qualitative research originated in the disciplines of anthropology and sociology but is now used to study psychological topics as well. Qualitative researchers generally begin with a less focused research question, collect large amounts of relatively "unfiltered" data from a relatively small number of individuals, and describe their data using nonstatistical techniques, such as grounded theory, thematic analysis, critical discourse analysis, or interpretative phenomenological analysis. They are usually less concerned with drawing general conclusions about human behavior than with understanding in detail the *experience* of their research participants.

Consider, for example, a study by researcher Per Lindqvist and his colleagues, who wanted to learn how the families of adolescents who died by suicide cope with their loss (Lindqvist et al., 2008). They did not have a specific research question or hypothesis, such as, "What percentage of family members join suicide support groups?" Instead, they wanted to understand the variety of reactions that families had, with a focus on what it is like from *their* perspectives. To address this question, they interviewed the families of 10 adolescents who died by suicide in their homes in rural Sweden. The interviews were relatively unstructured, beginning with a general request for the families to talk about their loved one and ending with an invitation to talk about anything else that they wanted to tell the interviewer. One of the most important themes that emerged from these interviews was that even as life returned to "normal," the families continued to struggle with the question of why their loved one died by suicide. This struggle appeared to be especially difficult for families for whom the death was most unexpected.

16.2 The Purpose of Qualitative Research

Quantitative research is especially useful for answering clearly defined questions, measuring relationships among variables, and identifying patterns that may apply across large groups of people. However, its emphasis on numerical measurement can limit the amount of detail it provides about individuals' experiences and the contexts in which behavior occurs. Qualitative research complements these strengths by offering rich descriptions of particular people, groups, and situations. It can also reveal unexpected ideas and perspectives that researchers may use to develop new research questions and hypotheses. The research of Lindqvist and colleagues, for example, suggests that there may be a general relationship between how unexpected an adolescent's death by suicide is and how consumed the family is with trying to understand why it occurred. This relationship can now be explored using quantitative research. But it is unclear whether this question would have arisen at all without the researchers sitting down with the families and listening to what they themselves wanted to say about their experience. Qualitative research can also provide rich and detailed descriptions of human behavior in the real-world contexts in which it occurs. Among qualitative researchers, this depth is often referred to as "thick description" (Geertz, 1973). Similarly, qualitative research can convey a sense of what it is actually like to be a member of a particular group or in a particular situation—what qualitative researchers often refer to as the "lived experience" of the research participants. Lindqvist and colleagues, for example, describe how all the families spontaneously offered to show the interviewer their loved one's bedroom or the place where the death occurred—revealing the importance of these physical locations to the families. It seems unlikely that a quantitative study would have discovered this detail.

Table 16.1 *Some Contrasts Between Qualitative and Quantitative Research*[1]

Table 6.3 Some contrasts between qualitative and quantitative research

Qualitative	Quantitative
1. In-depth information about relatively few people	1. Less depth information with larger samples
2. Conclusions are based on interpretations drawn by the investigator	2. Conclusions are based on statistical analyses
3. Global and exploratory	3. Specific and focused

Figure 16.1: Table comparing qualitative and quantitative research. Qualitative research emphasizes in-depth information from relatively few participants, conclusions based on researcher interpretation, and exploratory approaches. Quantitative research emphasizes less detailed information from larger samples, conclusions based on statistical analyses, and focused, specific research questions.

16.3 Data Collection and Analysis in Qualitative Research

Data collection approaches in qualitative research are quite varied and can involve naturalistic observation, participant observation, archival data, artwork, and many other things. But one of the most common approaches, especially for psychological research, is to conduct ***interviews***. Interviews in qualitative research can be unstructured—consisting of a small number of general questions or prompts that allow participants to talk about what is of interest to them—or structured, where there is a strict script that the interviewer does not deviate from. Most interviews are in between the two and are called semi-structured interviews, where the researcher has a few consistent questions and can follow up by asking more detailed questions about the topics that come up. Such interviews can be lengthy and detailed, but they are usually conducted with a relatively small sample. The unstructured interview was the approach used by Lindqvist and colleagues in their research on the families of people who died by suicide because the researchers were aware that how much was disclosed about such a sensitive topic should be led by the families, not by the researchers.

Another approach used in qualitative research involves small groups of people who participate together in interviews focused on a particular topic or issue, known as ***focus groups***. The interaction among participants in a focus group can sometimes bring out more information than can be learned in a one-on-one interview. The use of focus groups has become a standard technique in business and industry among those who want to understand consumer tastes and preferences. Focus groups are commonly audio- or video-recorded and transcribed to facilitate later analyses. However, we know from social psychology that group dynamics are often at play in any group, including focus groups, and it is useful to be aware of those possibilities. For example, the desire to be liked by others can lead participants to provide inaccurate answers that they believe will be perceived favorably by the other participants. The same may be said for personality characteristics. For example, highly extraverted participants can sometimes dominate discussions within focus groups.

16.4 Data Analysis in Qualitative Research

Although quantitative and qualitative research generally differ along several important dimensions (e.g., the specificity of the research question, the type of data collected), it is the method of data analysis that distinguishes them more clearly than anything else. To illustrate this idea, imagine a team of researchers that conducts a series of unstructured interviews with people recovering from alcohol use disorder to learn about the role of their religious faith in their recovery. Although this project sounds like qualitative research, imagine further that once they collect the data, they code the data in terms of how often each participant mentions God (or a "higher power"), and they then use descriptive and inferential statistics to find out whether those who mention God more often are more successful in abstaining from alcohol. Now it sounds like quantitative research. In other words, the quantitative-qualitative distinction

depends more on what researchers do with the data they have collected than with why or how they collected the data.

But what does qualitative data analysis look like? Just as there are many ways to collect data in qualitative research, there are many ways to analyze data. Here we focus on one general approach called ***grounded theory*** (Glaser & Strauss, 1967). This approach was developed within the field of sociology in the 1960s and has gradually gained popularity in psychology. In quantitative research, it is typical for the researcher to start with a theory, derive a hypothesis from that theory, and then collect data to test that specific hypothesis. In qualitative research using grounded theory, researchers start with the data and develop a theory or an interpretation that is "grounded in" those data. They do this analysis in stages. First, they identify ideas that are repeated throughout the data. Then they organize these ideas into a smaller number of broader themes. Finally, they write a ***theoretical narrative***—an interpretation of the data in terms of the themes that they have identified. This theoretical narrative focuses on the subjective experience of the participants and is usually supported by many direct quotations from the participants themselves.

As an example, consider a study by researchers Laura Abrams and Laura Curran, who used the grounded theory approach to study the experience of postpartum depression symptoms among low-income mothers (Abrams & Curran, 2009). Their data were the result of unstructured interviews with 19 participants. Table 16.2 shows the five broad themes the researchers identified and the more specific repeating ideas that made up each of those themes. In their research report, they provide numerous quotations from their participants, such as this one from a participant identified as "Destiny":

> Well, just recently my apartment was broken into and the fact that his Medicaid for some reason was cancelled so a lot of things was happening within the last two weeks all at one time. So that in itself I don't want to say almost drove me mad but it put me in a funk….Like I really was depressed. (p. 357)

Their theoretical narrative focused on the participants' experience of their symptoms, not as an abstract "affective disorder" but as closely tied to the daily struggle of raising children alone under often difficult circumstances.

Table 16.2 *Themes and Repeating Ideas in a Study of Postpartum Depression Among Low-Income Mothers*[2]

Theme	Repeating ideas
Ambivalence	"I wasn't prepared for this baby," "I didn't want to have any more children."
Caregiving overload	"Please stop crying," "I need a break," "I can't do this anymore."
Juggling	"No time to breathe," "Everyone depends on me," "Navigating the maze."
Mothering alone	"I really don't have any help," "My baby has no father."
Real-life worry	"I don't have any money," "Will my baby be OK?" "It's not safe here."

Table 6.4 Themes and Repeating Ideas in a Study of Postpartum Depression Among Low-Income Mothers. Based on Research by Abrams and Curran (2009).

Figure 16.2: Table showing themes and recurring participant statements from a qualitative study of postpartum depression among low-income mothers. Themes include ambivalence, caregiving overload, juggling, mothering alone, and real-life worry.

16.5 The Quantitative-Qualitative "Debate"

Given their differences, it may come as no surprise that quantitative and qualitative research in psychology and related fields do not coexist in complete harmony. Some quantitative researchers criticize qualitative methods on the grounds that they lack objectivity, are difficult to evaluate in terms of reliability and validity, and do not allow generalization to people or situations other than those actually studied. At the same time, some qualitative researchers criticize quantitative methods on the grounds that they overlook the richness of human behavior and experience and instead answer simple questions about easily quantifiable variables.

In general, however, qualitative researchers are well aware of the issues of objectivity, reliability, validity, and generalizability. In fact, they have developed a number of frameworks for addressing these issues (which are beyond the scope of our discussion). And in general, quantitative researchers are well aware of the issue of oversimplification. They do not believe that all human behavior and experience can be adequately described in terms of a small number of variables and the statistical relationships among them. Instead, they use simplification as a strategy for uncovering general principles of human behavior.

Many researchers from both the quantitative and qualitative camps now agree that the two approaches can and should be combined into what has come to be called ***mixed-methods research*** (Todd et al., 2004). (In fact, the studies by Lindqvist and colleagues and by Abrams and Curran both combined quantitative and qualitative approaches.) One approach to combining quantitative and qualitative research is to use qualitative research for hypothesis generation and quantitative research for hypothesis testing. While a qualitative study might

suggest that families whose loved one dies unexpectedly by suicide have more difficulty resolving the question of why, a well-designed quantitative study could test a hypothesis by measuring these specific variables in a large sample. A second approach to combining quantitative and qualitative research is referred to as ***triangulation***. The idea is to use both quantitative and qualitative methods simultaneously to study the same general questions and to compare the results. If the results of the quantitative and qualitative methods converge on the same general conclusion, they reinforce and enrich each other. If the results diverge, then they suggest an interesting new question: Why do the results diverge and how can they be reconciled?

Using qualitative research can often help clarify quantitative results via triangulation. Trenor et al. (2008) investigated the experience of female engineering students at a university. In the first phase, female engineering students were asked to complete a survey, where they rated a number of their perceptions, including their sense of belonging. Their results were compared across the student ethnicities, and statistically, the various ethnic groups showed no differences in their ratings of their sense of belonging. One might look at that result and conclude that ethnicity does not have anything to do with one's sense of belonging. However, in the second phase, the authors also conducted interviews with the students, and in those interviews, many minority students reported how the diversity of cultures at the university enhanced their sense of belonging. Without the qualitative component, we might have drawn the wrong conclusion about the quantitative results. This example shows how qualitative and quantitative research work together to help us understand human behavior.

16.6 Media Attributions

[1] Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

[2] Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

16.7 Text Attributions

Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

16.8 References

Abrams, L. S., & Curran, L. (2009). "And you're telling me not to stress?" A grounded theory study of postpartum depression symptoms among low-income mothers. *Psychology of Women Quarterly, 33*(3), 351–362. https://doi.org/10.1111/j.1471-6402.2009.01506.x

Geertz, C. (1973). *The interpretation of cultures*. Basic Books.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine.

Lindqvist, P., Johansson, L., & Karlsson, U. (2008). In the aftermath of teenage suicide: A qualitative study of the psychosocial consequences for the surviving family members. *BMC Psychiatry, 8*, Article 26. https://doi.org/10.1186/1471-244X-8-26

Todd, Z., Nerlich, B., McKeown, S., & Clarke, D. D. (Eds.). (2004). *Mixing methods in psychology: The integration of qualitative and quantitative methods in theory and practice*. Psychology Press.

Trenor, J. M., Yu, S. L., Waight, C. L., Zerda, K. S., & Sha, T.-L. (2008). The relations of ethnicity to female engineering students' educational experiences and college and career plans in an ethnically diverse learning environment. *Journal of Engineering Education, 97*(4), 449–465. https://doi.org/10.1002/j.2168-9830.2008.tb00992.x

17 Chapter 17: Observational Research

17.1 What Is Observational Research?

The term ***observational research*** is used to refer to several different types of non-experimental studies in which behavior is systematically observed and recorded. The goal of observational research is to describe a variable or set of variables. More generally, the goal is to obtain a snapshot of specific characteristics of an individual, group, or setting. As described previously, observational research is non-experimental because nothing is manipulated or controlled, and as such we cannot arrive at causal conclusions using this approach. The data that are collected in observational research studies are often qualitative in nature but they may also be quantitative or both (mixed-methods). There are several different types of observational methods that will be described below.

17.2 Naturalistic Observation

Naturalistic observation is an observational method that involves observing people's behavior in the environment in which it typically occurs. Thus, naturalistic observation is a type of field research (as opposed to a type of laboratory research). Jane Goodall's famous research on chimpanzees is a classic example of naturalistic observation. Dr. Goodall spent three decades observing chimpanzees in their natural environment in East Africa. She examined such things as chimpanzees' social structure, mating patterns, gender roles, family structure, and care of offspring by observing them in the wild. However, naturalistic observation could more simply involve observing shoppers in a grocery store, children on a school playground, or psychiatric inpatients in their wards. Researchers engaged in naturalistic observation usually make their observations as unobtrusively as possible so that participants are not aware that they are being studied. Such an approach is called ***disguised naturalistic observation***. Ethically, this method is considered to be acceptable if the participants remain anonymous and the behavior occurs in a public setting where people would not normally have an expectation of privacy. Grocery shoppers putting items into their shopping carts, for example, are engaged in public behavior that is easily observable by store employees and other shoppers. For this reason, most researchers would consider it ethically acceptable to observe them for a study.

In cases where it is not ethical or practical to conduct disguised naturalistic observation, researchers can conduct ***undisguised naturalistic observation*** where the participants are

made aware of the researcher's presence and monitoring of their behavior. However, one concern with undisguised naturalistic observation is reactivity. ***Reactivity*** refers to when a measure changes participants' behavior. In the case of undisguised naturalistic observation, the concern with reactivity is that when people know they are being observed and studied, they may act differently than they normally would. This type of reactivity is known as the ***Hawthorne effect***. For instance, you may act differently in a bar if you know that someone is observing you and recording your behavior, which could threaten the study's validity. So disguised observation is less reactive and therefore can have higher validity because people are not aware that their behaviors are being observed and recorded. However, we now know that people often become used to being observed and with time they begin to behave naturally in the researcher's presence. In other words, over time people habituate to being observed. Reality shows such as Big Brother or Survivor illustrate habituation: participants may monitor their behavior at first but often begin acting more naturally after continued observation.

17.3 Participant Observation

Another approach to data collection in observational research is participant observation. In ***participant observation***, researchers become active participants in the group or situation they are studying. Participant observation is very similar to naturalistic observation in that it involves observing people's behavior in the environment in which it typically occurs. As with naturalistic observation, the data that are collected can include interviews (usually unstructured), notes based on their observations and interactions, documents, photographs, and other artifacts. The only difference between naturalistic observation and participant observation is that researchers engaged in participant observation become active members of the group or situations they are studying. The basic rationale for participant observation is that there may be important information that is only accessible to, or can be interpreted only by, someone who is an active participant in the group or situation. Like naturalistic observation, participant observation can be either disguised or undisguised. In ***disguised participant observation***, the researchers pretend to be members of the social group they are observing and conceal their true identity as researchers. In ***undisguised participant observation***, the researchers become a part of the group they are studying and they disclose their true identity as researchers to the group under investigation.

There are important ethical issues to consider with disguised participant observation. First, no informed consent can be obtained, and second, deception is being used. The researcher is deceiving the participants by intentionally withholding information about their motivations for being a part of the social group they are studying. But sometimes disguised participation is the only way to access a protective group (like a cult). Further, disguised participant observation is less prone to reactivity than undisguised participant observation.

One of the primary benefits of participant observation is that the researchers are in a much better position to understand the viewpoint and experiences of the people they are studying

when they are a part of the social group. The primary limitation with this approach is that the mere presence of the observer could affect the behavior of the people being observed. While this is also a concern with naturalistic observation, additional concerns arise when researchers become active members of the social group they are studying because they may change the social dynamics and/or influence the behavior of the people they are studying. Similarly, if the researcher acts as a participant observer there can be concerns with biases resulting from developing relationships with the participants. Consequently, the researcher may become less objective resulting in more experimenter bias.

17.4 Structured Observation

Another observational method is ***structured observation***. Here the investigator makes careful observations of one or more specific behaviors in a particular setting that is more structured than the settings used in naturalistic or participant observation. Often the setting in which the observations are made is not the natural setting. Instead, the researcher may observe people in the laboratory environment. Alternatively, the researcher may observe people in a natural setting (like a classroom setting) that they have structured in some way, for instance by introducing a specific task for participants to complete or by introducing a specific social situation or manipulation.

Structured observation resembles naturalistic and participant observation because all three involve the systematic observation of behavior. Unlike the other two approaches, however, structured observation may elicit target behaviors by arranging a task, situation, or manipulation. Its emphasis is usually on gathering quantitative rather than qualitative data. Researchers using this approach are interested in a limited set of behaviors, which allows them to quantify what they observe. In other words, structured observation is less global than naturalistic or participant observation because the researcher focuses on a small number of specific behaviors. Therefore, rather than recording everything that happens, the researcher records only the behaviors identified in advance.

17.5 Quantifying and Sampling Observed Behavior

Before data collection begins, researchers create a coding scheme that translates behavior into observable categories. Each category needs an ***operational definition*** that states exactly what counts as an instance, when an instance begins and ends, and how ambiguous cases will be handled. After defining the behavior, researchers choose a method for quantifying it and a method for sampling from the larger stream of people, events, and time.

The ***frequency method*** records how often a defined behavior occurs. A researcher might count the number of aggressive comments during a discussion or the number of times a child shares a toy. Raw frequencies are most meaningful when observation periods and opportunities

are comparable. When they differ, researchers often report a rate, such as responses per minute or incidents per class period.

The ***duration method*** records how long a behavior lasts. Observers may measure the total duration, the average duration of an episode, or the proportion of the observation period occupied by the behavior. Duration is useful for behaviors that extend over time, such as sustained attention, conversation, or walking a fixed distance. A precise rule is needed for deciding when the behavior starts, stops, and resumes after an interruption.

The ***interval method*** divides an observation period into equal intervals and codes each interval. With whole-interval recording, the behavior is coded only if it occurs throughout the interval; this tends to underestimate behavior. With partial-interval recording, it is coded if it occurs at any point; this tends to overestimate behavior. In momentary time sampling, the observer records whether the behavior is occurring at a designated instant, often the end of each interval. Shorter intervals usually provide a more precise picture but require more observer effort.

Sampling determines which portions of the behavioral stream are actually observed. The sampling plan should be chosen in advance and should represent the settings, people, and occasions to which the researcher hopes to generalize. Sampling and measurement are related but distinct: for example, time sampling selects when observations occur, whereas interval recording specifies how behavior is coded within an observation period.

In ***time sampling***, researchers observe during selected periods rather than continuously. Observation windows can be scheduled systematically (for example, ten minutes every hour), selected randomly, or stratified so that mornings, afternoons, weekdays, and weekends are represented. Time sampling makes large studies manageable, but it can miss behavior that occurs outside the selected windows or follows predictable cycles.

In ***individual sampling***, sometimes called focal sampling, one person is selected for close observation for a specified period before the observer moves to another person. Researchers need a rule for selecting and rotating among individuals so that highly visible or active people are not overrepresented. The method is especially useful when many people are present and continuous coding of everyone is impossible.

In ***event sampling***, observers record every occurrence of a particular, clearly defined event during the observation period, often along with its antecedents and consequences. Event sampling is efficient for uncommon or especially important behaviors, such as conflicts or helping episodes. It does not, by itself, show how common the event is relative to opportunities unless the amount of observation time and the number of opportunities are also recorded.

Researchers Robert Levine and Ara Norenzayan used structured observation to study differences in the "pace of life" across countries (Levine & Norenzayan, 1999). One of their measures involved observing pedestrians in a large city to see how long it took them to walk 60 feet. They found that people in some countries walked reliably faster than people in other countries. For example, people in Canada and Sweden covered 60 feet in just under 13 seconds on average, while people in Brazil and Romania took close to 17 seconds. When structured observation

takes place in the complex and even chaotic "real world," the questions of when, where, and under what conditions the observations will be made, and who exactly will be observed are important to consider. Levine and Norenzayan described their sampling process as follows:

> Levine and Norenzayan measured walking speed over 60 feet at a minimum of two central downtown locations in each city. Researchers collected observations during business hours on clear summer days, using flat, unobstructed, broad, relatively uncrowded sidewalks. To reduce the influence of social interaction, they timed only adults walking alone and excluded children, window-shoppers, and pedestrians whose visible physical impairments might affect walking speed. In most cities, the sample included 35 men and 35 women (Levine & Norenzayan, 1999, p. 186).

Precise specification of the sampling process in this way makes data collection manageable for the observers, and it also provides some control over important extraneous variables. For example, by making their observations on clear summer days in all countries, Levine and Norenzayan controlled for effects of the weather on people's walking speeds. In Levine and Norenzayan's study, measurement was relatively straightforward. They simply measured out a 60-foot distance along a city sidewalk and then used a stopwatch to time participants as they walked over that distance.

As another example, researchers Robert Kraut and Robert Johnston wanted to study bowlers' reactions to their shots, both when they were facing the pins and then when they turned toward their companions (Kraut & Johnston, 1979). But what "reactions" should they observe? Based on previous research and their own pilot testing, Kraut and Johnston created a list of reactions that included "closed smile," "open smile," "laugh," "neutral face," "look down," "look away," and "face cover" (covering one's face with one's hands). The observers committed this list to memory and then practiced by coding the reactions of bowlers who had been videotaped. During the actual study, the observers spoke into an audio recorder, describing the reactions they observed. Among the most interesting results of this study was that bowlers rarely smiled while they still faced the pins. They were much more likely to smile after they turned toward their companions, suggesting that smiling is not purely an expression of happiness but also a form of social communication.

In yet another example (this one in a laboratory environment), Dov Cohen and his colleagues had observers rate the emotional reactions of participants who had just been deliberately bumped and insulted by a confederate after they dropped off a completed questionnaire at the end of a hallway. The confederate was posing as someone who worked in the same building and who was frustrated by having to close a file drawer twice in order to permit the participants to walk past them (first to drop off the questionnaire at the end of the hallway and once again on their way back to the room where they believed the study they signed up for was taking place). The two observers were positioned at different ends of the hallway so that they could read the participants' body language and hear anything they might say. Interestingly, the researchers hypothesized that participants from the southern United States, which is one of several places in the world that has a "culture of honor," would react with more aggression than participants

from the northern United States, a prediction that was in fact supported by the observational data (Cohen et al., 1996).

When the observations require a judgment on the part of the observers—as in the studies by Kraut and Johnston and Cohen and his colleagues—a process referred to as ***coding*** is typically required. Coding generally requires clearly defining a set of target behaviors. The observers then categorize participants individually in terms of which behavior they have engaged in and the number of times they engaged in each behavior. The observers might even record the duration of each behavior. The target behaviors must be defined in such a way that guides different observers to code them in the same way. This difficulty with coding illustrates the issue of inter-rater reliability. Researchers are expected to demonstrate the inter-rater reliability of their coding procedure by having multiple raters code the same behaviors independently and then showing that the different observers are in close agreement. Kraut and Johnston, for example, video recorded a subset of their participants' reactions and had two observers independently code them. The two observers showed that they agreed on the reactions that were exhibited 97% of the time, indicating good inter-rater reliability.

One of the primary benefits of structured observation is that it is often more efficient than naturalistic or participant observation. Because researchers focus on specific behaviors, they can reduce the time and expense required for data collection. Researchers may also arrange the setting to increase opportunities for the behaviors of interest to occur, rather than waiting for those behaviors to arise naturally. In addition, structured observation gives researchers greater control over the environment. This control involves a tradeoff, however: A highly structured setting may feel less natural and therefore reduce external validity. For example, it may be unclear whether behavior observed in a laboratory will generalize to behavior in everyday settings. Structured observation may also raise concerns about reactivity because participants often know that they are being observed.

17.6 Reliability and Validity Concerns

Reliability concerns the consistency of an observational measure. Reliability depends on a detailed codebook, observer training, practice with examples and nonexamples, and periodic checks for observer drift. Reliability is necessary for a useful measure, but reliable observers can still record the wrong construct, so reliability does not by itself establish validity.

Test–retest reliability, or stability across repeated observation, can be examined by recording the same behavior on multiple occasions under comparable conditions. Similar scores suggest that the measure is stable when the underlying behavior is expected to be stable. Low consistency may reflect an unreliable procedure, real change in behavior, or differences in the settings and opportunities sampled. Researchers therefore repeat observations across enough occasions to distinguish a typical pattern from a single unusual episode. This differs from ***intra-rater reliability***, which assesses whether the same observer codes the same material consistently on separate occasions.

Inter-rater reliability concerns whether independent observers code the same units in the same way. At least two trained observers should code an overlapping subset of participants or recordings without consulting one another. The units, categories, and statistic should be selected before the codes are compared. The 97% agreement reported by Kraut and Johnston is an example of this type of evidence.

Percent agreement is calculated as follows: Percent agreement = (number of agreements ÷ number of coding decisions) × 100%. It is easy to understand and is useful for a quick check, but it does not correct for agreement expected by chance. It can also appear high when one category is very common, such as when both observers usually code that a rare behavior did not occur.

Cohen's kappa is commonly used when two observers assign nominal categories. It corrects observed agreement for the agreement expected by chance: 𝜅 = (P𝜅 − P𝜅) ÷ (1 − P𝜅), where P𝜅 is the observed proportion of agreement and P𝜅 is the chance-expected proportion. Kappa should be interpreted along with the category frequencies and percent agreement because very rare or very common categories can affect its value.

Correlation can assess consistency when observers produce quantitative scores such as frequencies, ratings, or durations. Pearson's r indicates whether observers rank cases similarly, but a high correlation can occur even when one observer consistently gives higher scores than another. When absolute agreement is important, an intraclass correlation coefficient or another agreement statistic is more informative than Pearson's r alone.

Resolving disagreements should occur only after independent codes have been retained and reliability has been calculated. Observers can discuss each discrepancy and reach consensus, or a third trained person can adjudicate using the codebook and original recording. Repeated disagreements may reveal an unclear definition; researchers should refine the codebook, retrain observers, and, when necessary, recode earlier observations. Reports should distinguish the pre-resolution reliability estimate from the final consensus data used in analysis.

Observer bias is a validity concern that occurs when observers' expectations or theoretical commitments influence what they notice or how they interpret ambiguous behavior. Risk can be reduced with specific operational definitions, standardized recording forms, observer training, blinding observers to hypotheses or participant conditions when possible, and periodic reliability checks. Video or audio records allow later review, although recordings can still omit context and must be handled ethically.

Internal validity concerns whether a study supports a causal conclusion. Observational research usually has limited internal validity because researchers do not randomly assign participants or manipulate the presumed cause, so confounding variables and alternative explanations remain. Structured settings, comparison groups, repeated measurement, consistent sampling, and statistical adjustment can reduce some alternatives, but they do not turn an observational association into proof of causation. Researchers should describe associations precisely and avoid causal language unless the design warrants it.

17.7 Case Studies

A ***case study*** is an in-depth examination of an individual. Sometimes case studies are also completed on social units (e.g., a cult) and events (e.g., a natural disaster). Most commonly in psychology, however, case studies provide a detailed description and analysis of an individual. Often the individual has a rare or unusual condition or disorder or has damage to a specific region of the brain.

Like many observational research methods, case studies tend to be primarily qualitative. A case study provides an in-depth examination of an individual, often over an extended period. Depending on the research question, the individual may be observed in a natural setting or studied in a therapist's office or research laboratory. Most case-study reports emphasize detailed descriptions of the person rather than statistical analyses. However, they may also include quantitative data. For example, a researcher might compare an individual's depression score with normative scores or compare the person's scores before and after treatment. Researchers can gather case-study information through interviews, naturalistic or structured observation, psychological testing (e.g., an IQ test), physiological measurements (e.g., brain scans), or a combination of these methods.

The case of H.M. (Henry Molaison) is one of the most influential case studies in psychology. Molaison experienced severe, medically intractable epilepsy. In 1953, surgeon William Scoville removed portions of both medial temporal lobes, including much of the hippocampal region, in an effort to reduce his seizures. The surgery greatly reduced his most severe seizures, while his general intellectual abilities and personality remained largely intact. However, he developed profound anterograde amnesia. H.M. could carry on a conversation and retain short strings of letters, digits, and words—that is, his short-term memory was preserved. Yet he could not form many new long-term declarative memories. This dissociation between short-term and long-term memory provided important evidence that these abilities depend on partly distinct brain systems and highlighted the medial temporal lobes' role in memory consolidation.

Case studies are useful because they provide a level of detailed analysis not found in many other research methods and greater insights may be gained from this more detailed analysis. As a result of the case study, the researcher may gain a sharpened understanding of what might become important to look at more extensively in future more controlled research. Case studies are also often the only way to study rare conditions because it may be impossible to find a large enough sample of individuals with the condition to use quantitative methods. Although at first glance a case study of a rare individual might seem to tell us little about ourselves, they often do provide insights into normal behavior. The case of H.M. provided important insights into the role of the hippocampus in memory consolidation.

Case studies can generate hypotheses and contribute converging or disconfirming evidence for theories, but they cannot establish causation by themselves. Because they typically lack random assignment, experimental manipulation, and comparison conditions, alternative explanations often remain. For example, H.M.'s surgery affected multiple structures in the medial temporal

lobes, so researchers must be cautious about attributing all of his memory changes to a single structure. Case studies also have limited external validity because findings from one individual, particularly someone with an unusual condition, may not generalize to other people. Finally, investigators' theoretical expectations may influence how a case is interpreted or reported. Researchers can reduce this risk by documenting procedures, preserving records, considering competing explanations, and seeking converging evidence from other cases and methods.

17.8 Archival Research

Another approach that is often considered observational research involves analyzing archival data that have already been collected for some other purpose. An example is a study by Brett Pelham and his colleagues on "implicit egotism"—the tendency for people to prefer people, places, and things that are similar to themselves (Pelham et al., 2005). In one study, they examined Social Security records to show that women with the names Virginia, Georgia, Louise, and Florence were especially likely to have moved to the states of Virginia, Georgia, Louisiana, and Florida, respectively.

As with naturalistic observation, measurement can be more or less straightforward when working with archival data. For example, counting the number of people named Virginia who live in various states based on Social Security records is relatively straightforward. But consider a study by Christopher Peterson and his colleagues on the relationship between optimism and health using data that had been collected many years before for a study on adult development (Peterson et al., 1988). In the 1940s, healthy male college students had completed an open-ended questionnaire about difficult wartime experiences. In the late 1980s, Peterson and his colleagues reviewed the men's questionnaire responses to obtain a measure of explanatory style—their habitual ways of explaining bad events that happen to them. More pessimistic people tend to blame themselves and expect long-term negative consequences that affect many aspects of their lives, while more optimistic people tend to blame outside forces and expect limited negative consequences. To obtain a measure of explanatory style for each participant, the researchers used a procedure in which all negative events mentioned in the questionnaire responses and any causal explanations for them were identified and written on index cards. These were given to a separate group of raters who rated each explanation in terms of three separate dimensions of optimism-pessimism. These ratings were then averaged to produce an explanatory style score for each participant. The researchers then assessed the statistical relationship between the men's explanatory style as undergraduate students and archival measures of their health at approximately 60 years of age. The primary result was that the more optimistic the men were as undergraduate students, the healthier they were as older men. Pearson's r was +.25.

This method is an example of ***content analysis***—a family of systematic approaches to measurement using complex archival data. Just as structured observation requires specifying the behaviors of interest and then noting them as they occur, content analysis requires specifying keywords, phrases, or ideas and then finding all occurrences of them in the data. These

occurrences can then be counted, timed (e.g., the amount of time devoted to entertainment topics on the nightly news show), or analyzed in a variety of other ways.

17.9 Text Attributions

Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

17.10 References

Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An "experimental ethnography." *Journal of Personality and Social Psychology, 70*(5), 945–960. https://doi.org/10.1037/0022-3514.70.5.945

Kraut, R. E., & Johnston, R. E. (1979). Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology, 37*(9), 1539–1553. https://doi.org/10.1037/0022-3514.37.9.1539

Levine, R. V., & Norenzayan, A. (1999). The pace of life in 31 countries. *Journal of Cross-Cultural Psychology, 30*(2), 178–205. https://doi.org/10.1177/0022022199030002003

Pelham, B. W., Carvallo, M., & Jones, J. T. (2005). Implicit egotism. *Current Directions in Psychological Science, 14*(2), 106–110. https://doi.org/10.1111/j.0963-7214.2005.00344.x

Peterson, C., Seligman, M. E. P., & Vaillant, G. E. (1988). Pessimistic explanatory style is a risk factor for physical illness: A thirty-five-year longitudinal study. *Journal of Personality and Social Psychology, 55*(1), 23–27. https://doi.org/10.1037/0022-3514.55.1.23

18 Chapter 18: Survey Research

Shortly after the terrorist attacks in New York City and Washington, DC, in September of 2001, researcher Jennifer Lerner and her colleagues conducted an internet-based survey of nearly 2,000 American teens and adults ranging in age from 13 to 88 (Lerner et al., 2003). They asked participants about their reactions to the attacks and for their judgments of various terrorism-related and other risks. Among the results were that the participants tended to overestimate most risks, that females did so more than males, and that there were no differences between teens and adults. The most interesting result, however, had to do with the fact that some participants were "primed" to feel anger by asking them what made them angry about the attacks and by presenting them with a photograph and audio clip intended to evoke anger. Others were primed to feel fear by asking them what made them fearful about the attacks and by presenting them with a photograph and audio clip intended to evoke fear. As the researchers hypothesized, the participants who were primed to feel anger perceived less risk than the participants who had been primed to feel fear—showing how risk perceptions are strongly tied to specific emotions.

The study by Lerner and her colleagues is an example of survey research in psychology—the topic of this chapter. We begin with the purposes, benefits, and limitations of surveys. We then examine survey responding as a psychological process, the construction of effective questions and rating scales, and the reliability and validity of survey measures. Finally, we compare common methods of administering surveys and consider sampling and data-collection decisions.

18.1 What Is Survey Research?

Survey research is a quantitative and qualitative method with two important characteristics. First, the variables of interest are measured using self-reports, usually questionnaires or interviews. Survey researchers ask ***respondents*** to report directly on their thoughts, feelings, and behaviors. Second, survey research gives careful attention to sampling. Probability sampling and strong response rates can support population estimates, whereas a large sample alone cannot correct coverage or nonresponse bias. Beyond these characteristics, surveys can be long or short; administered in person, by telephone, through the mail, or online; and focused on voting intentions, consumer preferences, social attitudes, health, or any other topic about which people can provide meaningful answers. Although survey data are often analyzed statistically, some survey questions are better suited to qualitative analysis.

Most survey research is nonexperimental. It is used to describe single variables (e.g., the percentage of voters who prefer one presidential candidate or another, the prevalence of schizophrenia in the general population, etc.) and also to assess statistical relationships between variables (e.g., the relationship between income and health). But surveys can also be used within experimental research. The study by Lerner and her colleagues is a good example. Their use of self-report measures and a large national sample identifies their work as survey research. But their manipulation of an independent variable (anger vs. fear) to assess its effect on a dependent variable (risk judgments) also identifies their work as experimental.

18.2 Why Researchers Use Surveys

Self-report gives researchers direct access to respondents' attitudes, values, beliefs, intentions, experiences, and behaviors. Surveys can also describe a population by estimating demographic characteristics such as age, education, employment, household composition, and other attributes relevant to the research question. Because sensitive demographic questions may affect participation, researchers should ask only for information that has a clear purpose and should provide inclusive response options.

Surveys can be descriptive or can test hypotheses about relationships between variables. In a nonexperimental survey, a ***predictor variable*** is used to predict or explain variation in a ***criterion variable***. These variables are sometimes labeled the independent variable (IV) and dependent variable (DV), respectively, but predictor and criterion are more precise terms when neither variable is manipulated. A relationship can support a prediction, but a cross-sectional survey alone cannot establish that the predictor caused the criterion.

18.2.1 Benefits of Survey Research

Surveys are flexible, efficient, and often less costly than direct observation or in-person testing. Standardized questions allow researchers to collect comparable information from many people, including geographically dispersed populations, and probability sampling can support population estimates. Surveys are especially valuable for subjective constructs—such as beliefs, emotions, preferences, and private experiences—that may not be directly observable.

18.2.2 Limitations of Survey Research

Survey results depend on what respondents understand, remember, and are willing to disclose. Ambiguous wording, question order, social desirability, acquiescence, imperfect recall, and careless responding can introduce measurement error. Coverage error and nonresponse can also make a sample unrepresentative even when the number of respondents is large. Finally,

associations in nonexperimental survey data may reflect reverse direction, third variables, or selection effects rather than causal relationships.

18.3 History and Uses of Survey Research

Survey research may have its roots in English and American "social surveys" conducted around the turn of the 20th century by researchers and reformers who wanted to document the extent of social problems such as poverty (Converse, 1987). By the 1930s, the US government was conducting surveys to document economic and social conditions in the country. The need to draw conclusions about the entire population helped spur advances in sampling procedures. At about the same time, several researchers who had already made a name for themselves in market research, studying consumer preferences for American businesses, turned their attention to election polling. A watershed event was the presidential election of 1936 between Alf Landon and Franklin Roosevelt. A magazine called Literary Digest conducted a survey by sending ballots (which were also subscription requests) to millions of Americans. Based on this "straw poll," the editors predicted that Landon would win in a landslide. At the same time, the new pollsters were using scientific methods with much smaller samples to predict just the opposite—that Roosevelt would win in a landslide. In fact, one of them, George Gallup, publicly criticized the methods of Literary Digest before the election and all but guaranteed that his prediction would be correct. And of course, it was, demonstrating the effectiveness of careful survey methodology. Gallup's demonstration of the power of careful survey methods encouraged additional local surveys and, in 1948, the first national election survey by the Survey Research Center at the University of Michigan. This work eventually became the American National Election Studies (https://electionstudies.org/) as a collaboration of Stanford University and the University of Michigan, and these studies continue today.

From market research and election polling, survey research made its way into several academic fields, including political science, sociology, and public health—where it continues to be one of the primary approaches to collecting new data. Beginning in the 1930s, psychologists made important advances in questionnaire design, including techniques that are still used today, such as the Likert scale. Survey research has a strong historical association with the social psychological study of attitudes, stereotypes, and prejudice. Early attitude researchers were also among the first psychologists to seek larger and more diverse samples than the convenience samples of university students that were routinely used in psychology (and still are).

Survey research continues to be important in psychology today. For example, survey data have been instrumental in estimating the prevalence of various mental disorders and identifying statistical relationships among those disorders and with various other factors. The National Comorbidity Survey is a large-scale mental health survey conducted in the United States (see http://www.hcp.med.harvard.edu/ncs). In just one part of this survey, nearly 10,000 adults were given a structured mental health interview in their homes in 2002 and 2003. Table 18.1 presents results on the lifetime prevalence of some anxiety, mood, and substance use disorders.

(Lifetime prevalence is the percentage of the population that develops the problem sometime in their lifetime.) Obviously, this kind of information can be of great use both to basic researchers seeking to understand the causes and correlates of mental disorders as well as to clinicians and policymakers who need to understand exactly how common these disorders are.

Table 18.1 *Some Lifetime Prevalence Results From the National Comorbidity Survey*[1]

Lifetime prevalence*			
Disorder	**Total**	**Female**	**Male**
Generalized anxiety disorder	5.7	7.1	4.2
Obsessive-compulsive disorder	2.3	3.1	1.6
Major depressive disorder	16.9	20.2	13.2
Bipolar disorder	4.4	4.5	4.3
Alcohol abuse	13.2	7.5	19.6
Drug abuse	8.0	4.8	11.6
***The lifetime prevalence of a disorder is the percentage of people in the population that develop that disorder at any time in their lives.**			

Table 7.1 Some Lifetime Prevalence Results From the National Comorbidity Survey

Figure 18.1: Table showing the lifetime prevalence (percentage of people who experience a disorder at any point in their lives) of six mental disorders for the total population and separately for females and males. Disorders include generalized anxiety disorder, obsessive-compulsive disorder, major depressive disorder, bipolar disorder, alcohol abuse, and drug abuse. Anxiety, obsessive-compulsive disorder, and major depressive disorder are more prevalent among females, while alcohol abuse and drug abuse are more prevalent among males. Bipolar disorder shows similar prevalence across sexes.

And as the opening example makes clear, survey research can even be used as a data collection method within experimental research to test specific hypotheses about causal relationships between variables. Such studies, when conducted on large and diverse samples, can be a useful supplement to laboratory studies conducted on university students. Survey research is thus a flexible approach that can be used to study a variety of basic and applied research questions.

18.4 Constructing Surveys

The heart of any survey research project is the survey itself. Although it is easy to think of interesting questions to ask people, constructing a good survey is not easy at all. The problem is that the answers people give can be influenced in unintended ways by the wording of the items, the order of the items, the response options provided, and many other factors. At best, these influences add noise to the data. At worst, they result in systematic biases and misleading results. In this section, therefore, we consider some principles for constructing surveys to minimize these unintended effects and thereby maximize the reliability and validity of respondents' answers.

18.4.1 Survey Responding as a Psychological Process

Before looking at specific principles of survey construction, it will help to consider survey responding as a psychological process.

18.4.1.1 A Cognitive Model

Figure 18.1 presents a model of the cognitive processes that people engage in when responding to a survey item (Sudman et al., 1996). Respondents must interpret the question, retrieve relevant information from memory, form a tentative judgment, convert the tentative judgment into one of the response options provided (e.g., a rating on a 1-to-7 scale), and finally edit their response as necessary.

Figure 18.2: Flowchart showing the steps respondents use to answer a survey question: question interpretation, information retrieval, judgment formation, response formatting, and response editing.

Figure 18.1 *Model of the Cognitive Processes Involved in Responding to a Survey Item*[2]

Consider, for example, the following questionnaire item:

How many alcoholic drinks do you consume in a typical day?

- _____ a lot more than average
- _____ somewhat more than average

- _____ average
- _____ somewhat fewer than average
- _____ a lot fewer than average

Although this item at first seems straightforward, it poses several difficulties for respondents. First, they must interpret the question. For example, they must decide whether "alcoholic drinks" include beer and wine (as opposed to just hard liquor) and whether a "typical day" is a typical weekday, typical weekend day, or both. Chang and Krosnick (2003) found that asking about a "typical week" can produce more valid reports than asking about the "past week," although that comparison may not generalize to questions that distinguish typical weekdays from typical weekend days. Once respondents have interpreted the question, they must retrieve relevant information from memory to answer it. But what information should they retrieve, and how should they go about retrieving it? They might think vaguely about some recent occasions on which they drank alcohol, they might carefully try to recall and count the number of alcoholic drinks they consumed last week, or they might retrieve some existing beliefs that they have about themselves (e.g., "I am not much of a drinker"). Then they must use this information to arrive at a tentative judgment about how many alcoholic drinks they consume in a typical day. For example, this mental calculation might mean dividing the number of alcoholic drinks they consumed last week by seven to come up with an average number per day. Then they must format this tentative answer in terms of the response options actually provided. In this case, the options pose additional problems of interpretation. For example, what does "average" mean, and what would count as "somewhat more" than average? Finally, they must decide whether they want to report the response they have come up with or whether they want to edit it in some way. For example, if they believe that they drink a lot more than average, they might not want to report that for fear of looking bad in the eyes of the researcher, so instead, they may opt to select the "somewhat more than average" response option.

From this perspective, what at first appears to be a simple matter of asking people how much they drink (and receiving a straightforward answer from them) turns out to be much more complex.

18.4.2 Context Effects on Survey Responses

Again, this complexity can lead to unintended influences on respondents' answers. These are often referred to as ***context effects*** because they arise from the setting or sequence in which an item appears rather than from its substantive content (Schwarz & Strack, 1990). An ***item-order effect*** occurs when an earlier item changes how respondents interpret a later item or which information they retrieve. For example, Strack and colleagues asked college students about their general life satisfaction and dating frequency (Strack et al., 1988). When life satisfaction was asked first, the correlation was only −.12. When dating frequency was asked first, the correlation was +.66 because dating information had become more accessible when respondents judged their lives overall.

Response options can also affect answers (Schwarz, 1999). When people are asked how often they are "really irritated" and the options range from "less than once a year" to "more than once a month," they tend to think of major irritations and report them infrequently. When the options range from "less than once a day" to "several times a month," they are more likely to think of minor irritations and report them frequently. Respondents may also treat a scale's middle categories as signals of what is normal; for example, they report more television viewing when the response scale is centered on 4 hours rather than 2 hours. When item or response-option order has no substantive rationale, researchers can rotate, counterbalance, or randomize the order. These procedures distribute order effects across conditions rather than allowing one position to receive a systematic advantage. In elections, for example, appearing first on the ballot produced an average advantage of about 2.5 percentage points among undecided voters (Miller & Krosnick, 1998).

18.5 Writing Survey Items

18.5.1 Types of Items

Questionnaire items can be open-ended, closed-ended, or partially closed-ended. ***Open-ended items*** simply ask a question and allow respondents to answer in whatever way they choose. The following are examples of open-ended questionnaire items.

- "What is the most important thing to teach children to prepare them for life?"
- "Please describe a time when you were discriminated against because of your age."
- "Is there anything else you would like to tell us about?"

Open-ended items are useful when researchers do not know how respondents might answer or want to avoid constraining their responses. They are often used in the early stages of a project or when detailed explanations are important. Open-ended items are relatively easy to write, but they require more effort from respondents and are more likely to be skipped. They also take more effort to analyze because answers must be transcribed, coded, and examined using qualitative methods such as content analysis. They are most appropriate when the likely range of answers is uncertain or when numeric answers can be categorized later.

Closed-ended items, also called restricted-response items, ask a question and provide a fixed set of response options from which respondents choose. The alcohol item just mentioned is an example, as are the following:

How old are you?

- _____ Under 18
- _____ 18 to 34

- _____ 35 to 49
- _____ 50 to 70
- _____ Over 70

On a scale of 0 (no pain at all) to 10 (worst pain ever experienced), how much pain are you in right now?

Have you ever in your adult life been depressed for a period of 2 weeks or more? Yes No

Closed-ended items are used when researchers have a good idea of the different responses that respondents might make. They are more quantitative in nature, so they are also used when researchers are interested in a well-defined variable or construct such as respondents' level of agreement with some statement, perceptions of risk, or frequency of a particular behavior. Closed-ended items are more difficult to write because they must include an appropriate set of response options. However, they are relatively quick and easy for respondents to complete. They are also much easier for researchers to analyze because the responses can be easily converted to numbers and entered into a spreadsheet. For these reasons, closed-ended items are much more common.

All closed-ended items include response options from which a respondent must choose. For categorical variables, categories should be listed and respondents should be told whether they may select one or more. For quantitative variables, researchers often use a ***rating scale***—an ordered set of response options. Figure 18.2 shows several examples. Five- and seven-point scales are common, but the number of options should reflect how precisely respondents can make the judgment. A unipolar frequency scale might range from Never to Always, whereas a bipolar evaluation scale might range from Strongly Dislike to Strongly Like. Labels should be balanced, clear, and applied consistently; numerical codes can be assigned during analysis. Figure 18.2 also shows a visual analog scale, on which respondents mark a position along a continuous line.

Partially closed-ended items, also called partially open-ended items, combine fixed response options with an open response such as "Other (please specify)." They are useful when common answers are known but the list may not be exhaustive. For all categorical items, response options should be mutually exclusive when only one response is allowed and collectively exhaustive. If categories can overlap, respondents should be instructed to select all that apply.

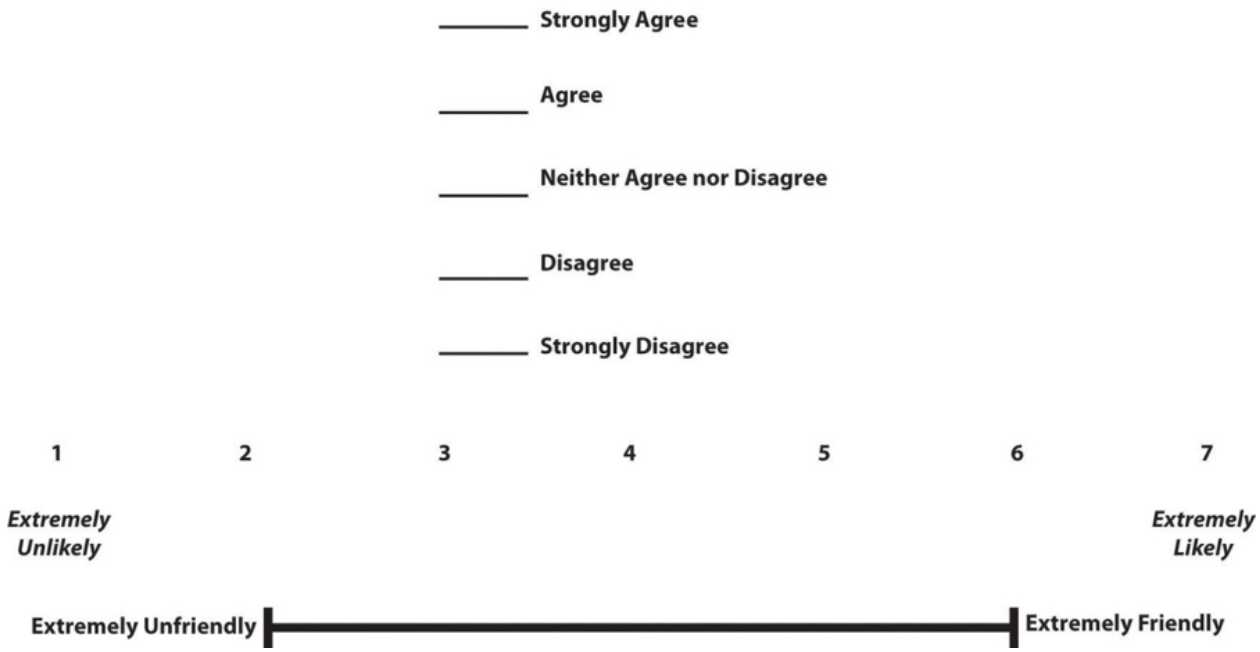


Figure 18.3: Examples of three common rating scales used in survey questions: a five-point agreement scale ranging from Strongly Agree to Strongly Disagree, a seven-point likelihood scale ranging from Extremely Unlikely to Extremely Likely, and a visual analog scale with endpoints labeled Extremely Unfriendly and Extremely Friendly.

Figure 18.2 *Example Rating Scales for Closed-Ended Questionnaire Items*[3]

18.5.2 What Is a Likert Scale?

A ***Likert item*** asks respondents to indicate their level of agreement with a statement, usually on a five- or seven-point ordered scale. A ***Likert scale*** combines scores from several Likert items intended to measure the same attitude or construct. The term should not be used for every rating scale; for example, a single 0-to-10 pain item is a rating scale but is not, by itself, a Likert scale.

Rensis Likert developed this approach to attitude measurement in the 1930s (Likert, 1932). Respondents indicate agreement with several statements, often using options such as Strongly Agree, Agree, Neither Agree nor Disagree, Disagree, and Strongly Disagree. Responses are assigned numerical values and combined across items. Items written in the opposite direction must be reverse-scored before the total or average is calculated so that higher scores have a consistent meaning.

18.5.3 Writing Effective Items

We can now consider some principles of writing questionnaire items that minimize unintended context effects and maximize the reliability and validity of respondents' responses. A rough guideline for writing questionnaire items is provided by the ***BRUSO*** model (Peterson, 2000). An acronym, BRUSO stands for "brief," "relevant," "unambiguous," "specific," and "objective." Effective questionnaire items are brief and to the point. They avoid long, overly technical, or unnecessary words. This brevity makes them easier for respondents to understand and faster for them to complete. Effective questionnaire items are also relevant to the research question. If a respondent's sexual orientation, marital status, or income is not relevant, then items on them should probably not be included. Again, this makes the questionnaire faster to complete, but it also avoids annoying respondents with what they will rightly perceive as irrelevant or even "nosy" questions. Effective questionnaire items are also unambiguous; they can be interpreted in only one way. Part of the problem with the alcohol item presented earlier in this section is that different respondents might have different ideas about what constitutes "an alcoholic drink" or "a typical day." Effective questionnaire items are also specific so that it is clear to respondents what their response should be about and clear to researchers what it is about. A common problem here is closed-ended items that are "double-barreled." They ask about two conceptually separate issues but allow only one response. For example, "Please rate the extent to which you have been feeling anxious and depressed." This item should probably be split into two separate items—one about anxiety and one about depression. Finally, effective questionnaire items are objective in the sense that they do not reveal the researcher's own opinions or lead respondents to answer in a particular way. Table 18.2 shows some examples of poor and effective questionnaire items based on the BRUSO criteria. The best way to know how people interpret the wording of the question is to conduct a pilot test and ask a few people to explain how they interpreted the question.

Table 18.2 *BRUSO Model of Writing Effective Questionnaire Items, Plus Examples*[4]

Criterion	Poor	Effective
B—Brief	"Are you now or have you ever been the possessor of a firearm?"	"Have you ever owned a gun?"
R—Relevant	"What is your sexual orientation?"	Do not include this item unless it is clearly relevant to the research.
U—Unambiguous	"Are you a gun person?"	"Do you currently own a gun?"
S—Specific	"How much have you read about the new gun control measure and sales tax?"	"How much have you read about the new sales tax?"
O—Objective	"How much do you support the new gun control measure?"	"What is your view of the new gun control measure?"

Table 7.2 BRUSO Model of Writing Effective Questionnaire Items, Plus Examples

Figure 18.4: Table comparing poor and effective questionnaire items using the BRUSO model. The criteria are Brief, Relevant, Unambiguous, Specific, and Objective. For each criterion, the table provides an example of a poorly written survey question alongside a revised, more effective version.

For rating scales, five or seven response options generally provide useful precision without demanding distinctions that respondents cannot make. More options may be appropriate for familiar judgments such as pain or likelihood on a 0-to-10 scale. Endpoints should be balanced around a neutral or modal midpoint when the construct is bipolar.

An unbalanced likelihood scale might look like this:

Unlikely | *Somewhat Likely* | *Likely* | *Very Likely* | *Extremely Likely*

A balanced version might look like this:

Extremely Unlikely | Somewhat Unlikely | As Likely as Not | Somewhat Likely | Extremely Likely

Note, however, that a middle or neutral response option does not have to be included. Researchers sometimes choose to leave it out because they want to encourage respondents to think more deeply about their response and not simply choose the middle option by default. However, including middle alternatives on bipolar dimensions can be used to allow people to choose an option that is neither.

18.6 Formatting the Survey

Writing effective items is only one part of constructing a survey. For one thing, every survey should have a written or spoken introduction that serves two basic functions (Peterson, 2000). One is to encourage respondents to participate in the survey. In many types of research, such encouragement is not necessary either because participants do not know they are in a study (as in naturalistic observation) or because they are part of a subject pool and have already shown their willingness to participate by signing up and showing up for the study. Survey research usually catches respondents by surprise when they answer their phone, go to their mailbox, or check their email—and the researcher must make a good case for why they should agree to participate. Thus, the introduction should briefly explain the purpose of the survey and its importance, provide information about the sponsor of the survey (university-based surveys tend to generate higher response rates), acknowledge the importance of the respondent's participation, and describe any incentives for participating.

The second function of the introduction is to establish informed consent. This involves describing to respondents everything that might affect their decision to participate. This includes the topics covered by the survey, the amount of time it is likely to take, the respondent's option to withdraw at any time, confidentiality issues, and so on. Written consent forms are not always used in survey research (when the research is of minimal risk and completion of the survey instrument is often accepted by the IRB as evidence of consent to participate), so it is important that this part of the introduction be well-documented and presented clearly and in its entirety to every respondent.

The introduction should be followed by the substantive questionnaire items. But first, it is important to present clear instructions for completing the questionnaire, including examples of how to use any unusual response scales. Remember that the introduction is the point at which respondents are usually most interested and least fatigued, so it is good practice to start with the most important items for purposes of the research and proceed to less important items. Items should also be grouped by topic or by type. For example, items using the same rating scale (e.g., a 5-point agreement scale) should be grouped together, if possible, to make things faster and easier for respondents. Demographic items are often presented last because they are least interesting to respondents but also easy to answer in the event respondents have become tired or bored. Of course, any survey should end with an expression of appreciation to the respondent.

18.7 Evaluating Survey Reliability and Validity

18.7.1 Reliability of Survey Measures

Reliability is the consistency of a measure. ***Test-retest reliability*** is estimated by administering the same measure to the same respondents on two occasions and correlating the scores.

It is most informative when the construct should be stable and the interval is long enough to reduce memory effects but short enough to limit genuine change. ***Parallel-forms reliability*** compares scores from two equivalent versions designed to measure the same construct. Equivalent forms can reduce memory effects, but they are difficult to create because item content and difficulty must be closely matched.

Internal consistency concerns whether items intended to measure the same construct produce coherent scores in a single administration. ***Cronbach's alpha*** is a common coefficient for multi-item scales, but a high alpha can also result from many redundant items and does not prove that a scale is unidimensional or valid. ***Split-half reliability*** correlates scores from two comparable halves of a measure. ***KR-20*** is a related internal-consistency coefficient for items scored dichotomously, such as correct/incorrect or yes/no. It uses information from all items rather than depending on one arbitrary split and is equivalent to alpha when items are scored 0 or 1.

18.7.1.1 Increasing Reliability

Reliability can often be improved by using several well-targeted items for each construct, provided that the items add relevant content rather than merely repeat the same wording. Administration conditions and instructions should be standardized so that respondents receive the same prompts, time limits, definitions, and response options. Questions should use clear, concise language; avoid double-barreled or ambiguous wording; and specify the relevant time frame.

Scoring should follow explicit rules established before analysis. Reverse-worded items must be recoded correctly, missing responses must be handled consistently, and any judgments required in coding open-ended answers should use trained coders and a clear rubric. Pilot testing can identify items that respondents misinterpret or that do not contribute to a stable score. A representative sample reduces sampling bias and improves population estimates, but representativeness is not itself a form of score reliability; it strengthens generalizability and the accuracy of descriptive conclusions.

18.7.2 Checking Validity

Validity concerns whether the evidence supports the intended interpretation and use of survey scores. ***Content validity*** is supported when the items adequately represent the full domain of the construct. Researchers commonly define the domain in advance, compare items with that definition, and obtain judgments from subject-matter experts and members of the intended population.

Construct validity is evaluated by testing whether scores relate to other variables as theory predicts. Convergent evidence is shown by strong relationships with other measures of the same or closely related constructs; divergent (or discriminant) evidence is shown by weak

relationships with measures of distinct constructs. When related constructs are measured at the same time, the resulting concurrent evidence can contribute to the construct-validity argument, although the term concurrent validity is also used for a form of criterion evidence.

Criterion-related validity is supported when a survey score relates to a meaningful external criterion. ***Concurrent validity*** compares the survey score with a criterion measured at approximately the same time, such as a screening score and a current clinical assessment. ***Predictive validity*** is supported when the score forecasts a later outcome, such as whether an admissions measure predicts subsequent performance. Validity is purpose-specific and cumulative; no single coefficient establishes that a survey is valid for every population or use.

18.8 Survey Administration Methods

In-person interviews use a trained interviewer to ask questions and record answers. Structured interviews can clarify standardized instructions, reach respondents who might not complete a written form, and permit relevant observations. They are usually the most expensive method and are vulnerable to interviewer effects and social-desirability responding. ***Telephone surveys*** provide some personal contact at lower cost, but screening calls, unknown numbers, unequal phone access, and the absence of a complete sampling frame can reduce coverage and response rates.

Mail surveys can reach respondents without internet access and allow them to answer privately at their own pace. Printing, postage, delayed returns, and nonresponse are important limitations. ***Internet or online surveys*** can use branching, validation rules, randomization, and rapid automated data capture at relatively low cost. Their quality still depends on the sampling frame, device compatibility, accessibility, identity checks, privacy protections, and the extent of nonresponse or self-selection.

Group-administered surveys are completed by many respondents at the same time in a classroom, workplace, clinic, or other shared setting. They are efficient and allow standardized instructions, but attendance may define the sample, privacy can be difficult to protect, and the presence of peers or authority figures can influence answers. Across all modes, researchers should standardize administration as much as possible and select the method that best fits the population, sensitivity of the questions, sampling plan, and available resources.

Finally, it is important to note that some of the concerns that people have about collecting data online (e.g., that internet-based findings differ from those obtained with other methods) have been found to be myths. Table 18.3 (adapted from Gosling et al., 2004) addresses three such preconceptions about data collected in web-based studies:

Table 18.3 *Some Preconceptions and Findings Pertaining to Web-Based Studies*[5]

Table 7.3 Some Preconceptions and Findings Pertaining to Web-based Studies

Preconception	Finding
Internet samples are not demographically diverse	Internet samples are more diverse than traditional samples in many domains, although they are not completely representative of the population
Internet samples are maladjusted, socially isolated, or depressed	Internet users do not differs from nonusers on markers of adjustment and depression
Internet-based findings differ from those obtained with other methods	Evidence so far suggests that internet-based findings are consistent with findings based on traditional methods (e.g., on self-esteem, personality), but more data are needed.

18.9 Online Survey Creation

Online platforms change quickly, and their access models, pricing, security controls, and features vary by plan and institution. Researchers should verify that a platform supports the study's branching, randomization, accessibility, consent, export, and data-security requirements before collecting data. Current examples include:

- SurveyMonkey—a general-purpose commercial survey builder; available features depend on the plan.
- Qualtrics—a commercial research and survey system with advanced logic, distribution, and analysis tools.
- REDCap—an institutionally supported research data-collection platform; access commonly depends on a participating organization.
- LimeSurvey—an open-source survey platform with hosted and self-hosted options.
- PsyToolkit—a noncommercial platform designed for psychology surveys and browser-based experiments.

18.9.1 Privacy and Data Governance

Platform selection should be based on the full data flow rather than only the provider's headquarters or server country. Researchers should identify what data are collected; where primary data, backups, and logs are processed and stored; who can access them; how data are encrypted; how long they are retained; how deletion requests are handled; and whether

subprocessors or artificial-intelligence features receive survey content. Consent materials should accurately describe the practices that matter to respondents.

Before data collection, researchers should follow applicable law and institutional policy, obtain required ethics or institutional review board approval, and consult their privacy or information-security office when sensitive data are involved. Institutional accounts, written data-processing terms, role-based access, retention schedules, and incident-response procedures help turn privacy requirements into an ongoing risk-management process rather than a one-time location decision (National Institute of Standards and Technology [NIST], 2020).

18.9.2 Online Recruitment Platforms

Survey software collects responses, whereas recruitment platforms connect researchers with potential participants. Prolific is designed to recruit and manage participants for studies hosted in another survey or experiment tool. Amazon Mechanical Turk (MTurk) is a general crowdsourcing marketplace that supports behavioral studies among many other task types (Amazon Web Services, n.d.; Prolific, 2025). On MTurk, requesters choose the reward, and Amazon charges service fees on rewards and bonuses, with additional fees for some task configurations and qualification options; researchers should consult the current pricing page when budgeting (Amazon Mechanical Turk, n.d.).

Online recruitment pools are self-selected and should not be assumed to represent a population. Researchers should pilot the study, use transparent and fair compensation, prespecify exclusion rules, prevent duplicate participation when appropriate, and combine several indicators when screening for inattention, bots, or misrepresentation. Quality controls should protect the data without unfairly excluding legitimate participants (Aguinis et al., 2021; Peer et al., 2022).

18.10 Media Attributions

[1-5] Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

18.11 Text Attributions

Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

18.12 References

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management, 47*(4), 823–837. https://doi.org/10.1177/0149206320969787

Amazon Mechanical Turk. (n.d.). *Amazon Mechanical Turk pricing.* Retrieved July 30, 2026, from https://requester.mturk.com/pricing

Amazon Web Services. (n.d.). *What is Amazon Mechanical Turk?* https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester/WhatIs.html

Chang, L., & Krosnick, J. A. (2003). Measuring the frequency of regular behaviors: Comparing the "typical week" to the "past week." *Sociological Methodology, 33*(1), 55–80.

Converse, J. M. (1987). *Survey research in the United States: Roots and emergence, 1890–1960.* University of California Press.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*(2), 93–104. https://doi.org/10.1037/0003-066X.59.2.93

Lerner, J. S., Gonzalez, R. M., Small, D. A., & Fischhoff, B. (2003). Effects of fear and anger on perceived risks of terrorism: A national field experiment. *Psychological Science, 14*(2), 144–150. https://doi.org/10.1111/1467-9280.01433

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 1–55.

Miller, J. M., & Krosnick, J. A. (1998). The impact of candidate name order on election outcomes. *Public Opinion Quarterly, 62*(3), 291–330. https://doi.org/10.1086/297848

National Institute of Standards and Technology. (2020). *NIST Privacy Framework: A tool for improving privacy through enterprise risk management (Version 1.0).* https://doi.org/10.6028/NIST.CSWP.01162020

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54*, 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Peterson, R. A. (2000). *Constructing effective questionnaires.* Sage.

Prolific. (2025, October 17). *About Prolific.* https://researcher-help.prolific.com/en/articles/449608-about-prolific

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93–105. https://doi.org/10.1037/0003-066X.54.2.93

Schwarz, N., & Strack, F. (1990). Context effects in attitude surveys: Applying cognitive theory to social research. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 2, pp. 31–50). Wiley.

Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology, 18*(5), 429–442. https://doi.org/10.1002/ejsp.2420180505

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* Jossey-Bass.

19 Chapter 19: Correlational Research

19.1 What Is Correlational Research?

Correlational research is a type of non-experimental research in which the researcher measures two variables (binary or continuous) and assesses the statistical relationship (i.e., the correlation) between them with little or no effort to control extraneous variables. There are many reasons that researchers interested in statistical relationships between variables would choose to conduct a correlational study rather than an experiment. The first is that they do not believe that the statistical relationship is a causal one or are not interested in causal relationships. Two goals of science are to describe and to predict and the correlational research strategy allows researchers to achieve both of these goals. Specifically, this strategy can be used to describe the strength and direction of the relationship between two variables and if there is a relationship between the variables then the researchers can use scores on one variable to predict scores on the other (using a statistical technique called regression).

Another reason that researchers would choose to use a correlational study rather than an experiment is that the statistical relationship of interest is thought to be causal, but the researcher *cannot* manipulate the independent variable because it is impossible, impractical, or unethical. For example, while a researcher might be interested in the relationship between how frequently people use cannabis and their memory abilities, they cannot ethically manipulate how frequently people use cannabis. As such, they must rely on the correlational research strategy; they must simply measure how frequently people use cannabis, assess their memory abilities using a standardized test, and then determine whether the frequency of cannabis use is statistically related to memory test performance.

Correlation is also used to establish the reliability and validity of measurements. For example, a researcher might evaluate the validity of a brief extraversion test by administering it to a large group of participants along with a longer extraversion test that has already been shown to be valid. This researcher might then check to see whether participants' scores on the brief test are strongly correlated with their scores on the longer one. Neither test score is thought to cause the other, so there is no independent variable to manipulate. In fact, the terms *independent variable* and *dependent variable* do not apply to this kind of research.

Another strength of correlational research is that it is often higher in external validity than experimental research. There is typically a trade-off between internal validity and external validity. As greater controls are added to experiments, internal validity is increased but often at the expense of external validity as artificial conditions are introduced that do not exist in

reality. In contrast, correlational studies typically have low internal validity because nothing is manipulated or controlled but they often have high external validity. Since nothing is manipulated or controlled by the experimenter the results are more likely to reflect relationships that exist in the real world.

Finally, extending upon this trade-off between internal and external validity, correlational research can help to provide converging evidence for a theory. If a theory is supported by a true experiment that is high in internal validity as well as by a correlational study that is high in external validity then the researchers can have more confidence in the validity of their theory.

19.1.1 Does Correlational Research Always Involve Quantitative Variables?

A common misconception among beginning researchers is that correlational research must involve two quantitative variables, such as scores on two extraversion tests or the number of daily hassles and number of symptoms people have experienced. However, the defining feature of correlational research is that the two variables are measured—neither one is manipulated—and this is true regardless of whether the variables are quantitative or categorical. Imagine, for example, that a researcher administers the Rosenberg Self-Esteem Scale to 50 American college students and 50 Japanese college students. Although this "feels" like a between-subjects experiment, it is a correlational study because the researcher did not manipulate the students' nationalities.

Figure 19.1 shows data from a hypothetical study on the relationship between whether people make a daily list of things to do (a "to-do list") and stress. Notice that it is unclear whether this is an experiment or a correlational study because it is unclear whether the independent variable was manipulated. If the researcher randomly assigned some participants to make daily to-do lists and others not to, then it is an experiment. If the researcher simply asked participants whether they made daily to-do lists, then it is a correlational study. The distinction is important because if the study was an experiment, then it could be concluded that making the daily to-do lists reduced participants' stress. But if it was a correlational study, it could only be concluded that these variables are statistically related. Perhaps being stressed has a negative effect on people's ability to plan ahead. Or perhaps people who are more conscientious are more likely to make to-do lists and less likely to be stressed. The crucial point is that what defines a study as experimental or correlational is not the variables being studied, nor whether the variables are quantitative or categorical, nor the type of graph or statistics used to analyze the data. What defines a study is *how* the study is conducted.

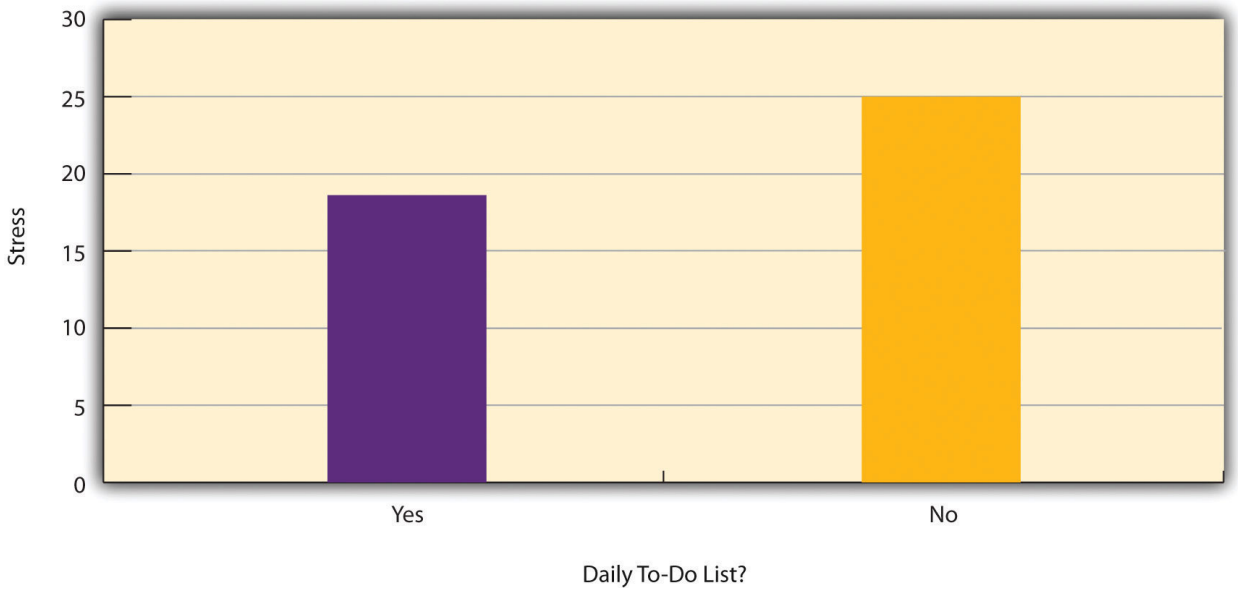


Figure 19.1: Bar graph comparing average stress levels between people who make a daily to-do list and those who do not. Participants who make a daily to-do list show lower average stress levels than participants who do not make a daily to-do list.

Figure 19.1 *Results of a Hypothetical Study on Whether People Who Make Daily To-Do Lists Experience Less Stress Than People Who Do Not Make Such Lists*[1]

19.2 Data Collection in Correlational Research

The defining feature of correlational research is that neither variable is manipulated. It does not matter how or where the variables are measured. A researcher could have participants come to a laboratory to complete a computerized backward digit span task and a computerized risky decision-making task and then assess the relationship between participants' scores on the two tasks. Or a researcher could go to a shopping mall to ask people about their attitudes toward the environment and their shopping habits and then assess the relationship between these two variables. Both of these studies would be correlational because no independent variable is manipulated.

19.3 Correlations Between Quantitative Variables

Correlations between quantitative variables are often presented using ***scatterplots***. Figure 19.2 shows some hypothetical data on the relationship between the amount of stress people

are under and the number of physical symptoms they have. Each point in the scatterplot represents one person's score on both variables. For example, the circled point in Figure 19.2 represents a person whose stress score was 10 and who had three physical symptoms. Taking all the points into account, one can see that people under more stress tend to have more physical symptoms. This is a good example of a ***positive relationship***, in which higher scores on one variable tend to be associated with higher scores on the other. In other words, they move in the same direction, either both up or both down. A ***negative relationship*** is one in which higher scores on one variable tend to be associated with lower scores on the other. In other words, they move in opposite directions. There is a negative relationship between stress and immune system functioning, for example, because higher stress is associated with lower immune system functioning.

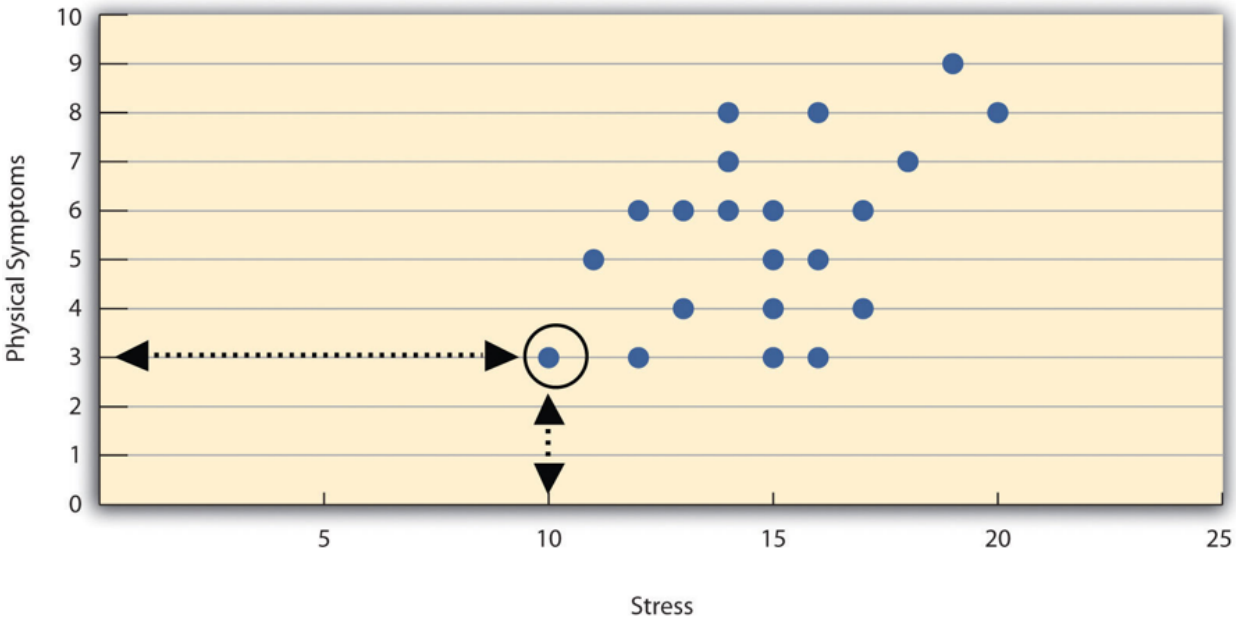


Figure 19.2: Scatterplot illustrating a positive relationship between stress and physical symptoms. As stress increases, the number of physical symptoms generally increases. A circled point identifies a participant with a stress score of 10 and three physical symptoms.

Figure 19.2 *Scatterplot Showing a Hypothetical Positive Relationship Between Stress and Number of Physical Symptoms*[2]

The strength of a correlation between quantitative variables is typically measured using a statistic called ***Pearson's Correlation Coefficient (or Pearson's r)***. As Figure 19.3 shows, Pearson's r ranges from −1.00 (the strongest possible negative relationship) to +1.00 (the strongest possible positive relationship). A value of 0 means there is no relationship between

the two variables. When Pearson's r is 0, the points on a scatterplot form a shapeless "cloud." As its value moves toward −1.00 or +1.00, the points come closer and closer to falling on a single straight line. Correlation coefficients near ±.10 are considered small, values near ±.30 are considered medium, and values near ±.50 are considered large. Notice that the sign of Pearson's r is unrelated to its strength. Pearson's r values of +.30 and −.30, for example, are equally strong; it is just that one represents a moderate positive relationship and the other a moderate negative relationship. With the exception of reliability coefficients, most correlations that we find in psychology are small or moderate in size. The R Psychologist interactive correlation visualization, created by Kristoffer Magnusson, provides an excellent interactive visualization of correlations that permits you to adjust the strength and direction of a correlation while witnessing the corresponding changes to the scatterplot.

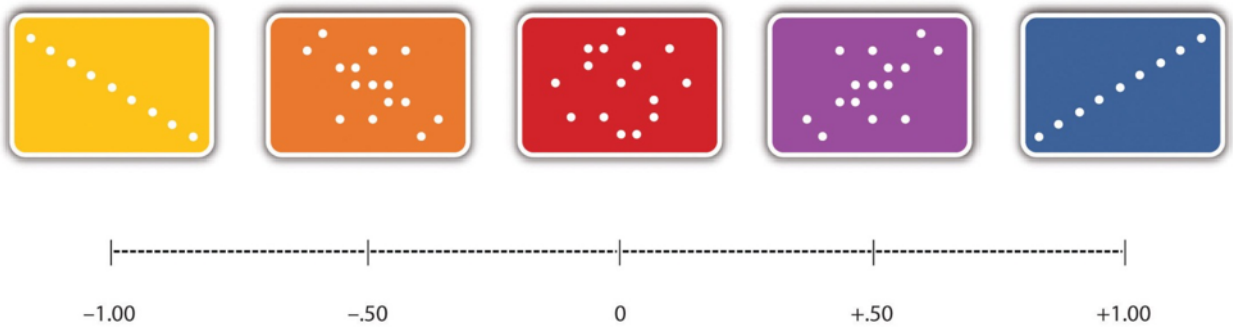


Figure 19.3: Five scatterplots demonstrating Pearson's correlation coefficient (r). The figure shows examples of strong negative, moderate negative, no, moderate positive, and strong positive correlations, illustrating how the direction and strength of a relationship change as Pearson's r ranges from −1.00 to +1.00.

Figure 19.3 *Range of Pearson's r, From −1.00 (Strongest Possible Negative Relationship), Through 0 (No Relationship), to +1.00 (Strongest Possible Positive Relationship)*[3]

There are several situations in which the value of Pearson's r can be misleading. Pearson's r is a good measure only for linear relationships, in which the points are best approximated by a straight line. It is not a good measure for nonlinear relationships, in which the points are better approximated by a curved line. Figure 19.4, for example, shows a hypothetical relationship between the amount of sleep people get per night and their level of depression. In this example, the line that best approximates the points is a U-shaped curve because people who get about eight hours of sleep tend to be the least depressed. Those who get too little sleep and those who get too much sleep tend to be more depressed. Even though Figure 19.4 shows a fairly strong relationship between depression and sleep, Pearson's r would be close to zero because the points in the scatterplot are not well fit by a single straight line. This means that it is important to make a scatterplot and confirm that a relationship is approximately linear before using Pearson's r.

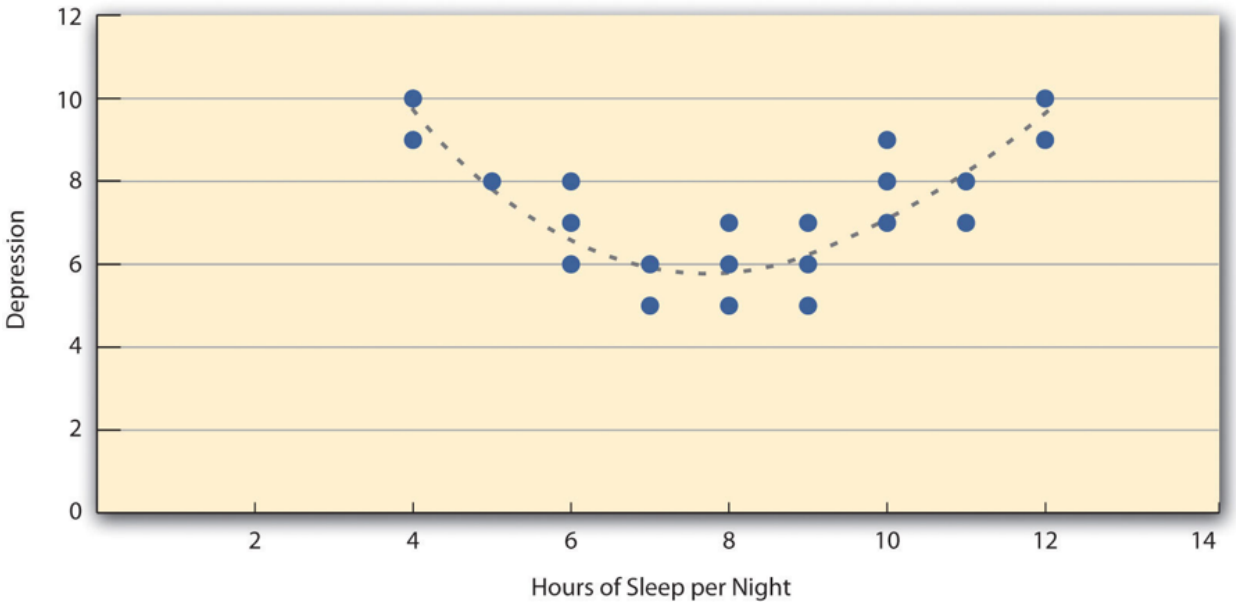


Figure 19.4: Scatterplot showing the relationship between hours of sleep per night and depression. The x-axis represents hours of sleep per night and the y-axis represents depression. The data form a U-shaped curve, indicating a nonlinear relationship in which depression is lowest for people who get about eight hours of sleep and higher for people who get either less or more sleep.

Figure 19.4 *Hypothetical Nonlinear Relationship Between Sleep and Depression*[4]

Another common situation in which the value of Pearson's r can be misleading is when one or both of the variables have a limited range in the sample relative to the population. This problem is referred to as ***restriction of range***. Assume, for example, that there is a strong negative correlation between people's age and their enjoyment of hip hop music as shown by the scatterplot in Figure 19.5. Pearson's r here is $-.77$. However, if we were to collect data only from 18- to 24-year-olds—represented by the shaded area of Figure 19.5—then the relationship would seem to be quite weak. In fact, Pearson's r for this restricted range of ages is 0. It is a good idea, therefore, to design studies to avoid restriction of range. For example, if age is one of your primary variables, then you can plan to collect data from people of a wide range of ages. Because restriction of range is not always anticipated or easily avoidable, however, it is good practice to examine your data for possible restriction of range and to interpret Pearson's r in light of it.

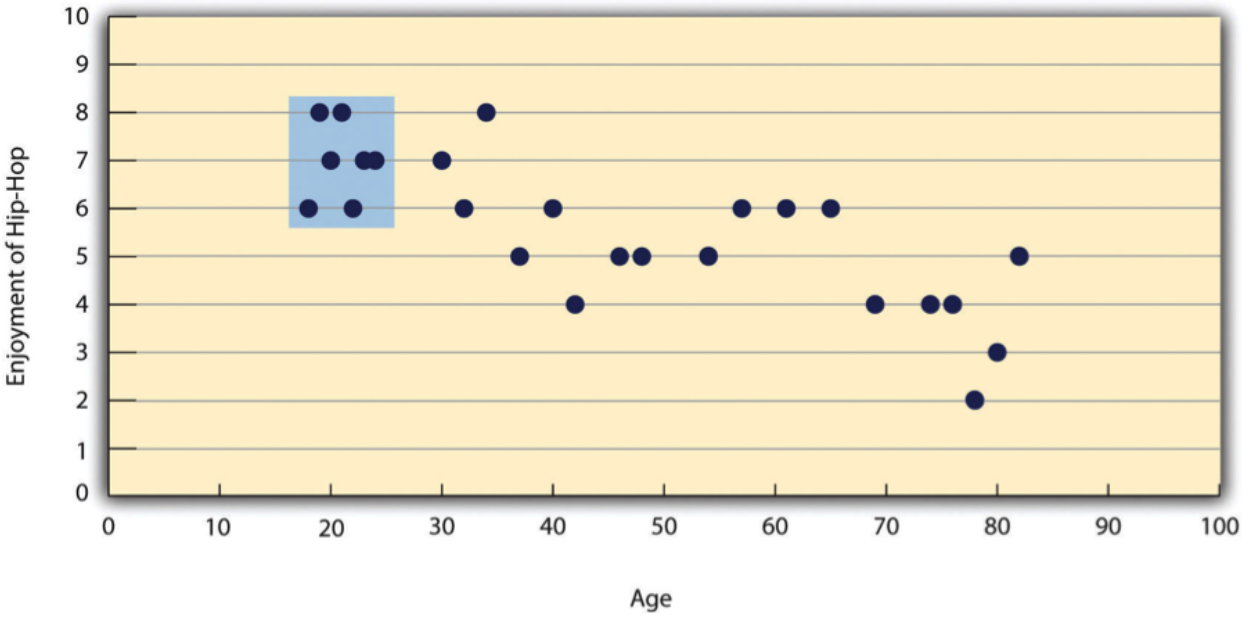


Figure 19.5: Scatterplot illustrating restriction of range. The overall data show a strong negative relationship between age and enjoyment of hip-hop music, but the highlighted subgroup of participants aged 18 to 24 shows little or no relationship, demonstrating how restricting the range of a variable can weaken an observed correlation.

Figure 19.5 *Hypothetical Data Showing How a Strong Overall Correlation Can Appear to Be Weak When One Variable Has a Restricted Range* [5]

Another issue that can affect a correlation is the reliability of the measurements. If the measures produce inconsistent scores, the correlation may appear weaker than the true relationship between the variables. Also, common statistical tests for correlations generally assume that the scores form roughly bell-shaped distributions without extreme outliers, and the spread of the scores is reasonably consistent across the relationship. When these assumptions are not met, researchers may need to use a different type of correlation or another statistical technique.

19.3.1 Correlation Does Not Imply Causation

You have probably heard repeatedly that "Correlation does not imply causation." An amusing example of this comes from a 2012 study that showed a positive correlation (Pearson's $r = 0.79$) between the per capita chocolate consumption of a nation and the number of Nobel prizes awarded to citizens of that nation (Messerli, 2012). It seems clear, however, that this does not mean that eating chocolate causes people to win Nobel prizes, and it would not make sense to try to increase the number of Nobel prizes won by recommending that parents feed their children more chocolate.

There are two reasons that correlation does not imply causation. The first is called the ***directionality problem***. Two variables, X and Y, can be statistically related because X causes Y or because Y causes X. Consider, for example, a study showing that whether or not people exercise is statistically related to how happy they are—such that people who exercise are happier on average than people who do not. This statistical relationship is consistent with the idea that exercising causes happiness, but it is also consistent with the idea that happiness causes exercise. Perhaps being happy gives people more energy or leads them to seek opportunities to socialize with others by going to the gym. The second reason that correlation does not imply causation is called the ***third-variable problem***. Two variables, X and Y, can be statistically related not because X causes Y, or because Y causes X, but because some third variable, Z, causes both X and Y. For example, the fact that nations that have won more Nobel prizes tend to have higher chocolate consumption probably reflects geography in that European countries tend to have higher rates of per capita chocolate consumption *and* invest more in education and technology (once again, per capita) than many other countries in the world. Similarly, the statistical relationship between exercise and happiness could mean that some third variable, such as physical health, causes both of the others. Being physically healthy could cause people to exercise and cause them to be happier. Correlations that are a result of a third-variable are often referred to as ***spurious correlations***.

Excellent and amusing examples of spurious correlations can be found on Tyler Vigen's Spurious Correlations website.

19.4 Complex Correlation

Researchers conduct correlational studies rather than experiments when they are interested in noncausal relationships or when they are interested in causal relationships but the independent variable cannot be manipulated for practical or ethical reasons. In this section, we look at some approaches to complex correlational research that involve measuring several variables and assessing the relationships among them.

19.4.1 Assessing Relationships Among Multiple Variables

Most complex correlational research involves measuring several variables—either binary or continuous—and then assessing the statistical relationships among them. For example, researchers Nathan Radcliffe and William Klein studied a sample of middle-aged adults to see how their level of optimism (measured by using a short questionnaire called the Life Orientation Test) relates to several other variables related to having a heart attack (Radcliffe & Klein, 2002). These included their health, their knowledge of heart attack risk factors, and their beliefs about their own risk of having a heart attack. They found that more optimistic participants were healthier (e.g., they exercised more and had lower blood pressure), knew about heart attack risk factors, and correctly believed their own risk to be lower than that of their peers.

In another example, Ernest Jouriles and his colleagues measured adolescents' experiences of physical and psychological relationship aggression and their psychological distress. Because measures of physical aggression (such as the Conflict in Adolescent Dating Relationships Inventory and the Relationship Violence Interview) often tend to result in highly skewed distributions, the researchers transformed their measures of physical aggression into a dichotomous (i.e., binary) measure (0 = did not occur, 1 = did occur). They did the same with their measures of psychological aggression and then measured the correlations among these variables, finding that adolescents who experienced physical aggression were moderately likely to also have experienced psychological aggression and that experiencing psychological aggression was related to symptoms of psychological distress (Jouriles et al., 2009).

This approach is often used to assess the validity of new psychological measures. For example, when John Cacioppo and Richard Petty created their Need for Cognition Scale—a measure of the extent to which people like to think and value thinking—they used it to measure the need for cognition for a large sample of college students, along with three other variables: intelligence, socially desirable responding (the tendency to give what one thinks is the "appropriate" response), and dogmatism (Cacioppo & Petty, 1982). The results of this study are summarized in Table 19.1, which is a ***correlation matrix*** showing the correlation (Pearson's r) between every possible pair of variables in the study. For example, the correlation between the need for cognition and intelligence was +.39, the correlation between intelligence and socially desirable responding was +.02, and so on. (Only half the matrix is filled in because the other half would contain exactly the same information. Also, because the correlation between a variable and itself is always +1.00, these values are replaced with dashes throughout the matrix.) In this case, the overall pattern of correlations was consistent with the researchers' ideas about how scores on the need for cognition should be related to these other constructs.

Table 19.1 *Correlation Matrix Showing Correlations Among the Need for Cognition and Three Other Variables Based on Research by Cacioppo and Petty (1982)*[6]

Table 6.1 Correlation Matrix Showing Correlations Among the Need for Cognition and Three Other Variables Based on Research by Cacioppo and Petty (1982)

	Need for cognition	Intelligence	Social desirability	Dogmatism
Need for cognition	—			
Intelligence	+.39	—		
Social desirability	+.08	+.02	—	
Dogmatism	−.27	−.23	+.03	—

Figure 19.6: Correlation matrix showing Pearson's correlation coefficients among need for cognition, intelligence, social desirability, and dogmatism. The table reports the correlation for each pair of variables, with the strongest positive correlation between need for cognition and intelligence (r = .39) and negative correlations between dogmatism and both need for cognition (r = −.27) and intelligence (r = −.23). Dashes indicate correlations of variables with themselves.

19.4.2 Principal Components Analysis and Factor Analysis

When researchers collect measurements on many conceptually related variables, they often want to determine whether a smaller number of broader patterns can explain the relationships among those variables. Two techniques used for this purpose are **factor analysis** and **principal components analysis (PCA)**. Although the results of these techniques can look similar, they answer different questions. PCA asks, "How can we summarize the information contained in all these variables using fewer scores?" It combines the original variables into a smaller number of **components**. Each component is a weighted combination of the variables that preserves as much of the overall variation in the data as possible. In this context, *variation* refers to the ways in which participants' scores differ from one another. PCA considers all the variation in each variable, including variation shared with other variables, variation unique to that variable, and variation caused by measurement error. A component is therefore primarily a statistical summary of the observed variables; it does not necessarily represent an unobserved psychological characteristic.

Factor analysis asks a different question: "What underlying characteristics might explain why these variables are correlated?" It focuses on **shared variance**, meaning the variation that two or more variables have in common. When several measures are strongly correlated, people who score high on one of them also tend to score high on the others. Factor analysis attempts to explain this pattern by estimating a smaller number of unobserved variables called **factors** or **latent constructs**. A latent construct cannot be measured directly but is inferred from people's scores on observable measures. For example, extraversion cannot be observed in the same direct way as height or reaction time. Instead, researchers infer a person's level

of extraversion from answers to questions about sociability, activity, positive emotions, and related behaviors.

Both factor analysis and PCA usually produce a table of numbers called **loadings**. A loading indicates how strongly an observed variable is associated with a factor or component. A large positive loading means that participants with high scores on the variable also tend to have high scores on that factor or component. Variables that have large loadings on the same factor or component form a recognizable group. Because factor-analysis and PCA loading tables have a similar appearance—and because the two methods sometimes produce similar groupings—it can be easy to treat factors and components as interchangeable. Conceptually, however, they are different. A factor is interpreted as an underlying construct that helps explain the correlations among variables. A component is a mathematical combination created to summarize the observed variables efficiently.

Consider a simplified study in which participants complete several mathematical and verbal tasks. Scores on arithmetic, quantitative estimation, and spatial-reasoning tasks might be strongly correlated with one another. Scores on grammar, reading-comprehension, and vocabulary tasks might form another group of correlations. Factor analysis might represent these patterns with two factors that researchers interpret as mathematical ability and verbal ability. The factors are not directly observed. Instead, they are proposed explanations for why performance on tasks within each group is related.

The **Big Five personality factors** provide another example. Researchers began with scores on a large number of specific personality characteristics. Measures of warmth, gregariousness, activity level, and positive emotions tended to be correlated. Factor analysis grouped these related measures together, and researchers interpreted the underlying factor as **extraversion**. In this interpretation, extraversion is the latent construct that helps explain why people who score highly on one of these characteristics often score highly on the others.

PCA can also reveal useful patterns, but the interpretation is somewhat different. Peter Rentfrow and Samuel Gosling asked more than 1,700 university students to rate how much they liked 14 popular genres of music (Rentfrow & Gosling, 2003). They used PCA to summarize the 14 separate ratings with four broader components. Genres that tended to receive similar ratings were grouped within the same component. For example, students who liked blues often also liked jazz, classical music, and folk music, so these genres contributed to a component the researchers named **Reflective and Complex**. The other components were **Intense and Rebellious** (rock, alternative, and heavy metal), **Upbeat and Conventional** (country, soundtrack, religious, and pop), and **Energetic and Rhythmic** (rap/hip-hop, soul/funk, and electronica); see Table 19.2. These components provided a concise way to summarize patterns across the 14 observed music-preference ratings. Unlike factors in a factor-analysis model, however, the components should not automatically be interpreted as unobserved psychological traits that caused the ratings.

Table 19.2 *Component Loadings of the 14 Music Genres on Four Principal Components*[7]

Table 6.2 Factor Loadings of the 14 Music Genres on Four Varimax-Rotated Principal Components. Based on Research by Rentfrow and Gosling (2003)

	Music-preference dimension			
Genre	***Reflective and Complex***	***Intense and Rebellious***	***Upbeat and Conventional***	***Energetic and Rhythmic***
Blues	**.85**	.01	−.09	.12
Jazz	**.83**	.04	.07	.15
Classical	**.66**	.14	.02	−.13
Folk	**.64**	.09	.15	−.16
Rock	.17	**.85**	−.04	−.07
Alternative	.02	**.80**	.13	.04
Heavy metal	.07	**.75**	−.11	.04
Country	−.06	.05	**.72**	−.03
Sound tracks	.01	.04	**.70**	.17
Religious	.23	−.21	**.64**	−.01
Pop	−.20	.06	**.59**	.45
Rap/hip-hop	−.19	−.12	.17	**.79**
Soul/funk	.39	−.11	.11	**.69**
Electronica/dance	−.02	.15	−.01	**.60**

Note. $N = 1{,}704$. All factor loadings .40 or larger are in italics; the highest factor loadings for each dimension are listed in boldface type.

Figure 19.7: Table showing factor loadings for 14 music genres across four music-preference dimensions: Reflective and Complex, Intense and Rebellious, Upbeat and Conventional, and Energetic and Rhythmic. The table illustrates which genres load most strongly on each factor, with bold values indicating the strongest loading for each genre and italicized values representing loadings of .40 or greater.

Two additional points about factor analysis are worth making. First, factors are ***continuous dimensions***, not categories. Factor analysis does not imply that people are either extraverted or conscientious or that they like either "reflective and complex" music or "intense and rebellious" music. Scores on different factors may be correlated or relatively independent, depending on the constructs and the statistical method used. Thus, a person who is high in extraversion might be high or low in conscientiousness, and a person who likes reflective and complex music

might or might not also like intense and rebellious music. Second, factor analysis reveals a statistical structure that researchers must interpret and label; it does not by itself explain why that structure exists. The distinct Big Five personality factors likely reflect complex combinations of genetic and environmental influences rather than each trait being controlled by a different set of genes (Plomin et al., 2008).

19.4.3 Exploring Causal Relationships

Another important use of complex correlational research is to explore possible causal relationships among variables. This might seem surprising given the oft-quoted saying that "correlation does not imply causation." Correlational research cannot unambiguously establish that one variable causes another. Complex correlational research can, however, help researchers evaluate whether some plausible alternative explanations remain consistent with the observed data. One approach is ***statistical adjustment*** for measured potential third variables. Instead of controlling these variables through random assignment or by holding them constant as in an experiment, researchers measure them and include them in an analysis such as ***partial correlation***. This technique estimates the relationship between two variables after adjusting statistically for one or more measured covariates. Statistical adjustment does not provide the same protection against alternative explanations as experimental control.

For example, assume that a researcher is interested in the relationship between watching violent television shows and aggressive behavior but is concerned that socioeconomic status (SES) might represent a third variable related to both. The researcher could measure participants' violent television viewing, acts of aggression, and SES. Suppose the unadjusted correlation between violent television viewing and aggression is +.35, a moderate-sized positive correlation. The researcher could then use partial correlation to reexamine the association after statistically adjusting for SES. If the partial correlation is +.34, the association between violent television viewing and aggression changed little after adjustment for SES. If the partial correlation drops to +.03, the unadjusted association is substantially accounted for by its statistical relationship with SES. If the partial correlation is +.20, SES statistically accounts for some, but not all, of the association. Partial correlation can adjust only for measured variables included in the model. It cannot establish causality, resolve directionality, or eliminate bias from unmeasured or poorly measured third variables, measurement error, or a misspecified model.

19.4.4 Regression

Once a relationship between two variables has been established, researchers can use that information to make predictions about the value of one variable given the value of another variable. For instance, once we have established that there is a correlation between IQ and GPA, we can use people's IQ scores to predict their GPA. Thus, while correlation coefficients can be used to describe the strength and direction of relationships between variables, ***regression*** is a statistical technique that allows researchers to predict one variable given another. Regression

can also be used to describe more complex relationships between more than two variables. Typically, the variable that is used to make the prediction is referred to as the ***predictor variable*** and the variable that is being predicted is called the ***outcome variable or criterion variable.*** This regression equation has the following general form:

$\hat{Y} = b_0 + b_1X_1$

$\hat{Y}$ represents the predicted score on the outcome variable. The intercept, b_0, is the predicted value of Y when X_1 equals 0. The coefficient b_1 is the slope, or regression coefficient, and X_1 is the predictor score. An observed score can be written as $Y = b_0 + b_1X_1 + e$, where e is the residual—the difference between the observed and predicted scores.

While ***simple regression*** uses one variable to predict another, ***multiple regression*** uses several predictor variables (X_1, X_2, X_3, …, X_k) to predict or describe an outcome variable (Y). The multiple-regression equation expresses the predicted outcome as an additive combination of predictor values:

$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$

The regression coefficients (b_1, b_2, and so on) indicate how much the predicted outcome changes for a one-unit increase in a predictor while all other predictors are held constant. The intercept b_0 is the predicted outcome when all predictors equal 0.

The advantage of multiple regression is that it can show whether a predictor variable makes a contribution to an outcome variable *over and above* the contributions made by other predictor variables (i.e., it can be used to show whether a predictor variable is related to an outcome variable after statistically controlling for other predictor variables). As a hypothetical example, imagine that a researcher wants to know how income and health relate to happiness. This is tricky because income and health are themselves related to each other. Thus, if people with greater incomes tend to be happier, then perhaps this is only because they tend to be healthier. Likewise, if people who are healthier tend to be happier, perhaps this is only because they tend to make more money. But a multiple regression analysis including both income and health as predictor variables would show whether each one makes a contribution to the prediction of happiness when the other is taken into account (when it is statistically controlled). In other words, multiple regression would allow the researcher to examine whether that part of income that is unrelated to health predicts or relates to happiness as well as whether that part of health that is unrelated to income predicts or relates to happiness.

The examples discussed in this section only scratch the surface of how researchers use complex correlational research to explore possible causal relationships among variables. It is important to keep in mind, however, that purely correlational approaches cannot unambiguously establish that one variable causes another. The best they can do is show patterns of relationships that are consistent with some causal interpretations and inconsistent with others.

19.5 Media Attributions

[1-7] Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

19.6 Text Attributions

Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

19.7 References

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Jouriles, E. N., Garrido, E., Rosenfield, D., & McDonald, R. (2009). Experiences of psychological and physical aggression in adolescent romantic relationships: Links to psychological distress. *Child Abuse & Neglect, 33*(7), 451–460. https://doi.org/10.1016/j.chiabu.2008.11.005

Messerli, F. H. (2012). Chocolate consumption, cognitive function, and Nobel laureates. *New England Journal of Medicine, 367*(16), 1562–1564. https://doi.org/10.1056/NEJMon1211064

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral genetics* (5th ed.). Worth.

Radcliffe, N. M., & Klein, W. M. P. (2002). Dispositional, unrealistic, and comparative optimism: Differential relations with the knowledge and processing of risk information and beliefs about personal risk. *Personality and Social Psychology Bulletin, 28*(6), 836–846. https://doi.org/10.1177/0146167202289012

Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology, 84*(6), 1236–1256. https://doi.org/10.1037/0022-3514.84.6.1236

20 Chapter 20: Experimental Research

In the late 1960s, social psychologists John Darley and Bibb Latané proposed a counterintuitive hypothesis. The more witnesses there are to an accident or a crime, the less likely any of them is to help the victim (Darley & Latané, 1968).

So to test their hypothesis, Darley and Latané created a simulated emergency situation in a laboratory. Each of their university student participants was isolated in a small room and told that they would be having a discussion about university life with other students via an intercom system. Early in the discussion, however, one of the students began having what seemed to be an epileptic seizure. Over the intercom came the following: "I could really-er-use some help so if somebody would-er-give me a little h-help-uh-er-er-er-er-er c-could somebody-er-er-help-er-uh-uh-uh (choking sounds)…I'm gonna die-er-er-I'm…gonna die-er-help-er-er-seizure-er- [chokes, then quiet]" (Darley & Latané, 1968, p. 379).

In actuality, there were no other students. These comments had been prerecorded and were played back to create the appearance of a real emergency. The key to the study was that some participants were told that the discussion involved only one other student (the victim), others were told that it involved two other students, and still others were told that it included five other students. Because this was the only difference between these three groups of participants, any difference in their tendency to help the victim would have to have been caused by it. And sure enough, the likelihood that the participant left the room to seek help for the "victim" decreased from 85% to 62% to 31% as the number of "witnesses" increased.

The research that Darley and Latané conducted was a particular kind of study called an experiment. Experiments are used to determine not only whether there is a meaningful relationship between two variables but also whether the relationship is a causal one that is supported by statistical analysis. For this reason, experiments are one of the most common and useful tools in the psychological researcher's toolbox. In this chapter, we look at experiments in detail. We will first consider what sets experiments apart from other kinds of studies and why they support causal conclusions while other kinds of studies do not. We then look at two basic ways of designing an experiment—between-subjects designs and within-subjects designs—and discuss their pros and cons. Finally, we consider several important practical issues that arise when conducting experiments.

20.1 Experiment Basics

20.1.1 What Is an Experiment?

An ***experiment*** is a type of study designed specifically to answer the question of whether there is a causal relationship between two variables. In other words, whether changes in one variable (referred to as an ***independent variable***) cause a change in another variable (referred to as a ***dependent variable***). Experiments have two fundamental features. The first is that the researchers manipulate, or systematically vary, the level of the independent variable. The different levels of the independent variable are called ***conditions***. For example, in Darley and Latané's experiment, the independent variable was the number of witnesses that participants believed to be present. The researchers manipulated this independent variable by telling participants that there were either one, two, or five other students involved in the discussion, thereby creating three conditions. For a new researcher, it is easy to confuse these terms by believing there are three independent variables in this situation: one, two, or five students involved in the discussion, but there is actually only one independent variable (number of witnesses) with three different levels or conditions (one, two, or five students). The second fundamental feature of an experiment is that the researcher exerts ***control*** over, or minimizes the variability in, variables other than the independent and dependent variable. These other variables are called ***extraneous variables***. Darley and Latané tested all their participants in the same room, exposed them to the same emergency situation, and so on. They also randomly assigned their participants to conditions so that the three groups would be similar to each other to begin with. Notice that although the words manipulation and control have similar meanings in everyday language, researchers make a clear distinction between them. They manipulate the independent variable by systematically changing its levels and control other variables by holding them constant.

20.1.2 Manipulation of the Independent Variable

Again, to ***manipulate*** an independent variable means to change its level systematically so that different groups of participants are exposed to different levels of that variable, or the same group of participants is exposed to different levels at different times. For example, to see whether expressive writing affects people's health, a researcher might instruct some participants to write about traumatic experiences and others to write about neutral experiences. The different levels of the independent variable are referred to as conditions, and researchers often give the conditions short descriptive names to make it easy to talk and write about them. In this case, the conditions might be called the "traumatic condition" and the "neutral condition."

Notice that the manipulation of an independent variable must involve the active intervention of the researcher. Comparing groups of people who differ on the independent variable before the study begins is not the same as manipulating that variable. For example, a researcher who compares the health of people who already keep a journal with the health of people who

do not keep a journal has not manipulated this variable and therefore has not conducted an experiment. This distinction is important because groups that already differ in one way at the beginning of a study are likely to differ in other ways too. For example, people who choose to keep journals might also be more conscientious, more introverted, or less stressed than people who do not. Therefore, any observed difference between the two groups in terms of their health might have been caused by whether or not they keep a journal, or it might have been caused by any of the other differences between people who do and do not keep journals. Thus, the active manipulation of the independent variable is crucial for eliminating potential alternative explanations for the results.

Of course, there are many situations in which the independent variable cannot be manipulated for practical or ethical reasons and therefore an experiment is not possible. For example, whether or not people have a significant early illness experience cannot be manipulated, making it impossible to conduct an experiment on the effect of early illness experiences on the development of health anxiety. This caveat does not mean it is impossible to study the relationship between early illness experiences and health anxiety—only that it must be done using nonexperimental approaches.

Independent variables can be manipulated to create two conditions. Experiments involving one independent variable with two conditions are often described as ***single-factor, two-level designs***. Sometimes, however, researchers gain greater insight by adding conditions. An experiment with one independent variable manipulated to produce more than two conditions is a ***single-factor, multilevel design***. Rather than comparing only a one-witness condition with a five-witness condition, Darley and Latané used a single-factor, multilevel design with one-witness, two-witness, and five-witness conditions.

20.1.3 Control of Extraneous Variables

An extraneous variable is anything that varies in the context of a study other than the independent and dependent variables. In an experiment on the effect of expressive writing on health, for example, extraneous variables would include participant variables (individual differences) such as their writing ability, their diet, and their gender. They would also include situational or task variables such as the time of day when participants write, whether they write by hand or on a computer, and the weather. Extraneous variables pose a problem because many of them are likely to have some effect on the dependent variable. For example, participants' health will be affected by many things other than whether or not they engage in expressive writing. This influencing factor can make it difficult to separate the effect of the independent variable from the effects of the extraneous variables, which is why it is important to control extraneous variables by holding them constant.

20.1.3.1 Extraneous Variables as "Noise"

Extraneous variables make it difficult to detect the effect of the independent variable in two ways. One is by adding variability or "noise" to the data. Imagine a simple experiment on the effect of mood (happy vs. sad) on the number of happy childhood events people are able to recall. Participants are put into a negative or positive mood (by showing them a happy or sad video clip) and then asked to recall as many happy childhood events as they can. The two leftmost columns of Table 20.1 show what the data might look like if there were no extraneous variables and the number of happy childhood events participants recalled was affected only by their moods. Every participant in the happy mood condition recalled exactly four happy childhood events, and every participant in the sad mood condition recalled exactly three. The effect of mood here is quite obvious. In reality, however, the data would probably look more like those in the two rightmost columns of Table 20.1. Even in the happy mood condition, some participants would recall fewer happy memories because they have fewer to draw on, use less effective recall strategies, or are less motivated. And even in the sad mood condition, some participants would recall more happy childhood memories because they have more happy memories to draw on, they use more effective recall strategies, or they are more motivated. Although the mean difference between the two groups is the same as in the idealized data, this difference is much less obvious in the context of the greater variability in the data. Thus, one reason researchers try to control extraneous variables is so their data look more like the idealized data in Table 20.1, which makes the effect of the independent variable easier to detect (although real data never look quite *that* good).

Table 20.1 *Hypothetical Noiseless Data Compared With Realistic Noisy Data*[1]

Idealized "noiseless" data		Realistic "noisy" data	
Happy mood	**Sad mood**	**Happy mood**	**Sad mood**
4	3	3	1
4	3	6	3
4	3	2	4
4	3	4	0
4	3	5	5
4	3	2	7
4	3	3	2
4	3	1	5
4	3	6	1
4	3	8	2
M = 4	M = 3	M = 4	M = 3

Table 5.1 Hypothetical Noiseless Data and Realistic Noisy Data

Figure 20.1: Table comparing hypothetical noiseless and realistic noisy recall data for happy- and sad-mood conditions. In both versions, the happy-mood mean is 4 and the sad-mood mean is 3, but the realistic data vary across participants.

One way to control extraneous variables is to hold them constant. This technique can mean holding situation or task variables constant by testing all participants in the same location, giving them identical instructions, treating them in the same way, and so on. It can also mean holding participant variables constant. For example, many studies of language limit participants to right-handed people, who generally have their language areas isolated in their left cerebral hemispheres. Left-handed people are more likely to have their language areas isolated in their right cerebral hemispheres or distributed across both hemispheres, which can change the way they process language and thereby add noise to the data.

In principle, researchers can reduce participant-related variability by limiting a sample to people who share selected characteristics, such as a particular age range, educational background, handedness, or first language. The obvious downside is that this approach can reduce ***external validity***—especially the extent to which the results generalize beyond the people actually studied. Results obtained from a narrowly defined group may not apply to people with different ages, backgrounds, identities, or life experiences. In many situations, the increased external validity of a diverse sample outweighs the reduction in noise achieved by a homogeneous one.

20.1.3.2 Extraneous Variables as Confounding Variables

The second way that extraneous variables can make it difficult to detect the effect of the independent variable is by becoming confounding variables. A ***confounding variable*** is an extraneous variable that differs on average *across* levels of the independent variable (i.e., it is an extraneous variable that varies systematically with the independent variable). For example, in almost all experiments, participants' intelligence quotients (IQs) will be an extraneous variable. But as long as there are participants with lower and higher IQs in each condition so that the average IQ is roughly equal across the conditions, then this variation is probably acceptable (and may even be desirable). What would be bad, however, would be for participants in one condition to have substantially lower IQs on average and participants in another condition to have substantially higher IQs on average. In this case, IQ would be a confounding variable.

To confound means to confuse, and this effect is exactly why confounding variables are undesirable. Because they differ systematically across conditions—just like the independent variable—they provide an alternative explanation for any observed difference in the dependent variable. Figure 20.1 shows the results of a hypothetical study, in which participants in a positive mood condition scored higher on a memory task than participants in a negative mood condition. But if IQ is a confounding variable—with participants in the positive mood condition having higher IQs on average than participants in the negative mood condition—then it is unclear whether it was the positive moods or the higher IQs that caused participants in the first condition to score higher. One way to avoid confounding variables is by holding extraneous variables constant. For example, one could prevent IQ from becoming a confounding variable by limiting participants only to those with IQs of exactly 100. But this approach is not always desirable because it sharply restricts the sample. A second and much more general approach is random assignment to conditions.

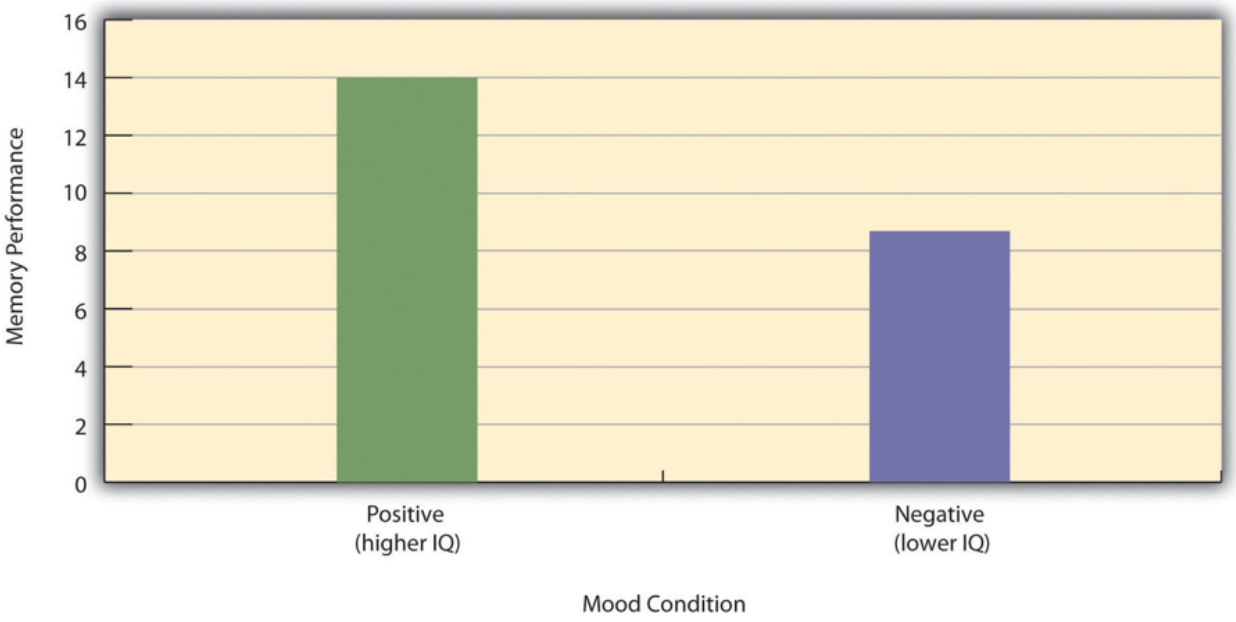


Figure 20.2: Bar graph showing hypothetical memory performance for two mood conditions. The positive mood group (higher IQ) has higher memory performance (about 14) than the negative mood group (lower IQ), which scores about 9. Because IQ differs between groups, the graph illustrates IQ as a confounding variable that could explain the observed difference in memory performance rather than mood alone.

Figure 20.1 *Hypothetical Results From a Study on the Effect of Mood on Memory*[2]

20.1.4 Treatment and Control Conditions

In psychological research, a ***treatment*** is any intervention meant to change people's behavior for the better. This intervention includes psychotherapies and medical treatments for psychological disorders but also interventions designed to improve learning, promote conservation, reduce prejudice, and so on. To determine whether a treatment works, participants are randomly assigned to either a ***treatment condition***, in which they receive the treatment, or a ***control condition***, in which they do not receive the treatment. If participants in the treatment condition end up better off than participants in the control condition—for example, they are less depressed, learn faster, conserve more, express less prejudice—then the researcher can conclude that the treatment works. In research on the effectiveness of psychotherapies and medical treatments, this type of experiment is often called a ***randomized clinical trial***.

There are different types of control conditions. In a ***no-treatment control condition***, participants receive no treatment whatsoever. One problem with this approach, however, is the

existence of placebo effects. A ***placebo*** is a simulated treatment that lacks any active ingredient or element that should make it effective, and a ***placebo effect*** is a positive effect of such a treatment. Many folk remedies that seem to work—such as eating chicken soup for a cold or placing soap under the bed sheets to stop nighttime leg cramps—are probably nothing more than placebos. Although placebo effects are not well understood, they are probably driven primarily by people's expectations that they will improve. Having the expectation to improve can result in reduced stress, anxiety, and depression, which can alter perceptions and even improve immune system functioning (Price et al., 2008).

Placebo effects are interesting in their own right (see Note "The Powerful Placebo"), but they also pose a serious problem for researchers who want to determine whether a treatment works. Figure 20.2 shows some hypothetical results in which participants in a treatment condition improved more on average than participants in a no-treatment control condition. If these conditions (the two leftmost bars in Figure 20.2) were the only conditions in this experiment, however, one could not conclude that the treatment worked. It could be instead that participants in the treatment group improved more because they expected to improve, while those in the no-treatment control condition did not.

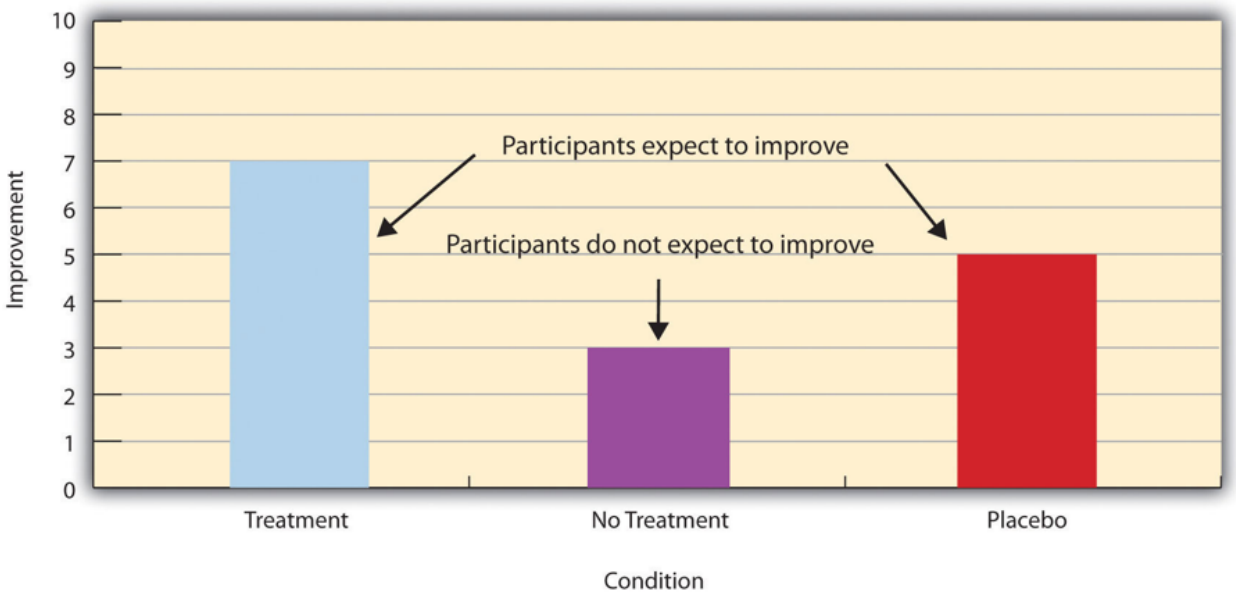


Figure 20.3: Bar graph comparing improvement across three conditions: treatment, no treatment, and placebo. The treatment group shows the greatest improvement (about 7), the no-treatment group shows the least improvement (about 3), and the placebo group shows moderate improvement (about 5). Arrows indicate that participants in the treatment and placebo groups expected to improve, whereas participants in the no-treatment group did not, illustrating how placebo effects can influence outcomes.

Figure 20.2 *Hypothetical Results From a Study Including Treatment, No-Treatment, and Placebo Conditions*[3]

Fortunately, there are several solutions to this problem. One is to include a ***placebo control condition***, in which participants receive a placebo that looks much like the treatment but lacks the active ingredient or element thought to be responsible for the treatment's effectiveness. When participants in a treatment condition take a pill, for example, then those in a placebo control condition would take an identical-looking pill that lacks the active ingredient in the treatment (a "sugar pill"). In research on psychotherapy effectiveness, the placebo might involve going to a psychotherapist and talking in an unstructured way about one's problems. The idea is that if participants in both the treatment and the placebo control groups expect to improve, then any improvement in the treatment group over and above that in the placebo control group must have been caused by the treatment and not by participants' expectations. This difference is what is shown by a comparison of the two outer bars in Figure 20.2.

Of course, the principle of informed consent requires that participants be told that they will be assigned to either a treatment or a placebo control condition—even though they cannot be told which until the experiment ends. In many cases, the participants who had been in the control condition are then offered an opportunity to have the real treatment. An alternative approach is to use a ***wait-list control condition***, in which participants are told that they will receive the treatment but must wait until the participants in the treatment condition have already received it. This disclosure allows researchers to compare participants who have received the treatment with participants who are not currently receiving it but who still expect to improve (eventually). A final solution to the problem of placebo effects is to leave out the control condition completely and compare any new treatment with the best available alternative treatment. For example, a new treatment for simple phobia could be compared with standard exposure therapy. Because participants in both conditions receive a treatment, their expectations about improvement should be similar. This approach also makes sense because once there is an effective treatment, the interesting question about a new treatment is not simply "Does it work?" but "Does it work better than what is already available?"

20.1.5 The Powerful Placebo

Many people are not surprised that placebos can have a positive effect on disorders that seem fundamentally psychological, including depression, anxiety, and insomnia. However, placebos can also have a positive effect on disorders that most people think of as fundamentally physiological. These include asthma, ulcers, and warts (Shapiro & Shapiro, 1999). There is even evidence that placebo surgery—also called "sham surgery"—can be as effective as actual surgery.

Medical researcher J. Bruce Moseley and his colleagues conducted a study on the effectiveness of two arthroscopic surgery procedures for osteoarthritis of the knee (Moseley et al., 2002). The control participants in this study were prepped for surgery, received a tranquilizer, and even

received three small incisions in their knees. But they did not receive the actual arthroscopic surgical procedure. Note that the IRB would have carefully considered the use of deception in this case and judged that the benefits of using it outweighed the risks and that there was no other way to answer the research question (about the effectiveness of a placebo procedure) without it. The surprising result was that all participants improved in terms of both knee pain and function, and the sham surgery group improved just as much as the treatment groups. According to the researchers, "This study provides strong evidence that arthroscopic lavage with or without débridement [the surgical procedures used] is not better than and appears to be equivalent to a placebo procedure in improving knee pain and self-reported function" (p. 85).

20.2 Experimental Design

In this section, we look at some different ways to design an experiment. The primary distinction we will make is between approaches in which each participant experiences one level of the independent variable and approaches in which each participant experiences all levels of the independent variable. The former are called between-subjects experiments and the latter are called within-subjects experiments.

20.2.1 Between-Subjects Experiments

In a ***between-subjects experiment***, each participant is tested in only one condition. For example, a researcher with a sample of 100 university students might assign half of them to write about a traumatic event and the other half write about a neutral event. Or a researcher with a sample of 60 people with severe agoraphobia (fear of open spaces) might assign 20 of them to receive each of three different treatments for that disorder. It is essential in a between-subjects experiment that the researcher assigns participants to conditions so that the different groups are, on average, highly similar to each other. Those in a trauma condition and a neutral condition, for example, should include a similar proportion of men and women, and they should have similar average IQs, similar average levels of motivation, similar average numbers of health problems, and so on. This matching is a matter of controlling these extraneous participant variables across conditions so that they do not become confounding variables.

20.2.1.1 Random Assignment

The primary way that researchers accomplish this kind of control of extraneous variables across conditions is called ***random assignment***, which means using a random process to decide which participants are tested in which conditions. Do not confuse random assignment with random sampling. Random sampling is a method for selecting a sample from a population, and it is rarely used in psychological research. Random assignment is a method for assigning

participants in a sample to the different conditions, and it is an important element of all experimental research in psychology and other fields too.

In its strictest sense, random assignment should meet two criteria. One is that each participant has an equal chance of being assigned to each condition (e.g., a 50% chance of being assigned to each of two conditions). The second is that each participant is assigned to a condition independently of other participants. Thus, one way to assign participants to two conditions would be to flip a coin for each one. If the coin lands heads, the participant is assigned to Condition A, and if it lands tails, the participant is assigned to Condition B. For three conditions, one could use a computer to generate a random integer from 1 to 3 for each participant. If the integer is 1, the participant is assigned to Condition A; if it is 2, the participant is assigned to Condition B; and if it is 3, the participant is assigned to Condition C. In practice, a full sequence of conditions—one for each participant expected to be in the experiment—is usually created ahead of time, and each new participant is assigned to the next condition in the sequence as they are tested. When the procedure is computerized, the computer program often handles the random assignment.

One problem with coin flipping and other strict procedures for random assignment is that they are likely to result in unequal sample sizes in the different conditions. Unequal sample sizes are generally not a serious problem, and you should never throw away data you have already collected to achieve equal sample sizes. However, for a fixed number of participants, it is statistically most efficient to divide them into equal-sized groups. It is standard practice, therefore, to use a kind of modified random assignment that keeps the number of participants in each group as similar as possible. One approach is ***block randomization***. In block randomization, all the conditions occur once in the sequence before any of them is repeated. Then they all occur again before any of them is repeated again. Within each of these "blocks," the conditions occur in a random order. Again, the sequence of conditions is usually generated before any participants are tested, and each new participant is assigned to the next condition in the sequence. Table 20.2 shows such a sequence for assigning nine participants to three conditions. Research Randomizer (website) will generate block randomization sequences for any number of participants and conditions. Again, when the procedure is computerized, the computer program often handles the block randomization.

Table 20.2 *Block Randomization Sequence for Assigning Nine Participants to Three Conditions*[4]

Participant	Condition
1	**A**
2	**C**
3	**B**
4	B
5	C
6	A
7	**C**
8	**B**
9	**A**

Table 5.2 Block Randomization Sequence for Assigning Nine Participants to Three Conditions

Figure 20.4: Table showing a block randomization sequence assigning nine participants to three experimental conditions. Participants 1–9 are assigned as follows: A, C, B, B, C, A, C, B, and A, demonstrating balanced random assignment across the three conditions.

Random assignment is not guaranteed to control all extraneous variables across conditions. The process is random, so it is always possible that just by chance, the participants in one condition might turn out to be substantially older, less tired, more motivated, or less depressed on average than the participants in another condition. However, there are some reasons that this possibility is not a major concern. One is that random assignment works better than one might expect, especially for large samples. Another is that the inferential statistics that researchers use to decide whether a difference between groups reflects a difference in the population takes the "fallibility" of random assignment into account. Yet another reason is that even if random assignment does result in a confounding variable and therefore produces misleading results, this confound is likely to be detected when the experiment is replicated. The upshot is that random assignment to conditions—although not infallible in terms of controlling extraneous variables—is always considered a strength of a research design.

20.2.2 Matched Groups

An alternative to simple random assignment of participants to conditions is the use of a ***matched-groups design***. Using this design, participants in the various conditions are

matched on the dependent variable or on some extraneous variable(s) prior to the manipulation of the independent variable. This guarantees that these variables will not be confounded across the experimental conditions. For instance, if we want to determine whether expressive writing affects people's health then we could start by measuring various health-related variables in our prospective research participants. We could then use that information to rank-order participants according to how healthy or unhealthy they are. Next, the two healthiest participants would be randomly assigned to complete different conditions (one would be randomly assigned to the traumatic experiences writing condition and the other to the neutral writing condition). The next two healthiest participants would then be randomly assigned to complete different conditions, and so on until the two least healthy participants. This method would ensure that participants in the traumatic experiences writing condition are matched to participants in the neutral writing condition with respect to health at the beginning of the study. If at the end of the experiment, a difference in health was detected across the two conditions, then we would know that it is due to the writing manipulation and not to pre-existing differences in health.

20.2.3 Within-Subjects Experiments

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

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

20.2.3.1 Carryover Effects and Counterbalancing

The primary disadvantage of within-subjects designs is that they can result in order effects. An ***order effect*** occurs when participants' responses in the various conditions are affected by the order of conditions to which they were exposed. One type of order effect is a carryover effect. A ***carryover effect*** is an effect of being tested in one condition on participants' behavior in later conditions. One type of carryover effect is a ***practice effect***, where participants perform a task better in later conditions because they have had a chance to practice it. Another type is a ***fatigue effect***, where participants perform a task worse in later conditions because they

become tired or bored. Being tested in one condition can also change how participants perceive stimuli or interpret their task in later conditions. This type of effect is called a ***context effect (or contrast effect)***. For example, an average-looking defendant might be judged more harshly when participants have just judged an attractive defendant than when they have just judged an unattractive defendant. Within-subjects experiments also make it easier for participants to guess the hypothesis. For example, a participant who is asked to judge the guilt of an attractive defendant and then is asked to judge the guilt of an unattractive defendant is likely to guess that the hypothesis is that defendant attractiveness affects judgments of guilt. This knowledge could lead the participant to judge the unattractive defendant more harshly because they think this is what they are expected to do. Or it could make participants judge the two defendants similarly in an effort to be "fair."

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

A common way to address order effects is ***counterbalancing***, in which different participants complete the conditions in different orders. With ***complete counterbalancing***, an equal number of participants completes each possible order. For two conditions, half of the participants would complete the attractive-defendant condition followed by the unattractive-defendant condition, and the other half would complete the reverse order. Three conditions produce six possible orders (ABC, ACB, BAC, BCA, CAB, and CBA); four produce 24; and five produce 120. Participants are randomly assigned to these orders. Thus, in a within-subjects design, randomization applies to the order of conditions rather than to condition membership, and counterbalancing distributes potential order effects across conditions.

A more efficient form of partial counterbalancing uses a ***Latin square***. A Latin square selects a limited set of orders so that each condition appears equally often in each ordinal position. For four conditions, the design requires four orders, represented by the rows of a 4×4 square. No condition repeats within a row, and each condition appears once in every column.

Table 20.3 *Latin Square Design for Four Versions of Four Treatments*[5]

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

Figure 20.5: Diagram illustrating a 4 × 4 square design for counterbalancing four treatment conditions labeled A, B, C, and D. The rows are: A B C D; B C D A; C D A B; and D A B C. Each condition appears exactly once in every row and column, demonstrating balanced counterbalancing across four conditions.

Table 20.3 shows that each condition appears once at each ordinal position: A appears first once, second once, third once, and fourth once, and the same is true for B, C, and D. This positional balance is the defining feature of a Latin square; the simple cyclic square shown here does not also balance every immediate predecessor-successor pair. A Latin square for an experiment with 6 conditions would be 6 × 6, one for an experiment with 8 conditions would be 8 × 8, and so on. Complete counterbalancing of 6 conditions would require 720 orders, whereas a Latin square would require only 6.

When the number of conditions is large, researchers can instead use ***random counterbalancing***, in which a condition order is generated randomly for each participant. This approach does not require researchers to list every possible order in advance. Because a particular sample may not represent all ordinal positions or adjacent condition pairs evenly, random counterbalancing generally controls order effects less precisely than complete counterbalancing or a carefully chosen Latin square. It can nevertheless be practical when there are many conditions and order effects are expected to be small.

There are two ways to think about what counterbalancing accomplishes. One is that it controls the order of conditions so that it is no longer a confounding variable. Instead of the attractive condition always being first and the unattractive condition always being second, the attractive condition comes first for some participants and second for others. Likewise, the unattractive condition comes first for some participants and second for others. Thus, any overall difference in the dependent variable between the two conditions cannot have been caused by the order of conditions. A second way to think about what counterbalancing accomplishes is that if there are carryover effects, it makes it possible to detect them. One can analyze the data separately for each order to see whether it had an effect.

20.2.3.2 When 9 Is "Larger" Than 221

Researcher Michael Birnbaum has argued that the lack of context provided by between-subjects designs is often a bigger problem than the context effects created by within-subjects designs. To demonstrate this problem, he asked participants to rate two numbers on how large they were on a scale of 1-to-10 where 1 was "very, very small" and 10 was "very, very large". One group of participants was asked to rate the number 9 and another group was asked to rate the number 221 (Birnbaum, 1999). Participants in this between-subjects design gave the number 9 a mean rating of 5.13 and the number 221 a mean rating of 3.10. In other words, they rated 9 as larger than 221! According to Birnbaum, this difference is because participants spontaneously compared 9 with other one-digit numbers (in which case it is *relatively* large) and compared 221 with other three-digit numbers (in which case it is *relatively* small).

20.2.3.3 Simultaneous Within-Subjects Designs

So far, we have discussed an approach to within-subjects designs in which participants are tested in one condition at a time. There is another approach, however, that is often used when participants make multiple responses in each condition. Imagine, for example, that participants judge the guilt of 10 attractive defendants and 10 unattractive defendants. Instead of having people make judgments about all 10 defendants of one type followed by all 10 defendants of the other type, the researcher could present all 20 defendants in a sequence that mixed the two types. The researcher could then compute each participant's mean rating for each type of defendant. Or imagine an experiment designed to see whether people with social anxiety disorder remember negative adjectives (e.g., "stupid," "incompetent") better than positive ones (e.g., "happy," "productive"). The researcher could have participants study a single list that includes both kinds of words and then have them try to recall as many words as possible. The researcher could then count the number of each type of word that was recalled.

20.2.4 Between-Subjects or Within-Subjects?

Almost every experiment can be conducted using either a between-subjects design or a within-subjects design. This possibility means that researchers must choose between the two approaches based on their relative merits for the particular situation.

Between-subjects experiments have the advantage of being conceptually simpler and requiring less testing time per participant. They also avoid carryover effects without the need for counterbalancing. Within-subjects experiments have the advantage of controlling extraneous participant variables, which generally reduces noise in the data and makes it easier to detect any effect of the independent variable upon the dependent variable. Within-subjects experiments also require fewer participants than between-subjects experiments to detect an effect of the same size.

A good rule of thumb, then, is that if it is possible to conduct a within-subjects experiment (with proper counterbalancing) in the time that is available per participant—and you have no serious concerns about carryover effects—this design is probably the best option. If a within-subjects design would be difficult or impossible to carry out, then you should consider a between-subjects design instead. For example, if you were testing participants in a doctor's waiting room or shoppers in line at a grocery store, you might not have enough time to test each participant in all conditions and therefore would opt for a between-subjects design. Or imagine you were trying to reduce people's level of prejudice by having them interact with someone of another race. A within-subjects design with counterbalancing would require testing some participants in the treatment condition first and then in a control condition. But if the treatment works and reduces people's level of prejudice, then they would no longer be suitable for testing in the control condition. This difficulty is true for many designs that involve a treatment meant to produce long-term change in participants' behavior (e.g., studies testing the effectiveness of psychotherapy). Clearly, a between-subjects design would be necessary here.

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

20.3 Experimentation and Validity

20.3.1 Four Big Validities

When we read about psychology experiments with a critical view, one question to ask is "is this study valid (accurate)?" However, that question is not as straightforward as it seems because, in psychology, there are many different kinds of validities. Researchers have focused on four validities to help assess whether an experiment is sound (Judd & Kenny, 1981; Morling, 2014): internal validity, external validity, construct validity, and statistical validity. We will explore each validity in depth.

20.3.2 Internal Validity

Two variables being statistically related does not necessarily mean that one causes the other. In your psychology education, you have probably heard the term, "Correlation does not imply causation." For example, if it were the case that people who exercise regularly are happier than people who do not exercise regularly, this implication would not necessarily mean that exercising increases people's happiness. It could mean instead that greater happiness causes people to exercise or that something like better physical health causes people to exercise *and* be happier.

The purpose of an experiment, however, is to show that two variables are statistically related and to do so in a way that supports the conclusion that the independent variable caused any observed differences in the dependent variable. The logic is based on this assumption: If the researcher creates two or more highly similar conditions and then manipulates the independent variable to produce just one difference between them, then any later difference between the conditions must have been caused by the independent variable. For example, because the only difference between Darley and Latané's conditions was the number of students that participants believed to be involved in the discussion, this difference in belief must have been responsible for differences in helping between the conditions.

An empirical study is said to be high in ***internal validity*** if the way it was conducted supports the conclusion that the independent variable caused any observed differences in the dependent variable. Thus, experiments are high in internal validity because the way they are conducted—with the manipulation of the independent variable and the control of extraneous variables (such as through the use of random assignment to minimize confounds)—provides strong support for causal conclusions. In contrast, non-experimental research designs (e.g., correlational designs), in which variables are measured but are not manipulated by an experimenter, are low in internal validity.

20.3.3 External Validity

At the same time, the way that experiments are conducted sometimes leads to a different kind of criticism. Specifically, the need to manipulate the independent variable and control extraneous variables means that experiments are often conducted under conditions that seem artificial (Bauman et al., 2014). In many psychology experiments, the participants are all undergraduate students and come to a classroom or laboratory to fill out a series of paper-and-pencil questionnaires or to perform a carefully designed computerized task. Consider, for example, an experiment in which researcher Barbara Fredrickson and her colleagues had undergraduate students come to a laboratory on campus and complete a math test while wearing a swimsuit (Fredrickson et al., 1998). At first, this manipulation might seem silly. When will undergraduate students ever have to complete math tests in their swimsuits outside of this experiment?

The issue we are confronting is that of ***external validity***. An empirical study is high in external validity if the way it was conducted supports generalizing the results to people and situations beyond those actually studied. As a general rule, studies are higher in external validity when the participants and the situation studied are similar to those that the researchers want to generalize to and participants encounter every day, often described as ***mundane realism***. Imagine, for example, that a group of researchers is interested in how shoppers in large grocery stores are affected by whether breakfast cereal is packaged in yellow or purple boxes. Their study would be high in external validity and have high mundane realism if they studied the decisions of ordinary people doing their weekly shopping in a real grocery store. If the shoppers bought much more cereal in purple boxes, the researchers would be fairly

confident that this increase would be true for other shoppers in other stores. Their study would be relatively low in external validity, however, if they studied a sample of undergraduate students in a laboratory at a selective university who merely judged the appeal of various colors presented on a computer screen; however, this study would have high ***psychological realism*** where the same mental process is used in both the laboratory and in the real world. If the students judged purple to be more appealing than yellow, the researchers would not be very confident that this preference is relevant to grocery shoppers' cereal-buying decisions because of low external validity, but they could be confident that the visual processing of colors has high psychological realism.

We should be careful, however, not to draw the blanket conclusion that experiments are low in external validity. One reason is that experiments need not seem artificial. Consider that Darley and Latané's experiment provided a reasonably good simulation of a real emergency situation. Or consider field experiments that are conducted entirely outside the laboratory. In one such experiment, Robert Cialdini and his colleagues studied whether hotel guests choose to reuse their towels for a second day as opposed to having them washed as a way of conserving water and energy (Cialdini, 2005). These researchers manipulated the message on a card left in a large sample of hotel rooms. One version of the message emphasized showing respect for the environment, another emphasized that the hotel would donate a portion of their savings to an environmental cause, and a third emphasized that most hotel guests choose to reuse their towels. The result was that guests who received the message that most hotel guests choose to reuse their towels, reused their own towels substantially more often than guests receiving either of the other two messages. Given the way they conducted their study, it seems very likely that their result would hold true for other guests in other hotels.

A second reason not to draw the blanket conclusion that experiments are low in external validity is that they are often conducted to learn about psychological processes that are likely to operate in a variety of people and situations. Let us return to the experiment by Fredrickson and colleagues. They found that the women in their study, but not the men, performed worse on the math test when they were wearing swimsuits. They argued that this gender difference was due to women's greater tendency to objectify themselves—to think about themselves from the perspective of an outside observer—which diverts their attention away from other tasks. They argued, furthermore, that this process of self-objectification and its effect on attention is likely to operate in a variety of women and situations—even if none of them ever finds herself taking a math test in her swimsuit.

20.3.4 Construct Validity

In addition to considering whether results generalize, researchers should evaluate the ***construct validity*** of an experiment—the extent to which its manipulations and measures adequately represent the intended constructs. Darley and Latané asked whether diffusion of responsibility reduces helping behavior. They hypothesized that participants would be less likely to help when they believed that more potential helpers were present. Translating an abstract construct into

a concrete procedure is called ***operationalization***. The researchers operationalized diffusion of responsibility by varying the number of potential helpers and operationalized helping by recording whether participants sought assistance. The procedure had strong construct validity because it created an apparent crisis, gave participants an opportunity to help, and systematically varied the number of other people believed to be available.

The number and range of conditions can also affect construct validity. With only two conditions—one other student or two—reduced helping might reflect the mere presence of another potential helper rather than a graded diffusion of responsibility. Adding more levels could reveal whether helping continues to decline or reaches a plateau. More conditions do not automatically improve construct validity, however; each level must contribute meaningfully to the intended test. When designing an experiment, consider how well the constructs are operationalized in your study.

20.3.5 Statistical Validity

Statistical conclusion validity concerns whether the analyses are appropriate and whether the statistical evidence supports the conclusions drawn from the data. Researchers can choose among many inferential statistical tests, including t tests, analyses of variance, regression models, and correlation coefficients. The appropriate analysis depends on the research design, the scale on which the dependent variable was measured, and the statistical assumptions of the method. Statistical conclusion validity is threatened when assumptions are violated and the violations are not addressed through a more suitable analysis, transformation, robust method, or transparent qualification of the conclusion.

One common critique of experiments is that a study did not have enough participants. The main reason for this criticism is that it is difficult to generalize about a population from a small sample. At the outset, it seems as though this critique is about external validity but there are studies where small sample sizes are not a problem. Therefore, small sample sizes are actually a critique of statistical validity. Statistical conclusion validity concerns whether the analyses are sound and support the conclusions that are drawn.

The proper statistical analysis should be conducted on the data to determine whether the difference or relationship that was predicted was indeed found. Interestingly, the likelihood of detecting an effect of the independent variable on the dependent variable depends on not just whether a relationship really exists between these variables, but also the number of conditions and the size of the sample. This is why it is important to conduct a power analysis when designing a study, which is a calculation that informs you of the number of participants you need to recruit to detect an effect of a specific size.

20.4 Prioritizing Validities

These four big validities–internal, external, construct, and statistical–are useful to keep in mind when both reading about other experiments and designing your own. However, researchers must prioritize and often it is not possible to have high validity in all four areas. In Cialdini's study on towel usage in hotels, the external validity was high but the statistical validity was more modest. This discrepancy does not invalidate the study but it shows where there may be room for improvement for future follow-up studies (Goldstein et al., 2008). Morling (2014) points out that many psychology studies have high internal and construct validity but sometimes sacrifice external validity.

20.5 Practical Considerations

The information presented so far in this chapter is enough to design a basic experiment. When it comes time to conduct that experiment, however, several additional practical issues arise. In this section, we consider some of these issues and how to deal with them. Much of this information applies to non-experimental studies as well as experimental ones.

20.5.1 Recruiting Participants

Researchers should consider recruitment feasibility at the beginning of a project. Some questions require ethical and practical access to a specific population, such as people living with schizophrenia or justice-involved adolescents. A study should not be designed around a population that the research team cannot appropriately reach, support, and protect. Even when researchers plan to use a convenience sample, they still need a clear recruitment strategy.

One approach is to recruit from a formal ***participant pool***—an established group of people who have agreed to receive information about research opportunities. At many colleges and universities, for example, introductory psychology students may participate in studies for course credit through an online system. When participation is tied to a course requirement, students should have a reasonable nonresearch alternative for earning equivalent credit. Researchers can also use advertisements, community partnerships, registries, or direct outreach to organizations serving the population of interest. Recruitment materials should describe the study accurately, avoid undue influence, and make clear that deciding not to participate will not result in a penalty or loss of benefits.

20.5.1.1 Volunteer Samples

Most research participants volunteer, even when they receive course credit, payment, or access to an intervention. This self-selection can produce ***volunteer bias*** when people who choose to

participate differ systematically from those who do not. Older research, including Rosenthal and Rosnow (1976), identified several possible differences, but these patterns depend on the topic, population, recruitment method, and incentive and should not be treated as fixed characteristics of all volunteers. Depending on the study, volunteers may:

- Have greater interest in the research topic.
- Be more familiar or comfortable with research settings.
- Have schedules, transportation, technology, or other resources that make participation easier.
- Respond differently to the compensation or course-credit structure.
- Be more willing to disclose information or interact with researchers.
- Differ demographically or socioeconomically from the target population.

Volunteer bias can limit ***external validity*** when characteristics related to volunteering also affect the behavior being studied. For example, people who volunteer for persuasion research may be more interested in the topic or more comfortable with research procedures than members of the broader target population. Researchers should describe how participants were recruited, compare the sample with the target population when possible, and avoid generalizing beyond the evidence provided by the sample.

In some ***field experiments***, researchers identify eligible participants in public settings rather than actively recruiting them. Guéguen and de Gail (2003), for example, studied whether being smiled at affected helping among supermarket shoppers. A confederate walking down a stairway looked directly at a shopper walking upward and either smiled or did not smile. Shortly afterward, the shopper encountered another confederate who dropped computer diskettes. The dependent variable was whether the shopper helped pick them up. An institutional review board would need to determine whether any waiver or alteration of informed consent met the applicable ethical requirements. Researchers would also need a prespecified selection rule to reduce bias. In the original study, the confederate approached the first person encountered who appeared to be between 20 and 50 years old, and only people who returned the gaze were included. Contemporary researchers should justify observable eligibility criteria, consider the risk of misclassification, and explain how the selection procedure may affect the sample and the generalizability of the results.

20.5.2 Standardizing the Procedure

It is surprisingly easy to introduce extraneous variables during the procedure. For example, the same experimenter might give clear instructions to one participant but vague instructions to another. Or one experimenter might greet participants warmly while another barely makes eye contact with them. To the extent that such variables affect participants' behavior, they

add noise to the data and make the effect of the independent variable more difficult to detect. If they vary systematically across conditions, they become confounding variables and provide alternative explanations for the results. For example, if participants in a treatment group are tested by a warm and friendly experimenter and participants in a control group are tested by a cold and unfriendly one, then what appears to be an effect of the treatment might actually be an effect of experimenter demeanor. When there are multiple experimenters, the possibility of introducing extraneous variables is even greater, but is often necessary for practical reasons.

Researcher Robert Rosenthal has spent much of his career showing that this kind of unintended variation in the procedure does, in fact, affect participants' behavior. Furthermore, one important source of such variation is the experimenter's expectations about how participants "should" behave in the experiment. This outcome is referred to as an ***experimenter expectancy effect*** (Rosenthal, 1976). For example, if an experimenter expects participants in a treatment group to perform better on a task than participants in a control group, then they might unintentionally give the treatment group participants clearer instructions or more encouragement or allow them more time to complete the task. In a striking example, Rosenthal and Kermit Fode had several students in a laboratory course in psychology train rats to run through a maze. Although the rats were genetically similar, some of the students were told that they were working with "maze-bright" rats that had been bred to be good learners, and other students were told that they were working with "maze-dull" rats that had been bred to be poor learners. Sure enough, over five days of training, the "maze-bright" rats made more correct responses, made the correct response more quickly, and improved more steadily than the "maze-dull" rats (Rosenthal & Fode, 1963). Clearly, it had to have been the students' expectations about how the rats would perform that made the difference. But how? Some clues come from data gathered at the end of the study, which showed that students who expected their rats to learn quickly felt more positively about their animals and reported behaving toward them in a more friendly manner (e.g., handling them more).

The way to minimize unintended variation in the procedure is to standardize it as much as possible so that it is carried out in the same way for all participants regardless of the condition they are in. Here are several ways to do this:

- Create a written protocol that specifies everything that the experimenters are to do and say from the time they greet participants to the time they dismiss them.
- Create standard instructions that participants read themselves or that are read to them word for word by the experimenter.
- Automate the rest of the procedure as much as possible by using software packages for this purpose or even simple computer slide shows.
- Anticipate participants' questions and either raise and answer them in the instructions or develop standard answers for them.
- Train multiple experimenters on the protocol together and have them practice on each other.

- Be sure that each experimenter tests participants in all conditions.

Another good practice is to arrange for the experimenters to be "blind" to the research question or to the condition in which each participant is tested. The idea is to minimize experimenter expectancy effects by minimizing the experimenters' expectations. For example, in a drug study in which each participant receives the drug or a placebo, it is often the case that neither the participants nor the experimenter who interacts with the participants knows which condition they have been assigned to complete. Because both the participants and the experimenters are blind to the condition, this technique is referred to as a ***double-blind study***. (A single-blind study is one in which only the participant is blind to the condition.) Of course, there are many times this blinding is not possible. For example, if you are both the investigator and the only experimenter, it is not possible for you to remain blind to the research question. Also, in many studies, the experimenter must know the condition because they must carry out the procedure in a different way in the different conditions.

20.5.3 Record Keeping

It is essential to keep good records when you conduct an experiment. It is typical for experimenters to generate a written sequence of conditions before the study begins and then to test each new participant in the next condition in the sequence. As you test them, it is a good idea to add to this list basic demographic information; the date, time, and place of testing; and the name of the experimenter who did the testing. It is also a good idea to have a place for the experimenter to write down comments about unusual occurrences (e.g., a confused or uncooperative participant) or questions that come up. This kind of information can be useful later if you decide to analyze sex differences or effects of different experimenters, or if a question arises about a particular participant or testing session.

Since participants' identities should be kept as confidential (or anonymous) as possible, their names and other identifying information should not be included with their data. In order to identify individual participants, it can, therefore, be useful to assign an identification number to each participant as you test them. Simply numbering them consecutively beginning with 1 is usually sufficient. This number can then also be written on any response sheets or questionnaires that participants generate, making it easier to keep them together.

20.5.4 Manipulation Check

In many experiments, the independent variable is a construct that can only be manipulated indirectly. For example, a researcher might try to manipulate participants' stress levels indirectly by telling some of them that they have five minutes to prepare a short speech that they will then have to give to an audience of other participants. In such situations, researchers often include a ***manipulation check*** in their procedure. A manipulation check is a separate measure of the construct the researcher is trying to manipulate. The purpose of a manipulation check is

to confirm that the independent variable was, in fact, successfully manipulated. For example, researchers trying to manipulate participants' stress levels might give them a paper-and-pencil stress questionnaire or take their blood pressure—perhaps right after the manipulation or at the end of the procedure—to verify that they successfully manipulated this variable.

Manipulation checks are particularly important when the results of an experiment turn out null. In cases where the results show no significant effect of the manipulation of the independent variable on the dependent variable, a manipulation check can help the experimenter determine whether the null result is due to a real absence of an effect of the independent variable on the dependent variable or if it is due to a problem with the manipulation of the independent variable. Imagine, for example, that you exposed participants to happy or sad movie music—intending to put them in happy or sad moods—but you found that this had no effect on the number of happy or sad childhood events they recalled. This could be because being in a happy or sad mood has no effect on memories for childhood events. But it could also be that the music was ineffective at putting participants in happy or sad moods. A manipulation check—in this case, a measure of participants' moods—would help resolve this uncertainty. If it showed that you had successfully manipulated participants' moods, then it would appear that there is indeed no effect of mood on memory for childhood events. But if it showed that you did not successfully manipulate participants' moods, then it would appear that you need a more effective manipulation to answer your research question.

Manipulation checks are usually done at the end of the procedure to be sure that the effect of the manipulation lasted throughout the entire procedure and to avoid calling unnecessary attention to the manipulation (to avoid a demand characteristic). However, researchers are wise to include a manipulation check in a pilot test of their experiment so that they avoid spending a lot of time and resources on an experiment that is doomed to fail and instead spend that time and energy finding a better manipulation of the independent variable.

20.5.5 Pilot Testing

It is always a good idea to conduct a ***pilot test*** of your experiment. A pilot test is a small-scale study conducted to make sure that a new procedure works as planned. In a pilot test, you can recruit participants formally (e.g., from an established participant pool) or you can recruit them informally from among family, friends, classmates, and so on. The number of participants can be small, but it should be enough to give you confidence that your procedure works as planned. There are several important questions that you can answer by conducting a pilot test:

- Do participants understand the instructions?
- What kind of misunderstandings do participants have, what kind of mistakes do they make, and what kind of questions do they ask?
- Do participants become bored or frustrated?

- Is an indirect manipulation effective? (You will need to include a manipulation check.)
- Can participants guess the research question or hypothesis (are there demand characteristics)?
- How long does the procedure take?
- Are computer programs or other automated procedures working properly?
- Are data being recorded correctly?

Of course, to answer some of these questions you will need to observe participants carefully during the procedure and talk with them about it afterward. Participants are often hesitant to criticize a study in front of the researcher, so be sure they understand that their participation is part of a pilot test and you are genuinely interested in feedback that will help you improve the procedure. If the procedure works as planned, then you can proceed with the actual study. If there are problems to be solved, you can solve them, pilot test the new procedure, and continue with this process until you are ready to proceed.

20.6 Media Attributions

[1-5] Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

20.7 Text Attributions

Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

20.8 References

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass, 8*(9), 536–554. https://doi.org/10.1111/spc3.12131

Birnbaum, M. H. (1999). How to show that $9 > 221$: Collect judgments in a between-subjects design. *Psychological Methods, 4*(3), 243–249. https://doi.org/10.1037/1082-989X.4.3.243

Cialdini, R. (2005, April 24). Don't throw in the towel: Use social influence research. *APS Observer*. https://www.psychologicalscience.org/observer/dont-throw-in-the-towel-use-social-influence-research

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*(4, Pt. 1), 377–383. https://doi.org/10.1037/h0025589

Fredrickson, B. L., Roberts, T.-A., Noll, S. M., Quinn, D. M., & Twenge, J. M. (1998). The swimsuit becomes you: Sex differences in self-objectification, restrained eating, and math performance. *Journal of Personality and Social Psychology, 75*(1), 269–284. https://doi.org/10.1037/0022-3514.75.1.269

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research, 35*(3), 472–482. https://doi.org/10.1086/586910

Guéguen, N., & de Gail, M.-A. (2003). The effect of smiling on helping behavior: Smiling and good Samaritan behavior. *Communication Reports, 16*(2), 133–140.

Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. Cambridge University Press.

Morling, B. (2014, March 31). Guide your students to become better research consumers. *APS Observer*. https://www.psychologicalscience.org/observer/teach-your-students-to-be-better-consumers

Moseley, J. B., O'Malley, K., Petersen, N. J., Menke, T. J., Brody, B. A., Kuykendall, D. H., Hollingsworth, J. C., Ashton, C. M., & Wray, N. P. (2002). A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *The New England Journal of Medicine, 347*(2), 81–88. https://doi.org/10.1056/NEJMoa013259

Price, D. D., Finniss, D. G., & Benedetti, F. (2008). A comprehensive review of the placebo effect: Recent advances and current thought. *Annual Review of Psychology, 59*, 565–590. https://doi.org/10.1146/annurev.psych.59.113006.095941

Rosenthal, R. (1976). *Experimenter effects in behavioral research* (Enlarged ed.). Wiley.

Rosenthal, R., & Fode, K. (1963). The effect of experimenter bias on performance of the albino rat. *Behavioral Science, 8*(3), 183–189. https://doi.org/10.1002/bs.3830080302

Rosenthal, R., & Rosnow, R. L. (1976). *The volunteer subject*. Wiley.

Shapiro, A. K., & Shapiro, E. (1999). *The powerful placebo: From ancient priest to modern physician*. Johns Hopkins University Press.

21 Chapter 21: Confounds

[In this chapter, we will look at many of the ways that aspects of the study design provide alternative explanations for research findings. When present, these aspects can threaten the internal and external validity of the study we are examining. Ideally, we can keep internal and external validity as high as possible, but often that involves a tradeoff between the two. We will look at several ways that these threats influence validity and ways to prevent these threats from arising.]

21.1 Confounds and Artifacts

If we look at the issue of validity in the most general fashion the two biggest worries that we have are *confounds* and *artifacts*. These two terms are defined in the following way:

- ***Confound*:** A confound is an additional, often unmeasured variable that turns out to be related to both the predictor variable and the outcome variable. The existence of confounds threatens the *internal validity* of the study because you cannot tell whether the predictor causes the outcome, or if the confounding variable causes it.
- ***Artifact*:** A result is said to be "artifactual" if it only holds in the special situation that you happened to test in your study. The possibility that your result is an artifact poses a threat to your *external validity*, because it raises the possibility that you cannot generalize or apply your results to the actual population that you care about.

As a general rule, confounds are a bigger concern for non-experimental studies, precisely because they're not proper experiments. By definition, you're leaving many things uncontrolled, so there's a lot of scope for confounds being present in your study. Properly conducted experimental research tends to be much less vulnerable to confounds. The more control you have over what happens during the study, the more you can prevent confounds from affecting the results. With random assignment, for example, confounds are distributed randomly, and evenly, between different groups.

However, artifacts can impact results just as much as confounds. For the most part, artifactual results tend to be more of a concern for experimental studies than for non-experimental studies. To see this, it helps to realize that the reason that a lot of studies are non-experimental is precisely because what the researcher is trying to do is examine human behavior in a more naturalistic context. By working in a more real-world context you lose experimental

control (making yourself vulnerable to confounds), but because you tend to be studying human psychology "in the wild" you reduce the chances of getting an artifactual result. Or, to put it another way, when you take psychology out of the wild and bring it into the lab (which we usually have to do to gain our experimental control), you always run the risk of accidentally studying something different to what you wanted to study.

Be warned though. The above is a rough guide only. It's absolutely possible to have confounds in an experiment, and to get artifactual results with non-experimental studies. This can happen for all sorts of reasons, not least of which is experimenter or researcher error. In practice, it's really hard to think everything through ahead of time and even very good researchers make mistakes.

Many threats to validity can be classified as either confounds or artifacts. Because these are broad concepts, it is helpful to examine some of the most common examples.

21.2 Specific Confounds and Artifacts

21.2.1 Selection effects

Selection bias is a pretty broad term. Suppose that you're running an experiment with two groups of participants where each group gets a different "treatment", and you want to see if the different treatments lead to different outcomes. However, suppose that, despite your best efforts, you ended up with a gender imbalance across groups (say, group A has 80% females and group B has 50% females). It might sound like this could never happen, but it can. This is an example of a selection bias, in which the people "selected into" the two groups have different characteristics. If any of those characteristics turns out to be relevant (say, your treatment works better on females than males) then you're in a lot of trouble.

You can avoid selection bias in multiple ways. One way is by ***randomly assigning*** participants to different conditions of your independent variable. Randomly assigning participants across conditions spreads participant characteristics out across conditions. In the above case, both groups A and B would have 50% females. Another way is by using a ***within-subjects design***, where participants complete every condition in your independent variable (see Chapter 20 for additional information). This way, you can reduce the influence of irrelevant characteristics by having each participant act as their own control participant.

21.2.2 Attrition

Attrition usually means something is getting reduced; and in research, that would mean participants are leaving or dropping out of the study for any reason. When thinking about the effects of attrition, it is sometimes helpful to distinguish between two different types. The first is ***homogeneous attrition***, in which the attrition effect is the same for all groups, treatments

or conditions. In the example I gave above, the attrition would be homogeneous if (and only if) the male participants drop out of all of the conditions in my experiment at about the same rate. In general, the main effect of homogeneous attrition is likely to be that it makes your sample *unrepresentative*. As such, the biggest worry that you'll have is that the generalizability of the results decreases. In other words, you lose external validity.

The second type of attrition is ***heterogeneous attrition***, in which the attrition effect is different for different groups. More often called ***differential attrition***, this is a kind of selection bias that is caused by the study itself. Suppose that, for the first time ever in the history of psychology, we manage to find a perfectly balanced and representative sample of people. We start running "our incredibly long and tedious experiment" on our perfect sample but then, because the study is incredibly long and tedious, many people start dropping out. We can't stop this. Participants absolutely have the right to stop doing any experiment, any time, for whatever reason they feel like, and as researchers we are morally (and professionally) obliged to remind people that they do have this right. So, suppose that "our incredibly long and tedious experiment" has a very high drop-out rate. What do you suppose the odds are that this drop out is random? Answer: zero. Almost certainly the people who remain are more conscientious, more tolerant of boredom, etc., than those who leave. To the extent that (say) conscientiousness is relevant to the psychological phenomenon that I care about, this attrition can decrease the internal validity of our results.

Let's look at another example. Let's say we wanted to look at how using math tutors affect high school SAT scores. We can obtain a group of 10th grade students, get their SAT scores, and randomize them to have a math tutor or not have a math tutor. Two years later, when the students are in 12th grade, we get their SAT scores again.

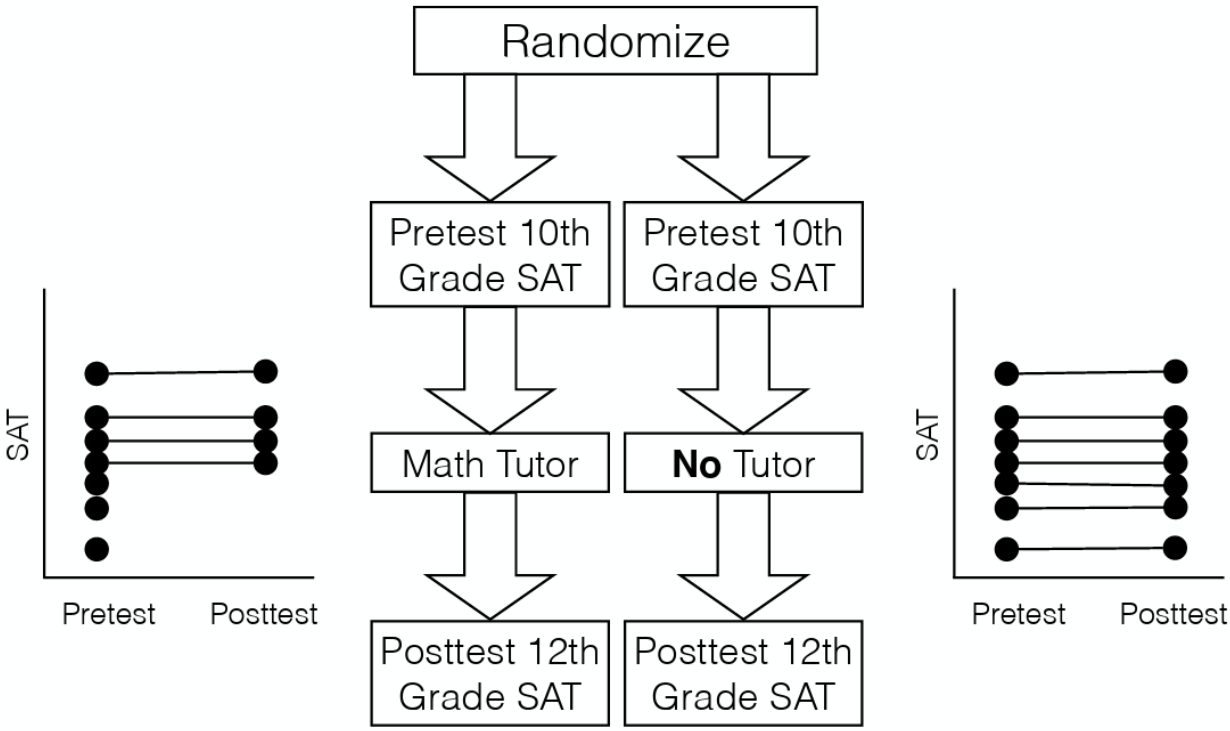


Figure 21.1: **Figure 21.1** Randomize 10th graders, pretest their SAT scores, assign them to Math Tutor or No Tutor, and Posttest their SAT scores in 12th grade[1]

Two years is a long time, and things can happen such as students moving away or experiencing hospitalization, things that can affect our ability to get the SAT scores of some of our participants. When we test for differences in posttest SAT scores we may see differences in posttest SAT scores, but those differences would have occurred for reasons beyond the use of a Math Tutor.

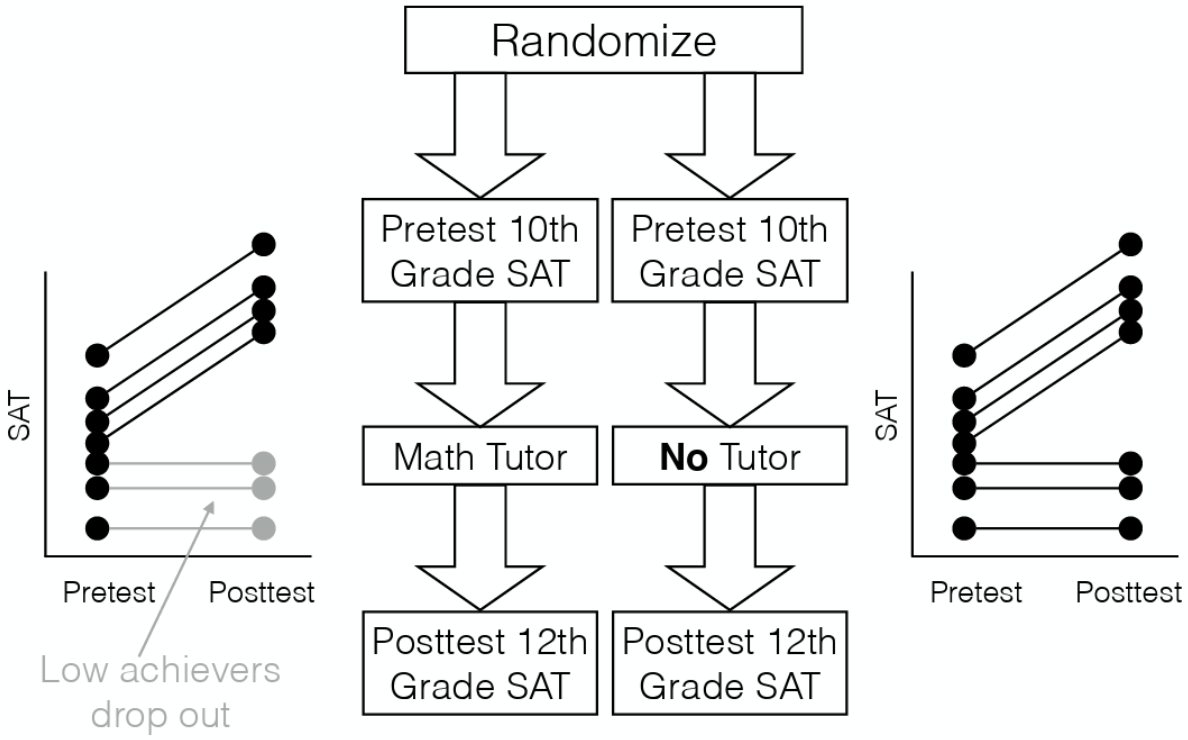


Figure 21.2: **Figure 21.2** More students in the Math Tutor group dropped out than students in the No Tutor group[2]

Addressing attrition can be tricky, but it is doable. One way is to shorten the length of your study, particularly if the study is long. If the study is short, then there is little chance of participants dropping out of the study. Another way is to motivate participants to complete your study. One easy-to-consider way is to provide monetary payment for completing the study. You can also provide other forms of motivation such as desired items (for example, candy for young children) or chances to win a prize.

A final way to address attrition is to check whether pretest scores differed for the group that dropped out and the group that remained. If scores are similar, then that is one way to suggest that participants in both groups did not differ in any especially relevant way and that random assignment worked.

21.2.3 Non-response bias

Non-response bias is closely related to selection bias and to differential attrition (for a reminder of what non-response bias is, see Chapter 6). The simplest version of the problem goes like this. You mail out a survey to 1000 people but only 300 of them reply. The 300 people who replied are almost certainly not a random subsample. People who respond to surveys are systematically different to people who don't. This introduces a problem when trying to generalize from those 300 people who replied to the population at large, since you now have a very non-random sample. The issue of non-response bias is more general than this, though. Among the (say) 300 people that did respond to the survey, you might find that not everyone answers every question. If (say) 80 people chose not to answer one of your questions, does this introduce problems? As always, the answer is maybe. If the question that wasn't answered was on the last page of the questionnaire, and those 80 surveys were returned with the last page missing, there's a good chance that the missing data isn't a big deal; probably the pages just fell off. However, if the question that 80 people didn't answer was the most confrontational or invasive personal question in the questionnaire, then almost certainly you've got a problem. In essence, what you're dealing with here is what's called the problem of ***missing data***. If the data that is missing was "lost" randomly, then it's not a big problem. If it's missing *systematically*, then it can be a big problem.

Addressing non-response bias can be like how you would address attrition: provide a reward for responding. Alternatively, you can use statistical analyses that help you take into account the issue of missing data. Discussion of these analyses is beyond the scope of this book.

Another way to address non-response bias is to remind participants of their confidentiality and anonymity. This is especially important if participants are concerned about any repercussions for answering certain questions the way they do.

21.2.4 Maturation

Maturation effects are fundamentally about change over time. However, maturation effects aren't in response to specific events. Rather, they relate to how people change on their own over time. We get older, we get tired, we get bored, and so on. Some examples of maturation effects are:

I. When doing developmental psychology research, you need to be aware that children grow up quite rapidly. So, suppose that you want to find out whether some educational trick helps with vocabulary size among 3-year-olds. One thing that you need to be aware of is that the vocabulary size of children that age is growing at an incredible rate (multiple words per day) all on its own. If you design your study without taking this maturational effect into account, then you won't be able to tell if your educational trick works.

II. When running a very long experiment in the lab (say, something that lasts for three hours) it's very likely that people will begin to get bored and tired, and that this maturational effect will cause performance to decline regardless of anything else going on in the experiment.

To control for maturation effects, you can add a randomly assigned control group. Any effects of maturation should be similar in each group, and ideally you will see a difference between the groups based on your independent variable.

21.2.5 History

History effects refer to the possibility that specific events may occur during the study that might influence the outcome measure. For instance, something might happen in- between a pretest and a post-test. Or in-between testing participant 23 and participant 24, such as a fire alarm interrupting data collection or a major news event that changes participants' mood or opinions. Examples of things that would count as history effects are:

- You're interested in how people think about risk and uncertainty. You started your data collection in January 2020. But finding participants and collecting data takes time, so you're still finding new people in January 2021. Unfortunately for you (and even more unfortunately for others), the COVID pandemic occurred during this time, killing many people and causing widespread social distancing for months. Not surprisingly, the people tested in January 2020 express quite different beliefs about handling risk than the people tested in January 2021. Which (if any) of these reflects the "true" beliefs of participants? I think the answer is probably both. The COVID pandemic genuinely changed the beliefs of the global public, though possibly only temporarily. The key thing here is that the "history" of the people tested in January 2020 is quite different to people tested in January 2021.
- You're testing the psychological effects of a new anti-anxiety drug. So what you do is measure anxiety before administering the drug (e.g., by self-report, and taking physiological measures). Then you administer the drug, and afterwards you take the same measures. In the interim however, because your lab is in Los Angeles, there's an earthquake which increases the anxiety of the participants.

There are multiple ways to control for history effects. Like with maturation effects, you can add a randomly assigned control group to the study, as all participants will have experienced the same historical events. Another way is to change the timing of the study. One especially effective way to do so is to conduct a longitudinal study (see Chapter 15) in which you measure your dependent variable before *and* after the historical event. This way, you will be able to assess whether the historical event caused any changes in participants, along with your independent variable.

21.2.6 Testing effects

Another threat to internal validity is testing effects. Suppose I want to take two measurements of some psychological construct, such as anxiety, at two different points in time. One thing I might be worried about is if the first measurement has an effect on the second measurement. This phenomenon is more common than you might expect. Examples of this include:

- **Practice or learning**: For example, participants may score higher on an intelligence test the second time simply because they become more familiar with the types of questions or developed better test-taking strategies, not because they actually became more intelligent.
- **Familiarity with the testing situation**: For example, if people are nervous at time 1, this might make performance go down. But after sitting through the first testing situation they are more comfortable with the testing environment and perform better simply because they are less anxious.
- **The first measurement changes later responses**: For example, if a questionnaire assessing mood is boring, then mood rating at measurement time 2 is more likely to be "bored" precisely because of the boring measurement made at time 1.

A common way to address this confound, especially if the concern is that participants' exposure to one measure influences their response to later measures, is to use a *within-subjects* design and randomize the order of the conditions that participants are exposed to.

Another way to address this confound is to use a *between-subjects* design, in which each participant is tested only once. However, this approach has an important limitation: Because different people are measured in each condition, researchers cannot examine how the same individuals change over time.

21.2.7 Order effects

Related to testing effects are order effects. Order effects occur when the exposure to one condition changes participants' response to a later condition. These can occur in within-subjects designs. There are different types of order effects. One type is a ***carryover effect***, which is a type of order effect in which some form of contamination (through exposure to the independent variable) carries over from one condition to the next. The previous treatment changes the participant, and those changes carry over into the subsequent treatment and change how the participant performs. This would be like what would happen if you drank orange juice right after brushing your teeth with mint toothpaste, rather than at other times—the taste of the orange juice would be different from usual (yuck!).

One type of carryover effect is the fatigue effect, where performance declines as participants get tired or bored over time after being exposed to the dependent variable. Another type of carryover effect is the practice effect, where performance improves as participants get better

over time after being exposed to the dependent variable, just like you would see with any skill.

Being tested in one condition can also change how participants perceive stimuli or interpret their task in later conditions. This type of effect is called a ***context effect (or contrast effect)***. For example, in a juror decision making study, an average-looking defendant might be judged more harshly when participants have just judged an attractive defendant than when they have just judged an unattractive defendant.

Because we can't prevent the possibility of carryover effects when we expose the same participant to more than one condition, we want to *balance* the influence of conditions on each other. We can do this through ***counterbalancing***. This involves randomizing the orders of the conditions can be exposed to. See Chapter 20 to learn more. Another way to address the issue of order effects is to use a between-subjects design. Doing so is especially effective for preventing issues like fatigue or context effects influencing results, as participants would be exposed to one condition of the independent variable.

21.2.8 Instrumentation

Instrumentation can be a threat to the internal validity of studies. Instrumentation refers to when the basic characteristics of the measuring instrument change over time. For example, let's say a body weight scale becomes mis-calibrated and begins to record weights that are 4 pounds heavier than a participant's true weight. This may cause any conclusions we can draw about weight to be somewhat inaccurate. We can also see instrumentation effects arise when we use human coders (for a reminder of behavioral coding, see Chapter 17). When human observers are used to measure behavior, they may, over time, get better at the task, become fatigued, or change the standards on which observations are based. This can also impact conclusions we can draw about behavior.

One way that instrumentation effects have appeared is through the diagnosis of Autism. Recent decades have shown increasing rates of Autism diagnoses (Centers for Disease Control and Prevention, 2025). Much of the increase can be attributed to changes in criteria as laid out in the Diagnostic and Statistical Manual of Mental Disorders (DSM), changing how clinicians diagnose Autism (Harris, 2023). The increase can also be attributed to more providers being aware of what behaviors to look for, compared to earlier decades.

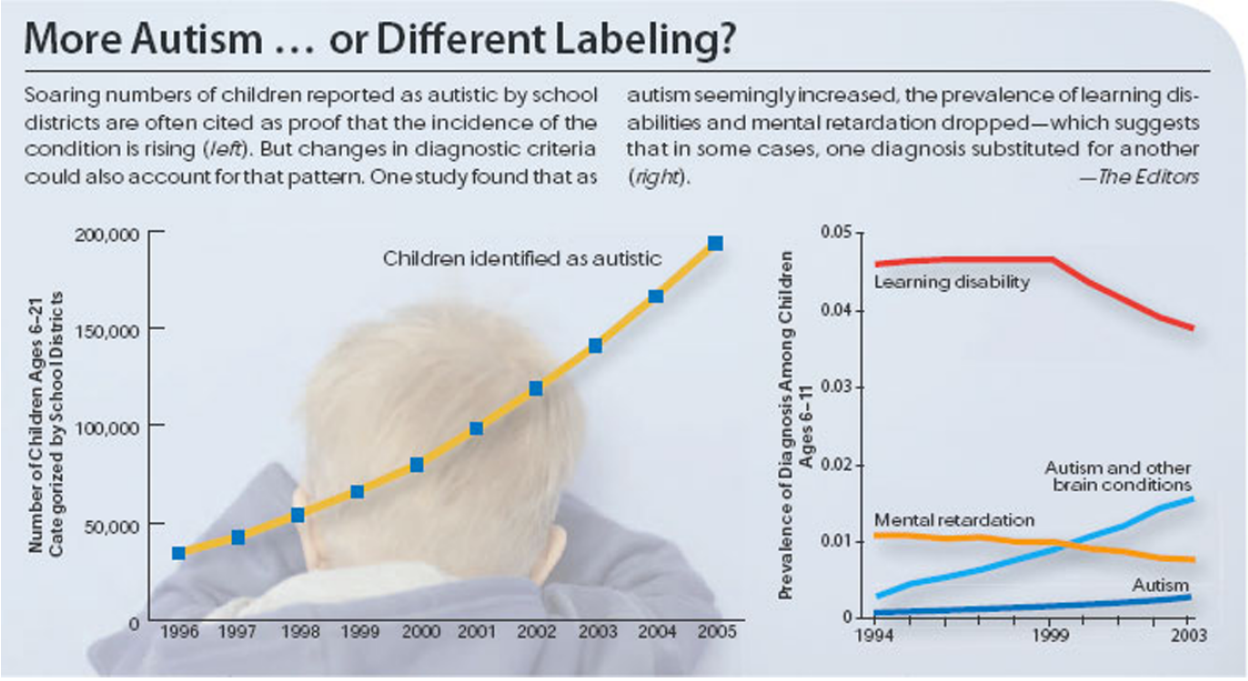


Figure 21.3: **Figure 21.3** Charts illustrating rising rates of children identified as autistic and reduction in rates of diagnosis of learning disability[3]

In many studies, the main way to address instrumentation effects is to keep the way you are measuring your variables the same throughout your study. For example, for studies using human coders to measure behavior, the coders should use standardized codebooks and take frequent breaks to prevent fatigue from affecting the nature of the coding. Another way to identify instrumentation effects is to include a randomly assigned control group. Like what you see with history or maturation effects, you assess whether any changes in your control and treatment groups occur only because of your independent variable. If the control group changes on your dependent variable even though it should not do so, that could be a sign of an instrumentation effect.

21.2.9 Demand characteristics

In Chapter 5, you learned about demand characteristics, cues provided by the researcher/situation that communicate the purpose of the study (that is, information telling participants how they should/should not behave). These are also called reactivity or demand effects. One of the things to remember when conducting psychological research is that people are not passive – they have their own experiences and motivations – they are active and responsive.

The basic idea is captured by the ***Hawthorne effect***: people alter their performance because of the attention that the study focuses on them. The effect takes its name from a study that took place in the "Hawthorne Works" factory outside of Chicago (see Adair, 1984). This study, from the 1920s, looked at the effects of factory lighting on worker productivity. But, importantly, change in worker behavior occurred because the workers knew they were being studied, rather than any effect of factory lighting.

To get a bit more specific about some of the ways in which the mere fact of being in a study can change how people behave, it helps to think like a social psychologist and look at some of the roles that people might adopt during an experiment but might not adopt if the corresponding events were occurring in the real world:

- The *good participant* tries to be too helpful to the researcher. He or she seeks to figure out the experimenter's hypotheses and confirm them.
- The *negative participant* does the exact opposite of the good participant. He or she seeks to break or destroy the study or the hypothesis in some way.
- The *faithful participant* is unnaturally obedient. He or she seeks to follow instructions perfectly, regardless of what might have happened in a more realistic setting.
- The *apprehensive participant* gets nervous about being tested or studied, so much so that his or her behavior becomes highly unnatural, or overly socially desirable.

There are multiple ways to reduce demand characteristics. One way is to eliminate any possible cues to your research question. You would need to ensure there are no cues in the area in which the study is conducted, such as providing distractor stimuli or tasks in the study. Additionally, you need to have a standard procedure, ensuring that experimenters avoid accidentally cueing behaviors from participants.

Another way to reduce demand characteristics is to use a between-subjects design with random assignment. Relatedly, you can avoid informing participants who may believe there are multiple conditions of the study they are in. If there are multiple conditions, you would avoid informing participants of which condition they are in. A study using this strategy is called a ***single-blind study***. You can go further by having both participants and experimenters not know the study condition that a given participant is in. A study using this strategy is called a ***double-blind study***.

A last way to avoid demand characteristics is to use *deception*. There are different ways to use deception. One way is to avoid informing participants of the true purpose of the study. This is called passive deception, and is a strategy used by many experimental psychologists to decrease demand effects. Another way is to provide false information verbally or by using detractors (also called confederates, people who act as participants). This type of strategy is called active deception and should only be used if there is no other way to study the research question of interest. As discussed in the Ethics chapter, debriefing is especially important if false information is provided.

21.2.10 Experimenter bias

Experimenter bias (also called ***observer bias***) can come in multiple forms. The basic idea is that the experimenter, despite the best of intentions, can accidentally end up influencing the results of the experiment by subtly communicating the "right answer" or the "desired behavior" to the participants. Typically, this occurs because the experimenter has special knowledge that the participant does not, for example the right answer to the questions being asked or knowledge of the expected pattern of performance for the condition that the participant is in. The classic example of this happening is the case study of "Clever Hans", which dates back to 1907 (Pfungst, 1911). Clever Hans was a horse that apparently was able to read and count and perform other human-like feats of intelligence. After Clever Hans became famous, psychologists started examining his behavior more closely. It turned out that, not surprisingly, Hans didn't know how to do math. Rather, Hans was responding to the human observers around him, because the humans did know how to count and the horse had learned to change its behavior when people changed theirs.

One way to solve experimenter bias is to *standardize* the procedure to prevent experimenters from subtly communicating what they want to see from the participants. Commonly, using experimental scripts will help experimenters perform the same actions in a study and say the same things to all participants. One form of standardization is using *constancy*. This is a technique where all participants in the study experience a single value of a variable. For example, all participants complete a study in the same room or with only one experimenter.

Another solution is to use *double-blind* studies, where neither the experimenter nor the participant knows which condition the participant is in or knows what the desired behavior is. This provides a very good solution to the problem, but it's important to recognize that it's not quite ideal, and hard to pull off perfectly. For instance, the obvious way that we could try to construct a double-blind study is to have one of our research assistants (one who doesn't know anything about the experiment) run the study. But because we have lab meetings where we talk about our ongoing studies that often examine similar topics, it can be difficult for the research assistant to avoid knowing at least a little bit about what is happening. This means that the research assistant may have guesses about the hypotheses of the study. When they act as an experimenter, those guesses may influence their behavior and thus participants' behavior.

Other knowledge that research assistants have about the tests typically used in the lab can also have an effect. Suppose the experimenter accidentally conveys the fact that the participants are expected to do well in a particular task. Well, there's a thing called the "Pygmalion effect", where if you expect great things from people they'll tend to rise to the occasion. But if you expect them to fail then they may do that too. In other words, the expectations become a self-fulfilling prophecy. Situations such as these emphasize the importance of having standardized experimenter scripts and using constancy.

Lastly, as with other confounds discussed above, random assignment to different conditions can

also help to reduce some experimenter influence. This strategy is especially effective if paired with using a double-blind study.

21.2.11 Placebo

The ***placebo effect*** is a specific type of demand effect that we worry a lot about. It refers to the situation where the mere fact of being treated causes an improvement in outcomes. The classic example comes from clinical trials. If you give people a completely chemically inert (inactive) drug and tell them that it's a cure for a disease, they will tend to get better faster than people who aren't treated at all. In other words, it is people's belief that they are being treated that causes the improved outcomes, not the drug.

However, the current consensus in medicine is that true placebo effects are quite rare and most of what was previously considered placebo effect is in fact some combination of natural healing (some people just get better on their own), regression to the mean (see below), and other quirks of study design. Of interest to psychology is that the strongest evidence for at least some placebo effect is in self-reported outcomes, most notably in treatment of pain (Hróbjartsson & Gøtzsche, 2010).

To control for the placebo effect, you should add a "nothing" control group. That is, make sure the control group does not experience the treatment condition.

21.2.12 Regression to the mean and spontaneous remission

Regression to the mean is another alternative explanation for a change in the dependent variable. This refers to the statistical fact that an individual who scores extremely high or extremely low on a variable on one occasion will tend to score less extremely on the next occasion. For example, a bowler with a long-term average of 150 who suddenly bowls a 220 will almost certainly score lower in the next game. Her score will "regress" toward her mean score of 150. Regression to the mean can be a problem when participants are selected for further study *because* of their extreme scores. Imagine, for example, that only students who scored especially high on the test of attitudes toward illegal drugs (those with extremely favorable attitudes toward drugs) were given the anti-drug program and then were retested. Regression to the mean all but guarantees that their scores will be lower at the posttest even if the training program has no effect.

Regression to the mean is surprisingly common. For instance, if two very tall people have kids their children will tend to be taller than average but not as tall as the parents. The reverse happens with very short parents. Two very short parents will tend to have short children, but nevertheless those kids will tend to be taller than the parents.

A closely related concept—and an extremely important one in psychological research—is ***spontaneous remission***. This is the tendency for many medical and psychological problems

to improve over time without any form of treatment. The common cold is a good example. If one were to measure symptom severity in 100 common cold sufferers today, give them a bowl of chicken soup every day, and then measure their symptom severity again in a week, they would probably be much improved. This does not mean that the chicken soup was responsible for the improvement, however, because they would have been much improved without any treatment at all. The same is true of many psychological problems. A group of severely depressed people today is likely to be less depressed on average in 6 months. In reviewing the results of several studies of treatments for depression, researchers Michael Posternak and Ivan Miller found that participants in waitlist control conditions improved an average of 10 to 15% before they received any treatment at all (Posternak & Miller, 2001). Thus, one must generally be very cautious about inferring causality from pretest-posttest designs.

To control for regression to the mean, you should add a randomly assigned control group. If both groups begin with similarly extreme scores, any natural regression toward the average should occur in both groups, making it easier to determine whether the treatment had an additional effect.

21.2.13 Diffusion of Treatment

Diffusion of treatment occurs when the communication across groups interferes with the manipulation of an independent variable. Suppose there are two groups in a study that looks at a new study habit, and there is a treatment and control group. The treatment group is given information about this new study habit and reasons why it works. Imagine the participants in this condition later pass on the helpful information to their friends, who happen to be in the control condition. Because the control condition has now been contaminated, any differences found between the two groups may not be due to the independent variable.

There are a few ways to address diffusion of treatment. One way to use random assignment, as seen with other confounds. Another way is to measure the dependent variable at only one time point, preventing possibility of information about the study from leaving the lab. A final way to address diffusion of treatment is to use a within-subjects design. That way, you can measure changes on your dependent variable, comparing participants to themselves; and participants can experience the benefits of any treatment that they are exposed to.

As you can see, there are many situations that can provide alternative explanations for any results you see in a study. These alternative explanations, confounds and artifacts, can reduce the internal and external validity of a study. Confounds raise issues of causal inference and impact internal validity. Artifacts raise issues of generalizability and impact external validity. Confounds and artifacts should be considered as early as possible when studies are created and throughout the time the study is occurring. Although they impact validity, there are ways to address them.

For brief overview of many of the above topics, please see the below video, "How to Evaluate Alternative Hypotheses".

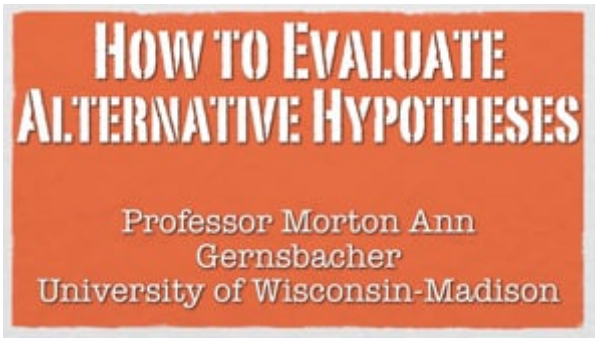


Figure 21.4: How To Evaluate Alternative Hypotheses

21.3 Media Attributions

[1-3] [Rottman,] B. M. (2026, July 7). *O[pen Source Research Methods for the Social Sciences]*[. https://canvas.pitt.edu/courses/124970.] Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

21.4 Text Attributions

D'Costa, S., Ukeye, M., O'Neil, M., Anguiano, R. (n.d.). Critical Research Methods in Psychology. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

Jhangiani, R.S., Chiang, I-C.A., Cuttler, C. & Leighton, D.C. (2019). Research methods in psychology. 4th edition. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[Navarro, D., & Foxcroft, D. (2025). Learning Statistics with JAMOVI: a tutorial for beginners in statistical analysis.] Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

Rottman, B. M. (2026, July 7). *Open Source Research Methods for the Social Sciences*. https://canvas.pitt.edu/courses/124970. Licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.]

Pearcey, Sharon; Kirsner, Beth; Randall, Christopher; Willard, Jen; Williamson, Adrienne; and Downtain, Tricia, "Experimental Design and Analysis: Ancillary Set" (2017). *Psychology, Sociology, Anthropology, and Social Work Ancillary Materials*. 4. https://oer.galileo.usg.edu/psychology-ancillary/4. This work is licensed under a Creative Commons Attribution 4.0 International License.

21.5 References

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology, 69*(2), 334–345. https://doi.org/10.1037/0021-9010.69.2.334

Centers for Disease Control and Prevention. (2025, May 27). *Data and Statistics on Autism Spectrum Disorder*. U.S. Department of Health & Human Services. https://www.cdc.gov/autism/data-research/index.html

Harris, E. (2023). Autism prevalence has been on the rise in the US for decades—and that's progress. *JAMA, 329*(20),1724–1726. https://doi.org/10.1001/jama.2023.6078

Hróbjartsson, A., & Gøtzsche, P. (2010). Placebo interventions for all clinical conditions. *Cochrane Database of Systematic Reviews*, *1*. https://doi.org/10.1002/14651858.cd003974.pub3

Pfungst, O. (1911). *Clever Hans:(the horse of Mr. Von Osten.) a contribution to xperimental animal and human psychology*. Holt, Rinehart and Winston.

Posternak, M. A., & Miller, I. (2001). Untreated short-term course of major depression: A meta-analysis of studies using outcomes from studies using wait-list control groups. *Journal of Affective Disorders, 66*, 139–146. https://doi.org/10.1016/S0165-0327(00)00304-9

22 Chapter 22: Quasi-Experimental Designs

The prefix *quasi* means "resembling." Thus, quasi-experimental research is research that resembles experimental research but is not true experimental research. Recall that with a true between-groups experiment, random assignment to conditions is used to ensure the groups are equivalent, and with a true within-subjects design, counterbalancing is used to guard against order effects. Quasi-experiments are missing one of these safeguards. Although an independent variable is manipulated, either a control group is missing or participants are not randomly assigned to conditions (Cook & Campbell, 1979).

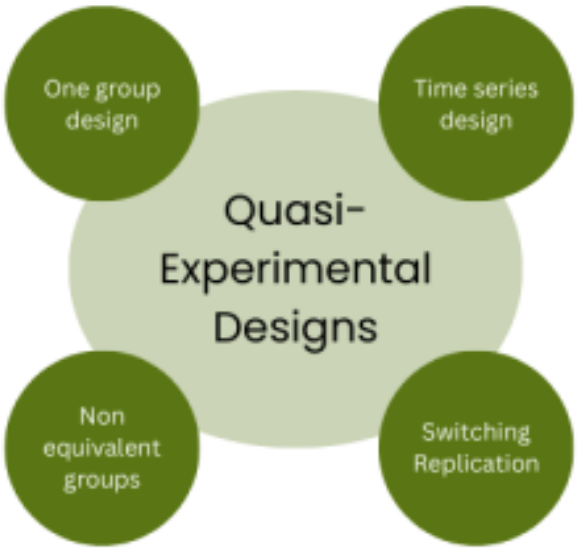


Figure 22.1: Graphic with main circle of examples of quasi experimental designs. Top left: one group design. Top right: time series design. Bottom left: non equivalent groups. Bottom right: switching replication.

Figure 22.1 *Four types of Quasi-Experimental Designs* [1]

Because the independent variable is manipulated before the dependent variable is measured, quasi-experimental research eliminates the directionality problem associated with non-experimental research. But because either counterbalancing techniques are not used or participants are not randomly assigned to conditions—making it likely that there are other differences between conditions—quasi-experimental research does not eliminate the problem of confounding variables. On a continuum of internal validity, therefore, quasi-experiments are generally somewhere between non experimental studies (lowest internal validity) and true experiments (highest internal validity). Recall the illustration you saw in Chapter 15:

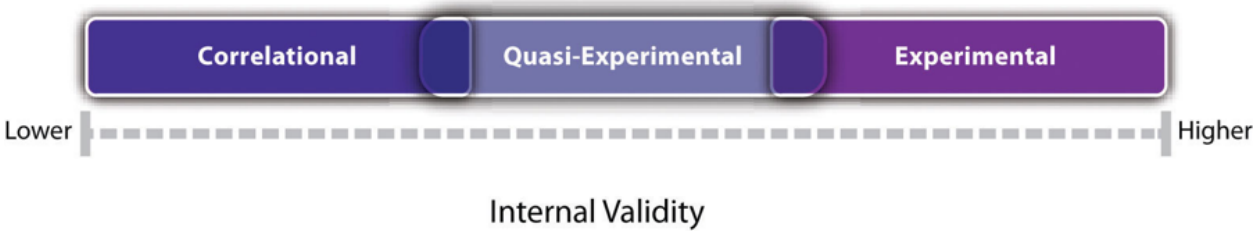


Figure 22.2: Internal validity continuum goes from low to high, with correlational on the left, then quasi-experimental in the middle, then experimental on the right.

Figure 22.2 *Internal Validity of Correlation, Quasi-Experimental, and Experimental Studies* [2]

Quasi-experiments are most likely to be conducted in field settings in which random assignment is difficult or impossible. They are often conducted to evaluate the effectiveness of a treatment—perhaps a type of psychotherapy or an educational intervention. There are many different kinds of quasi-experiments, but we will discuss just a few of the most common ones in this chapter.

22.1 One-Group Designs

22.1.1 One-Group Posttest Only Design

In a ***one-group posttest only design***, a treatment is implemented (or an independent variable is manipulated), and then a dependent variable is measured once after the treatment is implemented. Imagine, for example, a researcher who is interested in the effectiveness of an anti-drug education program on elementary school students' attitudes toward illegal drugs. The researcher could implement the anti-drug program, and then immediately after the program ends, the researcher could measure students' attitudes toward illegal drugs.

This is the weakest type of quasi-experimental design. A major limitation of this design is the lack of a control or comparison group. There is no way to determine what the attitudes of

these students would have been if they hadn't completed the anti-drug program. Despite this major limitation, results from this design are frequently reported in the media and are often misinterpreted by the general population. For instance, advertisers might claim that 80% of women noticed their skin looked brighter after using Brand X cleanser for a month. However, if there is no comparison group, then this statistic means little to nothing.

22.1.2 One-Group Pretest-Posttest Design

In a ***one-group pretest-posttest design***, the dependent variable is measured once before the treatment is implemented and once after it is implemented. Let's return to the example of a researcher who is interested in the effectiveness of an anti-drug education program on elementary school students' attitudes toward illegal drugs. The researcher could measure the attitudes of students at a particular elementary school during one week, implement the anti-drug program during the next week, and finally, measure their attitudes again the following week. The pretest-posttest design is much like a within-subjects experiment in which each participant is tested first under the control condition and then under the treatment condition. It is *unlike* a within-subjects experiment, however, in that the order of conditions is not counterbalanced because it typically is not possible for a participant to be tested in the treatment condition first and then in an "untreated" control condition.

If the average posttest score is better than the average pretest score (e.g., attitudes toward illegal drugs are more negative after the anti-drug educational program), then it makes sense to conclude that the treatment might be responsible for the improvement. Unfortunately, one often cannot conclude this with a high degree of certainty because there may be other explanations for why the posttest scores may have changed. These alternative explanations pose threats to internal validity.

A common approach to ruling out the threats to internal validity described above is by revisiting the research design to include a randomly assigned control group, one that does not receive the treatment effect. A control group would be subject to the same threats from history, maturation, testing, instrumentation, regression to the mean, and spontaneous remission, and so would allow the researcher to measure the actual effect of the treatment (if any). Of course, including a control group would mean that this is no longer a one-group design.

> **Does Psychotherapy Work?**
>
> Early studies on the effectiveness of psychotherapy tended to use one-group pretest-posttest designs. In a classic 1952 article, researcher Hans Eysenck summarized the results of 24 such studies showing that about two thirds of patients improved between the pretest and the posttest (Eysenck, 1952). But Eysenck also compared these results with archival data from state hospital and insurance company records showing that similar patients recovered at about the

same rate *without* receiving psychotherapy. This parallel suggested to Eysenck that the improvement that patients showed in the pretest-posttest studies might be no more than spontaneous remission. Note that Eysenck did not conclude that psychotherapy was ineffective. He merely concluded that there was no evidence that it was, and he wrote of "the necessity of properly planned and executed experimental studies into this important field" (p. 323). You can read the entire article here on the Classics in the History of Psychology website.

Fortunately, many other researchers took up Eysenck's challenge, and by 1980 hundreds of experiments had been conducted in which participants were randomly assigned to treatment and control conditions, and the results were summarized in a classic book by Mary Lee Smith, Gene Glass, and Thomas Miller (Smith, Glass, & Miller, 1980). They found that overall psychotherapy was quite effective, with about 80% of treatment participants improving more than the average control participant. Subsequent research has focused more on the conditions under which different types of psychotherapy are more or less effective.

22.1.3 Interrupted Time-Series Design

A variant of the pretest-posttest design is the ***interrupted time-series design***. A time-series is a set of measurements taken at intervals over a period of time. For example, a manufacturing company might measure its workers' productivity each week for a year. In an interrupted time-series design, a time-series like this one is "interrupted" by a treatment. In one classic example, the treatment was the reduction of the work shifts in a factory from 10 hours to 8 hours (Cook & Campbell, 1979). Because productivity increased rather quickly after the shortening of the work shifts, and because it remained elevated for many months afterward, the researcher concluded that the shortening of the shifts caused the increase in productivity. Notice that the interrupted time-series design is like a pretest-posttest design in that it includes measurements of the dependent variable both before and after the treatment. It is unlike the pretest-posttest design, however, in that it includes *multiple* pretest and posttest measurements.

Figure 22.3 shows data from a hypothetical interrupted time-series study. The dependent variable is the number of student absences per week in a research methods course. The treatment is that the instructor begins publicly taking attendance each day so that students know that the instructor is aware of who is present and who is absent. The top panel of Figure 22.3 shows how the data might look if this treatment worked. There is a consistently high number of absences before the treatment, and there is an immediate and sustained drop in absences after the treatment. The bottom panel of Figure 22.3 shows how the data might look if this treatment did not work. On average, the number of absences after the treatment is about the same as the number before. This figure also illustrates an advantage of the interrupted time-series design over a simpler pretest-posttest design. If there had been only

one measurement of absences before the treatment at Week 7 and one afterward at Week 8, then it would have looked as though the treatment were responsible for the reduction. The multiple measurements both before and after the treatment suggest that the reduction between Weeks 7 and 8 is nothing more than normal week-to-week variation.

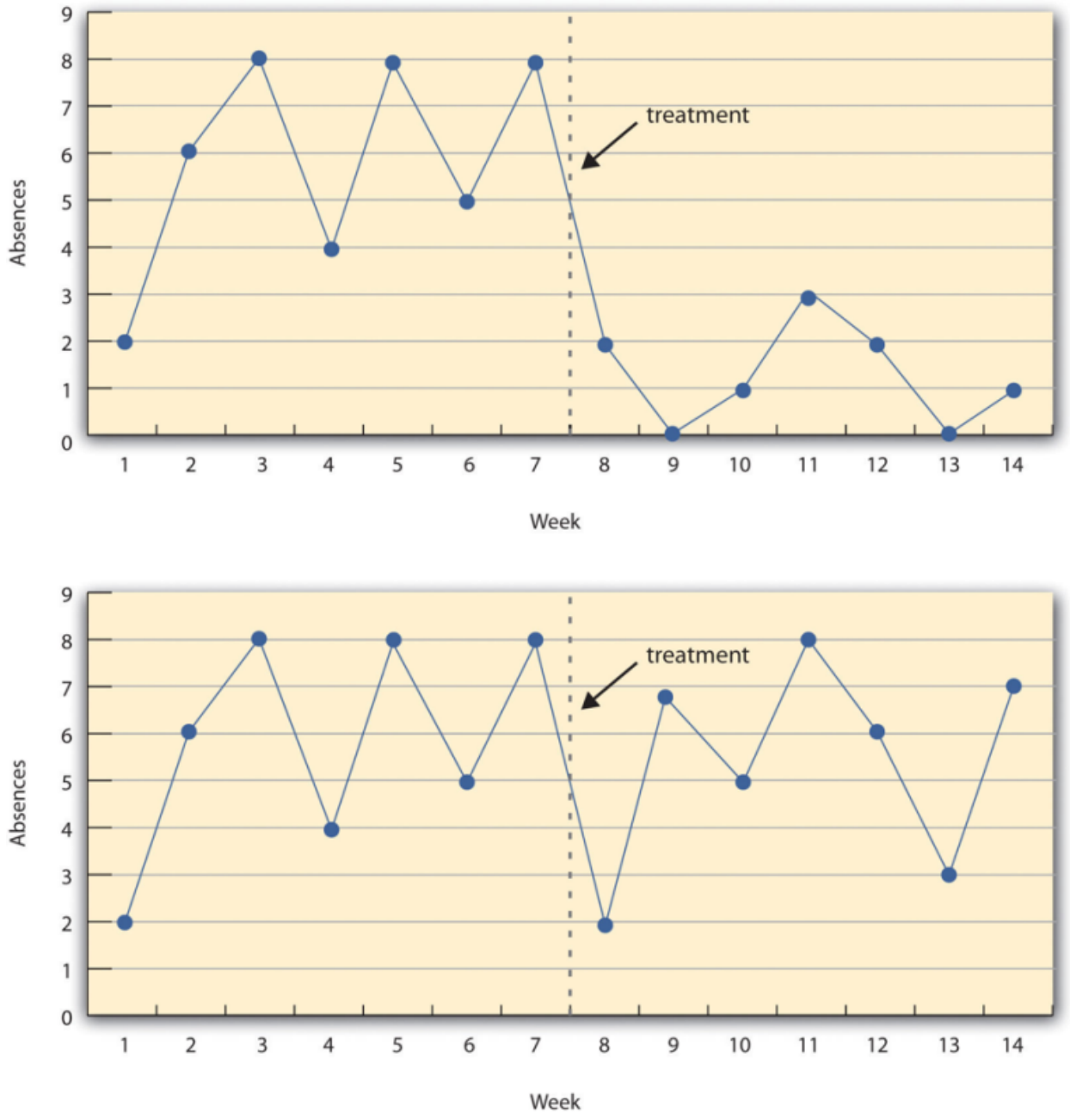


Figure 22.3: The figure contains two tan line charts labeled Week (x-axis 1–14) and Absences (y-axis 0–9). In both, a vertical dashed line at week 8 is annotated "treatment." Top panel: Before week 8, absences vary between about 2 and 8. After treatment, absences fall sharply—0 at weeks 9–10, then 1–3 with one week at 0—indicating a strong positive effect. Bottom panel: Absences before week 8 fluctuate between roughly 2 and 8. After treatment, the same up-and-down pattern continues with no downward shift, indicating little or no effect.

Figure 22.3 *A Hypothetical Interrupted Time-Series Design. The top panel shows data that suggest that the treatment caused a reduction in absences. The bottom panel shows data that suggest that it did not*[3]

22.2 Non-Equivalent Groups Designs

Recall that random assignment helps create groups that are similar before the study begins. Although the groups will never be perfectly identical, random assignment makes it less likely that they differ in systematic ways, especially when the sample size is large. Thus, researchers consider them to be equivalent. When participants are *not* randomly assigned to conditions, however, the resulting groups are likely to be dissimilar in some ways. For this reason, researchers consider them to be non-equivalent. A ***non-equivalent groups design***, then, is a between-subjects design in which participants have not been randomly assigned to conditions. There are several types of non-equivalent groups designs we will consider.

22.2.1 Posttest Only Non-Equivalent Groups Design

The first non-equivalent groups design we will consider is the ***posttest only non-equivalent groups design.*** In this design, participants in one group are exposed to a treatment, a non-equivalent group is not exposed to the treatment, and then the two groups are compared.

One way to illustrate the design is in Table 22.2. Here, we have two groups that were not randomly assigned to groups (N_1 and N_2). Group N_1 receives treatment X. Lastly, both groups complete posttest O_1.

Table 22.2 *Posttest Only Nonequivalent Groups Design*

Group	Treatment	Posttest
N_1	X	O_1
N_2		O_1

Imagine, for example, a researcher who wants to evaluate a new method of teaching fractions to third graders. One way would be to conduct a study with a treatment group consisting of one class of third-grade students and a control group consisting of another class of third-grade students. This design would be a non-equivalent groups design because the students are not randomly assigned to classes by the researcher, which means there could be important differences between them. For example, the parents of higher-achieving or more motivated students might have been more likely to request that their children be assigned to Ms. Williams's class. Or the principal might have assigned the "troublemakers" to Mr. Jones's class because he is a stronger disciplinarian. Of course, the teachers' styles and even the classroom environments might be very different and might cause different levels of achievement or motivation among

the students. If at the end of the study there was a difference in the two classes' knowledge of fractions, it might have been caused by the difference between the teaching methods—but it might have been caused by any of these confounding variables.

Of course, researchers using a posttest only non-equivalent groups design can take steps to ensure that their groups are as similar as possible. In the present example, the researcher could try to select two classes at the same school, where the students in the two classes have similar scores on a standardized math test and the teachers are the same sex, are close in age, and have similar teaching styles. Taking such steps would increase the internal validity of the study because it would eliminate some of the most important confounding variables. But without true random assignment of the students to conditions, there remains the possibility of other important confounding variables that the researcher was not able to control.

22.2.2 Pretest-Posttest Non-Equivalent Groups Design

Another way to improve upon the posttest only non-equivalent groups design is to add a pretest. In the ***pretest-posttest non-equivalent groups design***, there is a treatment group that is given a pretest, receives a treatment, and then is given a posttest. But at the same time, there is a non-equivalent control group that is given a pretest, does not receive the treatment, and then is given a posttest. The question, then, is not simply whether participants who receive the treatment improve, but whether they improve *more* than participants who do not receive the treatment.

One way to illustrate the design is in Table 22.3. Here, we have two groups that were not randomly assigned to groups (N_1 and N_2) that complete a pretest (O_1). Then, group N_1 receives treatment X. Lastly, both groups complete posttest O_2.

Table 22.3 *Nonequivalent Pretest-Posttest Design* [4]

Group	Pretest	Intervention	Posttest
N_1	O_1	X	O_2
N_2	O_1		O_2

Imagine, for example, that students in one school are given a pretest on their attitudes toward drugs, then are exposed to an anti-drug program, and finally, are given a posttest. Students in a similar school are given the pretest, not exposed to an anti-drug program, and finally, are given a posttest. Again, if students in the treatment condition become more negative toward drugs, this change in attitude could be an effect of the treatment, but it could also be a matter of history or maturation. If it really is an effect of the treatment, then students in the treatment condition should become more negative than students in the control condition. But if it is a matter of history (e.g., news of a celebrity drug overdose) or maturation (e.g., improved reasoning), then students in the two conditions would be likely to show similar amounts of change. This type of design does not completely eliminate the possibility of confounding

variables, however. Something could occur at one of the schools but not the other (e.g., a student drug overdose), so students at the first school would be affected by it while students at the other school would not.

Returning to the example of evaluating a new measure of teaching third graders, this study could be improved by adding a pretest of students' knowledge of fractions. The changes in scores from pretest to posttest would then be evaluated and compared across conditions to determine whether one group demonstrated a bigger improvement in knowledge of fractions than another. Of course, the teachers' styles and even the classroom environments might still be very different and might cause different levels of achievement or motivation among the students that are independent of the teaching intervention. Once again, differential history also represents a potential threat to internal validity. If asbestos is found in one of the schools causing it to be shut down for a month then this interruption in teaching could produce a difference across groups on posttest scores.

If participants in this kind of design are randomly assigned to conditions, it becomes a true between-groups experiment rather than a quasi-experiment. In fact, it is the kind of experiment that Eysenck called for—and that has now been conducted many times—to demonstrate the effectiveness of psychotherapy.

22.2.3 Interrupted Time-Series Design with Non-Equivalent Groups

One way to improve upon the interrupted time-series design is to add a control group. The ***interrupted time-series design with non-equivalent groups*** involves taking a set of measurements at intervals over a period of time, both before and after an intervention of interest in two or more non-equivalent groups. Once again, consider the manufacturing company that measures its workers' productivity each week for a year before and after reducing work shifts from 10 hours to 8 hours. This design could be improved by locating another manufacturing company who does not plan to change their shift length and using them as a non-equivalent control group. If productivity increased rather quickly after the shortening of the work shifts in the treatment group, but productivity remained consistent in the control group, then this provides better evidence for the effectiveness of the treatment.

Similarly, in the example of examining the effects of taking attendance on student absences in a research methods course, the design could be improved by using students in another section of the research methods course as a control group. If a consistently higher number of absences was found in the treatment group before the intervention, followed by a sustained drop in absences after the treatment, while the non-equivalent control group showed consistently high absences across the semester, then this would provide superior evidence for the effectiveness of the treatment in reducing absences.

22.2.4 Pretest-Posttest Design with Switching Replication

Some of these non-equivalent control group designs can be further improved by adding a switching replication. Using a ***pretest-posttest design with switching replication design***, non-equivalent groups are administered a pretest of the dependent variable, then one group receives a treatment while a non-equivalent control group does not receive a treatment, the dependent variable is assessed again, and then the treatment is added to the control group, and finally the dependent variable is assessed one last time.

One way to illustrate the design is in Table 22.4. Here, we have two groups that are not randomly assigned to groups (N_1 and N_2) that complete a pretest (O_1). Then, group N_1 receives treatment X and both groups complete posttest O_2. After that, group N_2 receives treatment X, while N_1 continues to receive the treatment. Finally, both groups complete posttest O_2. Overall, participants are measured on the dependent variable three times.

Table 22.4 *Pretest-Posttest Design with Switching Replication Design*

Group	**Pretest**	**Treatment**	**Posttest$_1$**	**Treatment**	**Posttest$_2$**
N_1	O_1	X	O_2	X	O_3
N_2	O_1		O_2	X	O_3

As a concrete example, let's say we wanted to introduce an exercise intervention for the treatment of depression. We recruit one group of patients experiencing depression and a non-equivalent control group of students experiencing depression. We first measure depression levels in both groups, and then we introduce the exercise intervention to the patients experiencing depression, but we hold off on introducing the treatment to the students. We then measure depression levels in both groups. If the treatment is effective, we should see a reduction in the depression levels of the patients (who received the treatment) but not in the students (who have not yet received the treatment). Finally, while the group of patients continues to engage in the treatment, we would introduce the treatment to the students with depression. Now and only now should we see the students' levels of depression decrease.

One of the strengths of this design is that it includes a built-in replication. In the example given, we would get evidence for the efficacy of the treatment in two different samples (patients and students). Another strength of this design is that it provides more control over history effects. It becomes rather unlikely that some outside event would perfectly coincide with the introduction of the treatment in the first group and with the delayed introduction of the treatment in the second group. For instance, if a change in the weather occurred when we first introduced the treatment to the patients, and this explained their reductions in depression the second time that depression was measured, then we would see depression levels decrease in both groups. Similarly, the switching replication helps to control for maturation and instrumentation. Both groups would be expected to show the same rates of spontaneous remission of depression, and if the instrument for assessing depression happened to change at some point in the study,

the change would be consistent across both groups. Of course, demand characteristics, placebo effects, and experimenter expectancy effects can still be problems. But they can be controlled for using some of the methods described in Chapter 21.

22.2.5 Switching Replication with Treatment Removal Design

In a basic pretest-posttest design with switching replication, the first group receives a treatment, and the second group receives the same treatment a little bit later on (while the initial group continues to receive the treatment). In contrast, in a ***switching replication with treatment removal design***, the treatment is removed from the first group when it is added to the second group. In other words, the original design is repeated or replicated temporally with treatment/control roles switched between the two groups. By the end of the study, all participants will have received the treatment *either* during the first or the second phase. This design is most feasible in organizational contexts where organizational programs (such as employee training) are implemented in a phased manner or are repeated at regular intervals.

One way to illustrate the design is in Table 22.5. Here, we have two groups that were not randomly assigned to groups (N_1 and N_2) that complete a pretest (O_1). Then, group N_1 receives treatment X and both groups complete posttest O_2. After that, group N_2 receives treatment X (while N_1 no longer receives treatment X). Finally, both groups complete posttest O_2. Overall, participants are measured on the dependent variable three times.

Table 22.5 *Switching Replication with Treatment Removal Design* [5]

Group	**Pretest**	**Treatment**	**Posttest$_1$**	**Treatment**	**Posttest$_2$**
N_1	O_1	X	O_2		O_3
N_2	O_1		O_2	X	O_3

Once again, let's assume we first measure the depression levels of patients with depression and students with depression. Then we introduce the exercise intervention to only the patients. After they have been exposed to the exercise intervention for a week, we assess depression levels again in both groups. If the intervention is effective, then we should see depression levels decrease in the patient group but not the student group (because the students haven't received the treatment yet). Next, we would remove the treatment from the group of patients with depression. So we would tell them to stop exercising. At the same time, we would tell the student group to start exercising. After a week of the students exercising and the patients not exercising, we would reassess depression levels. Now, if the intervention is effective, we should see that the depression levels have decreased in the student group but that they have increased in the patient group (because they are no longer exercising).

Demonstrating a treatment effect in two groups staggered over time and demonstrating the reversal of the treatment effect after the treatment has been removed can provide strong evidence for the efficacy of the treatment. In addition to providing evidence for the replicability

of the findings, this design can also provide evidence for whether the treatment continues to show effects after it has been withdrawn.

There are several takeaways from this discussion on quasi-experimental designs. In these types of designs, the independent variable is manipulated, but there is no random assignment to individual conditions of an independent variable. Because of the manipulation of the independent variable, the directionality problem is addressed. However, because of the lack of random assignment, it does not eliminate the problem of confounding variables. For these reasons, quasi-experimental designs are generally higher in internal validity than non-experimental designs but lower than true experiments. It is best to be cautious about making strong causal inferences from quasi-experimental designs.

To briefly many of the different types of quasi-experimental designs discussed in this chapter, see the video "Quasi-experimental designs".

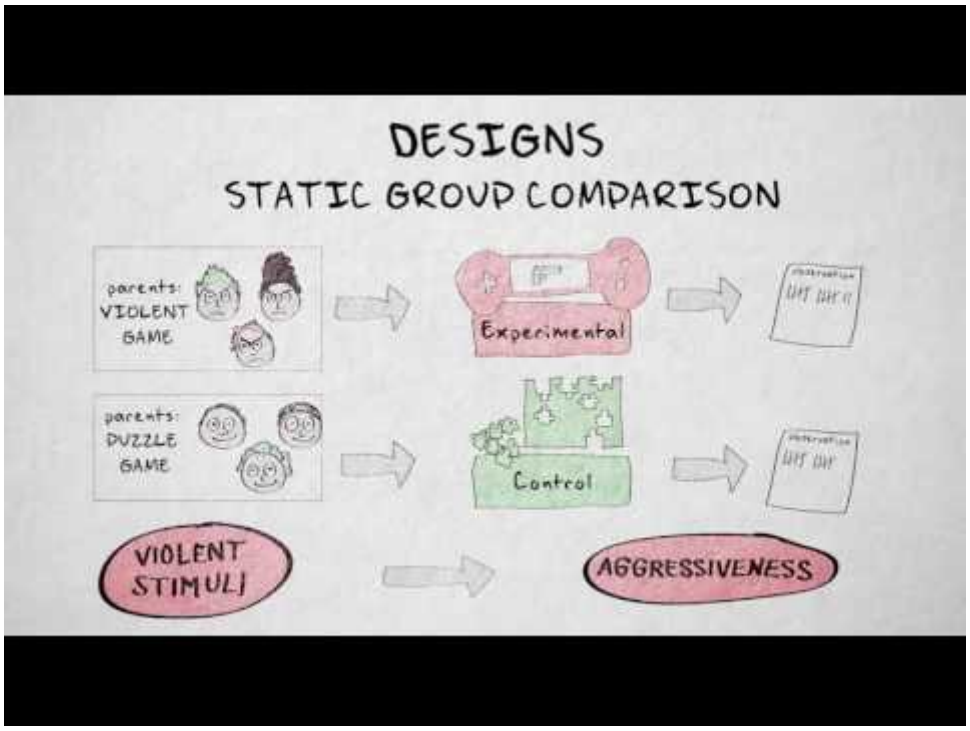


Figure 22.4: 3.9 Quasi-experimental designs | Quantitative methods | Research Designs | UvA

Activity: Addressing Immigration Stress

You are a researcher working with a local school district who wants to provide mental health services to
their Latine immigrant students who have been reporting high levels of stress after anti-

immigrant
sentiment has been occurring across the country. You provide students with the Hispanic Stress
Inventory- Adolescent
Version
as a pre-test and then provide an 8-week mental health intervention. At the conclusion of the
intervention, you give the students the HSI-A as a post-test. When you analyze the data, you find that
students report significantly lower stress levels after the completion of the intervention.

1. Can you say that the intervention caused the reduction in stress levels?
2. What threats to internal validity might impact the change in stress levels?
3. What can you do to mitigate the threats to internal validity for a one-group pretest-posttest design?
4. You decide to compare immigration stress levels across two schools in the district. Following a
   similar procedure as mentioned above, describe the benefits of adding a control group to your design.
5. What step would be needed to move this design from quasi-experimental to experimental?

22.3 Media Attributions

[1] D'Costa, S., Ukeye, M., O'Neil, M., & Anguiano, R. (n.d.). Critical Research Methods in Psychology. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

[2]Jhangiani, R.S., Chiang, I-C.A., Cuttler, C., & Leighton, D.C. (2019). Research methods in psychology. 4th edition. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

[3] D'Costa, S., Ukeye, M., O'Neil, M., & Anguiano, R. (no date). Critical Research Methods in Psychology. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

[4-5] Bhattacherjee, A. (n.d.). Chapter 10 Experimental Research. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Modified by current authors.

22.4 Text Attributions

Bhattacherjee, A. (n.d.). Chapter 10 Experimental Research. In *[Research Methods for the Social Sciences]*. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Modified by current authors.

D'Costa, S., Ukeye, M., O'Neil, M., & Anguiano, R. (no date). Critical Research Methods in Psychology. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Jhangiani, R.S., Chiang, I-C.A., Cuttler, C., & Leighton, D.C. (2019). Research methods in psychology. 4th edition. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

22.5 References

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues in field settings. Boston, MA: Houghton Mifflin.

Eysenck, H. J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology, 16*, 319–324.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore, MD: Johns Hopkins University Press.

23 Chapter 23: Factorial Designs

Just as it is common for studies in psychology to include multiple levels of a single independent variable (placebo, new drug, old drug), it is also common for them to include multiple independent variables. Researchers' inclusion of multiple independent variables in one experiment is further illustrated by the following actual titles from various professional journals:

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

Just as including multiple levels of a single independent variable allows one to answer more sophisticated research questions, so too does including multiple independent variables in the same experiment. But including multiple independent variables also allows the researcher to answer questions about whether the effect of one independent variable depends on the level of another. This is referred to as an interaction between the independent variables. Interactions are often among the most interesting results in psychological research.

By far the most common approach to including multiple independent variables (which are often called factors) in an experiment is the factorial design. In a ***factorial design***, each level of one independent variable is combined with each level of the others to produce all possible combinations. Each combination, then, becomes a condition in the experiment. Imagine, for example, an experiment on the effect of cell phone use (yes vs. no) and time of day (day vs. night) on driving ability. This is shown in the ***factorial design table*** in Figure 23.1. The columns of the table represent cell phone use, and the rows represent time of day. The four cells of the table represent the four possible combinations or conditions: using a cell phone during the day, not using a cell phone during the day, using a cell phone at night, and not using a cell phone at night. This particular design is referred to as a 2×2 (read "two-by-two") factorial design because it combines two variables, each of which has two levels.

If one of the independent variables had a third level (e.g., using a handheld cell phone, using a hands-free cell phone, and not using a cell phone), then it would be a 3×2 factorial design,

and there would be six distinct conditions. Notice that the number of possible conditions is the *product* of the numbers of levels. A 2 × 2 factorial design has four conditions, a 3 × 2 factorial design has six conditions, a 4 × 5 factorial design would have 20 conditions, and so on. Also notice that each number in the notation represents one factor, one independent variable. So by looking at *how many numbers are in the notation*, you can determine how many independent variables there are in the experiment. 2 x 2, 3 x 3, and 2 x 3 designs all have two numbers in the notation and therefore all have two independent variables. The *numerical value of each of the numbers* represents the number of levels of each independent variable. A 2 means that the independent variable has two levels, a 3 means that the independent variable has three levels, a 4 means it has four levels, etc. To illustrate, a 3 x 3 design has two independent variables, each with three levels, while a 2 x 2 x 2 design has three independent variables, each with two levels.

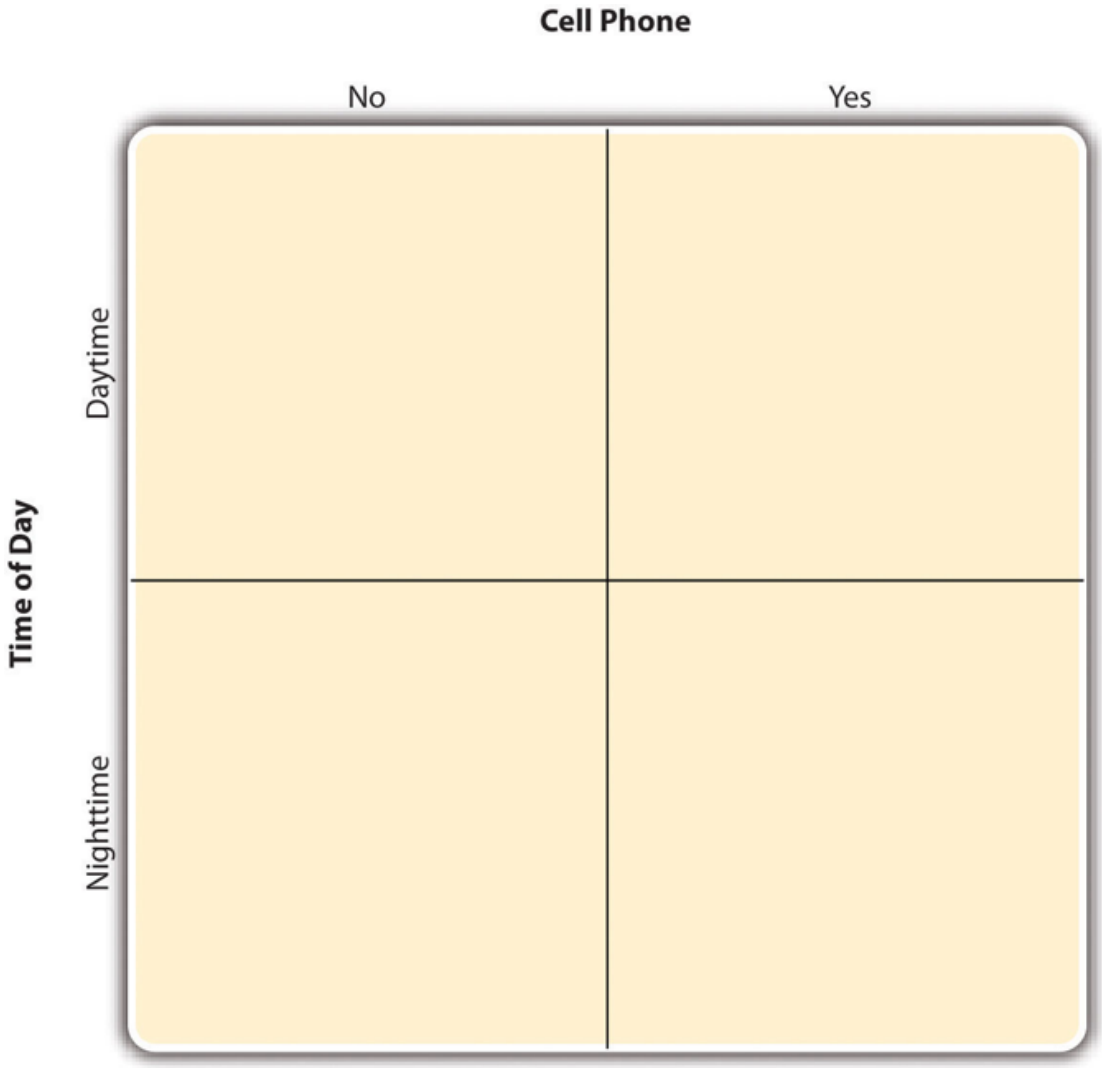


Figure 23.1: Diagram of a 2 × 2 factorial design table. The columns represent cell phone use (No and Yes), and the rows represent time of day (Daytime and Nighttime). The four empty cells represent the four combinations of the two independent variables: daytime/no cell phone, daytime/cell phone, nighttime/no cell phone, and nighttime/cell phone.

Figure 23.1 *Factorial Design Table Representing a 2 × 2 Factorial Design*[1]

In principle, factorial designs can include any number of independent variables with any number of levels. For example, an experiment could include the type of psychotherapy (cognitive vs. behavioral), the length of the psychotherapy (2 weeks vs. 2 months), and the sex of the psychotherapist (female vs. male). This would be a 2 × 2 × 2 factorial design and would

have eight conditions. Figure 23.2 shows one way to represent this design. In practice, it is unusual for there to be more than three independent variables with more than two or three levels each. This is for at least two reasons: For one, the number of conditions can quickly become unmanageable. For example, adding a fourth independent variable with three levels (e.g., therapist experience: low vs. medium vs. high) to the current example would make it a 2 $\times$ 2 $\times$ 2 $\times$ 3 factorial design with 24 distinct conditions. Second, the number of participants required to populate all of these conditions (while maintaining a reasonable ability to detect a real underlying effect) can render the design unfeasible. For clarity, the examples that follow focus on designs with two independent variables. The general principles discussed here extend in a straightforward way to more complex factorial designs.

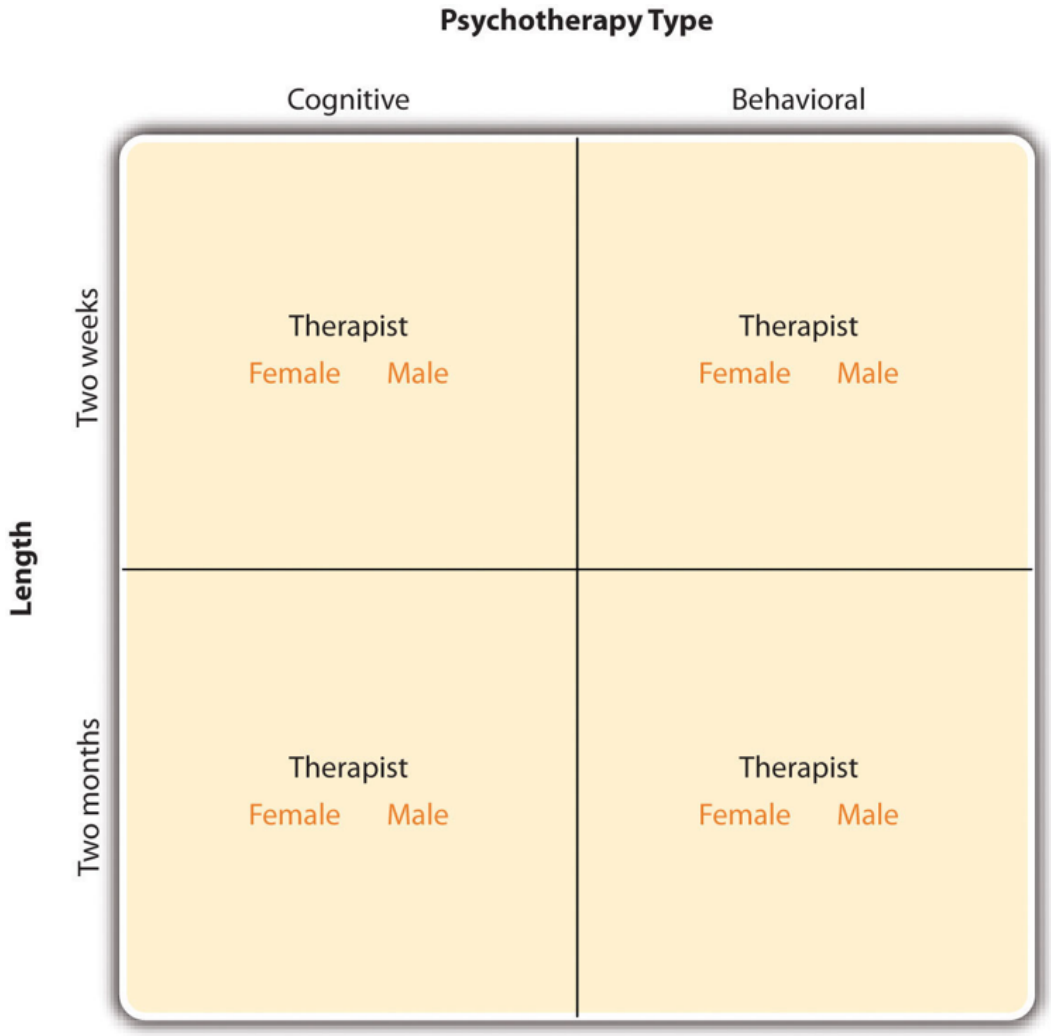


Figure 23.2: Diagram of a 2 × 2 × 2 factorial design. The columns represent therapy type (Cognitive and Behavioral), and the rows represent treatment length (Two weeks and Two months). Within each of the four cells, participants are divided by therapist sex (Female and Male), creating eight unique experimental conditions.

Figure 23.2 *Factorial Design Table Representing a 2 × 2 × 2 Factorial Design*[2]

23.1 Assigning Participants to Conditions

In a simple between-subjects design, each participant is tested in only one condition. In a simple within-subjects design, each participant is tested in all conditions. In a factorial

experiment, the decision to take the between-subjects or within-subjects approach must be made separately for each independent variable. In a ***between-subjects factorial design***, all of the independent variables are manipulated between subjects. For example, all participants could be tested either while using a cell phone *or* while not using a cell phone and either during the day *or* during the night. This would mean that each participant would be tested in one and only one condition. In a within-subjects factorial design, all of the independent variables are manipulated within subjects. All participants could be tested both while using a cell phone *and* while not using a cell phone and both during the day *and* during the night. This would mean that each participant would need to be tested in all four conditions. The between-subjects design is conceptually simpler, avoids order/carryover effects, and minimizes the time and effort of each participant. The within-subjects design is more efficient for the researcher, controls extraneous participant variables, and can increase statistical power.

Since factorial designs have more than one independent variable, it is also possible to manipulate one independent variable between subjects and another within subjects. This is called a ***mixed factorial design***. For example, a researcher might choose to treat cell phone use as a within-subjects factor by testing the same participants both while using a cell phone and while not using a cell phone (while counterbalancing the order of these two conditions). But they might choose to treat time of day as a between-subjects factor by testing each participant either during the day or during the night (perhaps because this only requires them to come in for testing once). Thus, each participant in this mixed design would be tested in two of the four conditions.

Regardless of whether the design is between subjects, within subjects, or mixed, the actual assignment of participants to conditions or orders of conditions is typically done randomly.

23.2 Non-Manipulated Independent Variables

In many factorial designs, one of the independent variables is a ***non-manipulated independent variable***. The researcher measures it but does not manipulate it. One example is a study by Halle Brown and colleagues in which participants were exposed to several words that they were later asked to recall (Brown et al., 1999). The manipulated independent variable was the type of word. Some were negative health-related words (e.g., *tumor, coronary*), and others were not health related (e.g., *election, geometry*). The non-manipulated independent variable was whether participants scored high or low on a measure of hypochondriasis—the term used in the original study for what is now often called health anxiety, or excessive concern about bodily symptoms. Participants higher in health anxiety recalled more health-related words than participants lower in health anxiety, but the groups did not differ in their recall of non-health-related words.

Such studies are extremely common, and there are several points worth making about them. First, non-manipulated independent variables are usually participant variables (private body consciousness, health anxiety, self-esteem, gender, and so on), and as such, they are by definition

between-subjects factors. For example, people are classified as either lower or higher in health anxiety; they cannot be tested in both of these conditions. Second, such studies are generally considered to be experiments as long as at least one independent variable is manipulated, regardless of how many non-manipulated independent variables are included. Third, it is important to remember that causal conclusions can only be drawn about the manipulated independent variable. Thus, it is important to be aware of which variables in a study are manipulated and which are not.

23.2.1 Non-Experimental Studies With Factorial Designs

Factorial experiments can include manipulated independent variables or a combination of manipulated and non-manipulated independent variables. But factorial designs can also include *only* non-manipulated independent variables, in which case they are no longer experiments but are instead non-experimental in nature. Consider a hypothetical study in which a researcher simply measures both the moods and the self-esteem of several participants—categorizing them as having either a positive or negative mood and as being either high or low in self-esteem—along with their willingness to have unprotected sexual intercourse. This can be conceptualized as a 2 × 2 factorial design with mood (positive vs. negative) and self-esteem (high vs. low) as non-manipulated between-subjects factors. Willingness to have unprotected sex is the dependent variable.

Because neither independent variable in this example was manipulated, it is a non-experimental study rather than an experiment. This is important because, as always, one must be cautious about inferring causality from non-experimental studies because of the directionality and third-variable problems. For example, an effect of participants' moods on their willingness to have unprotected sex might be caused by any other variable that happens to be correlated with their moods.

23.3 Interpreting the Results of a Factorial Experiment

23.3.1 Graphing the Results of Factorial Experiments

The results of factorial experiments with two independent variables can be graphed by representing one independent variable on the x-axis and representing the other by using different colored bars or lines. (The y-axis is always reserved for the dependent variable.) Figure 23.3 shows results for two hypothetical factorial experiments. The top panel shows the results of a 2 × 2 design. Time of day (day vs. night) is represented by different locations on the x-axis, and cell phone use (no vs. yes) is represented by different-colored bars. (It would also be possible to represent cell phone use on the x-axis and time of day as different-colored bars. The choice comes down to which way seems to communicate the results most clearly.) The bottom panel of Figure 23.3 shows the results of a 4 × 2 design in which one of the variables is quantitative.

This variable, psychotherapy length, is represented along the x-axis, and the other variable (psychotherapy type) is represented by differently formatted lines. This is a line graph rather than a bar graph because the variable on the x-axis is quantitative with a small number of distinct levels. Line graphs are also appropriate when representing measurements made over a time interval (also referred to as time series information) on the x-axis.

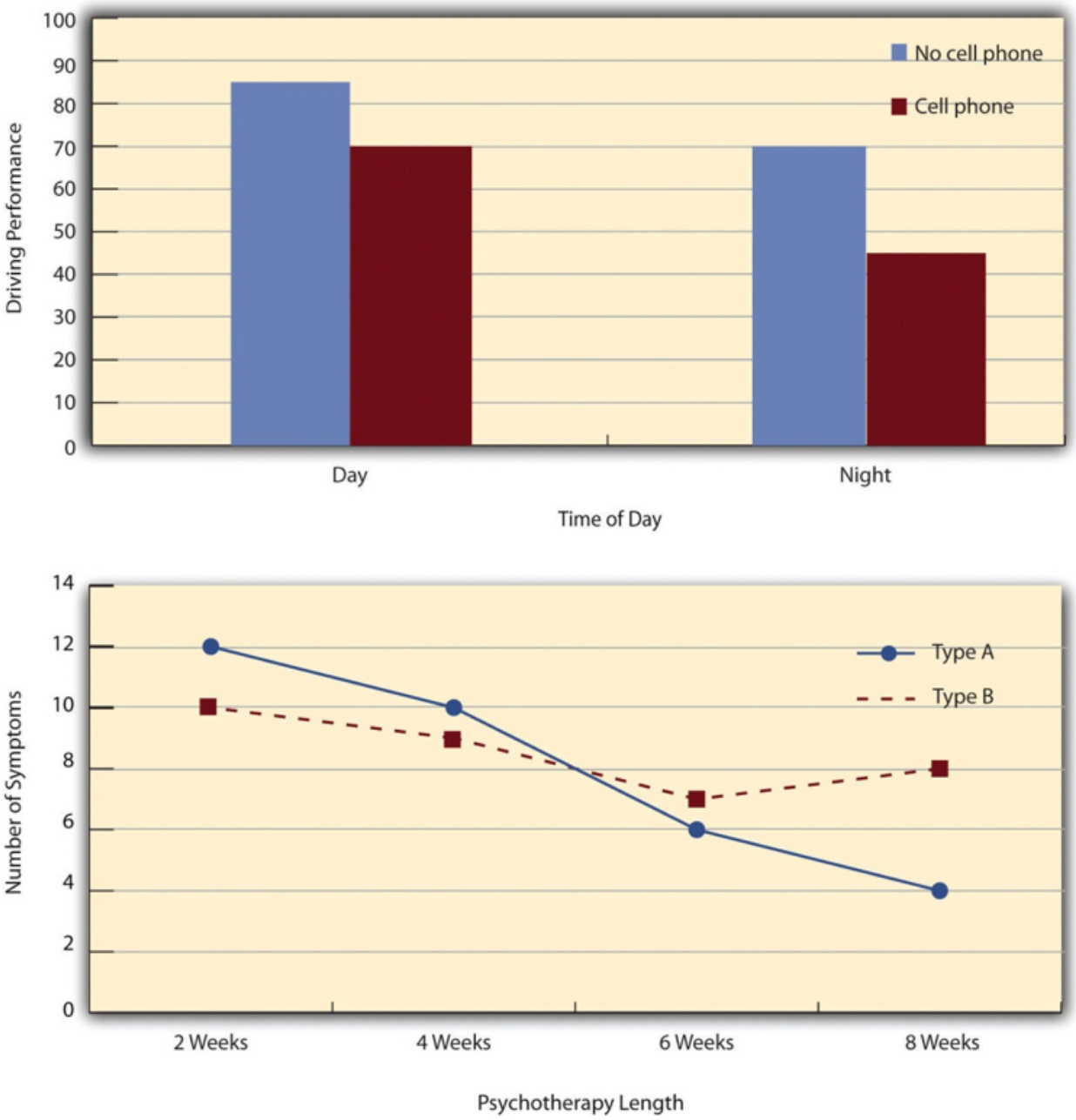


Figure 23.3: Two graphs illustrating results from factorial designs. The top bar graph shows driving performance by time of day (day and night) and cell phone use (cell phone vs. no cell phone). The lower line graph shows symptom counts at four psychotherapy durations ranging from 2 to 8 weeks for Therapy Types A and B.

Figure 23.3 *Two Ways to Plot the Results of a Factorial Experiment With Two Independent*

Variables[3]

23.3.2 Main Effects

In factorial designs, there are three kinds of results that are of interest: main effects, interaction effects, and simple effects. A ***main effect*** is the effect of one independent variable on the dependent variable—averaging across the levels of the other independent variable. Thus, there is one main effect to consider for each independent variable in the study. The top panel of Figure 23.3 shows a main effect of cell phone use because driving performance was better, on average, when participants were not using cell phones than when they were. The blue bars are, on average, higher than the red bars. It also shows a main effect of time of day because driving performance was better during the day than during the night—both when participants were using cell phones and when they were not. Main effects are independent of each other in the sense that whether or not there is a main effect of one independent variable says nothing about whether or not there is a main effect of the other. The bottom panel of Figure 23.3, for example, shows a clear main effect of psychotherapy length. The longer the psychotherapy, the better it worked.

Table 23.1 presents ***marginal means***, the average scores for each level of one independent variable, calculated by averaging across all levels of the other independent variable or variables. In our study of driving performance, to visually identify the main effect of Time of Day, compare the marginal mean for Day with the marginal mean for Night. Each marginal mean is calculated by averaging across the Cell Phone and No Cell Phone conditions.

Table 23.1: Table 23.1. Marginal means for the driving performance study[4]. Marginal mean driving-performance scores for day and night conditions, split by cell phone use. Scores are 70 and 45 at night and 85 and 70 during the day; the marginal means are 57.5 and 77.5, respectively.

Time of Day	No Cell Phone	Cell Phone	Marginal Mean
Night	70	45	57.5
Day	85	70	77.5

The difference between the Time of Day marginal means is: 77.5 – 57.5 = 20

Because the marginal means differ, the graph would show a main effect of Time of Day. Visually, the points or lines associated with Day would generally appear higher than those associated with Night.

23.3.3 Interactions

There is an ***interaction*** effect (or just "interaction") when the effect of one independent variable depends on the level of another. Although this might seem complicated, you already have an intuitive understanding of interactions. As an everyday example, assume your friend asks you to go to a movie with another friend. Your response to them is, "Well it depends on which movie you are going to see and who else is coming." You really want to see the big blockbuster summer hit but have little interest in seeing the cheesy romantic comedy. In other words, there is a main effect of type of movie on your decision. If your decision to go to see either of these movies further depends on who they are bringing with them, then there is an interaction. For instance, if you will go to see the cheesy romantic comedy if they bring their friend you want to get to know better, but you will not go to this movie if they bring anyone else, then there is an interaction. Interactions also appear in everyday learning. For example, a brief relaxation exercise might reduce test anxiety more for students who slept poorly than for students who slept well. In that case, the effect of the relaxation exercise would depend on sleep quality.

Let's now consider some examples of interactions from research. It probably would not surprise you to hear that the effect of receiving psychotherapy is stronger among people who are highly motivated to change than among people who are not motivated to change. This is an interaction because the effect of one independent variable (whether or not one receives psychotherapy) depends on the level of another (motivation to change).

In many studies, the primary research question is about an interaction. The study by Brown and her colleagues was inspired by the idea that people high in health anxiety may be especially attentive to negative health-related information. This led to the hypothesis that people high in health anxiety would recall negative health-related words more accurately than people low in health anxiety but would recall non-health-related words about as accurately as people low in health anxiety. And of course, this is exactly what happened in this study.

23.3.4 Types of Interactions

The effect of one independent variable can depend on the level of the other in several different ways. First, there can be ***spreading interactions***. Examples of spreading interactions are shown in the top two panels of Figure 23.4. In the top panel, independent variable "B" has an effect at level 1 of independent variable "A" (there is a difference in the height of the blue and red bars on the left side of the graph) but no effect at level 2 of independent variable "A." (There is no difference in the height of the blue and red bars on the right side of the graph.) In the middle panel, independent variable "B" has a stronger effect at level 1 of independent variable "A" than at level 2 (there is a larger difference in the height of the blue and red bars on the left side of the graph and a smaller difference in the height of the blue and red bars on the right side of the graph). This is like the hypothetical driving example where there was a strong effect of using a cell phone at night and a weaker effect of using a cell phone during the

day. So to summarize, for spreading interactions there is an effect of one independent variable at one level of the other independent variable and there is either a weak effect or no effect of that independent variable at the other level of the other independent variable.

The second type of interaction that can be found is a ***cross-over interaction.*** A cross-over interaction is depicted in the bottom panel of Figure 23.4. Independent variable "B" again has an effect at both levels of independent variable "A," but the effects are in opposite directions. Another example of a crossover interaction comes from a study by Kathy Gilliland on the effect of caffeine on the verbal test scores of introverts and extraverts (Gilliland, 1980). Introverts perform better than extraverts when they have not ingested any caffeine. But extraverts perform better than introverts when they have ingested 4 mg of caffeine per kilogram of body weight.

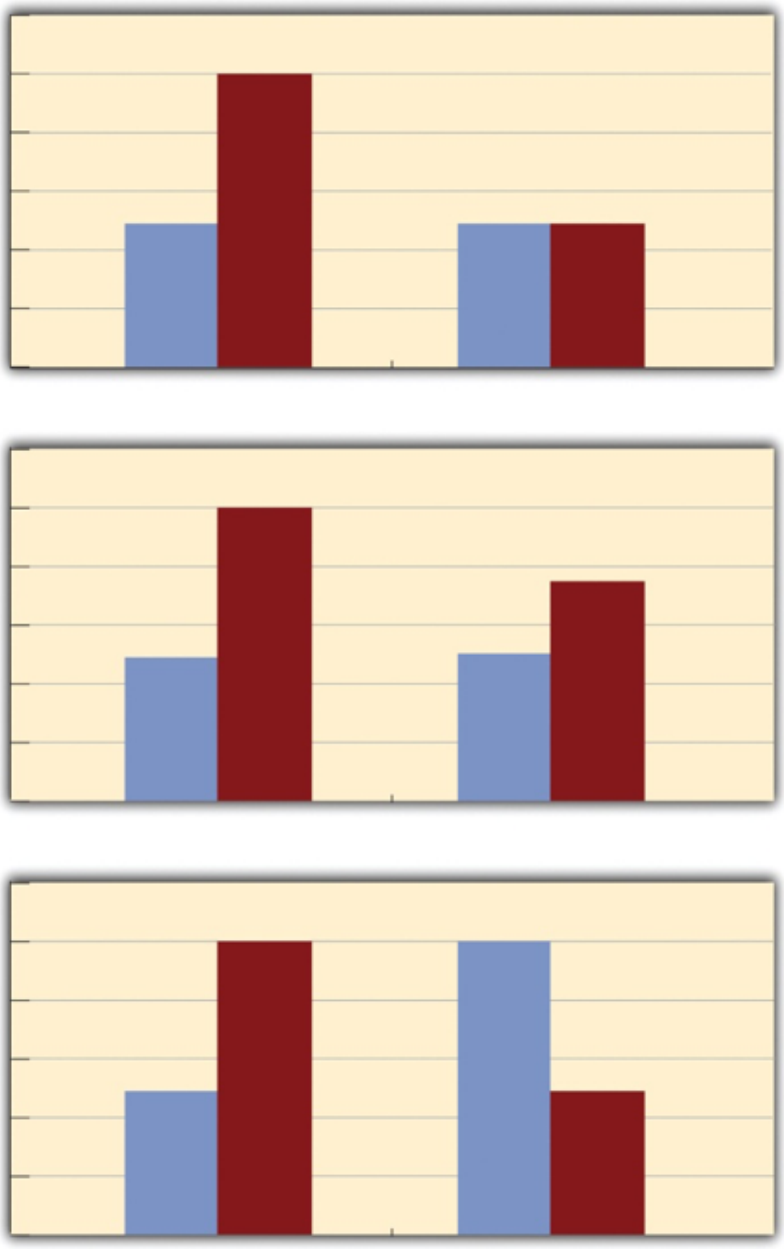


Figure 23.4: Three stacked bar graphs illustrating different interaction effects in factorial designs. The top graph shows an interaction at one level of a second independent variable but not the other. The middle graph shows one independent variable having a stronger effect at one level of the second variable. The bottom graph shows opposite effects of one independent variable across the two levels of the second variable.

Figure 23.4 *Bar Graphs Showing Three Types of Interactions*[5]

Figure 23.5 shows examples of these same kinds of interactions when one of the independent variables is quantitative and the results are plotted in a line graph. Note that the top two figures depict the two kinds of spreading interactions that can be found while the bottom figure depicts a crossover interaction (the two lines literally "cross over" each other).

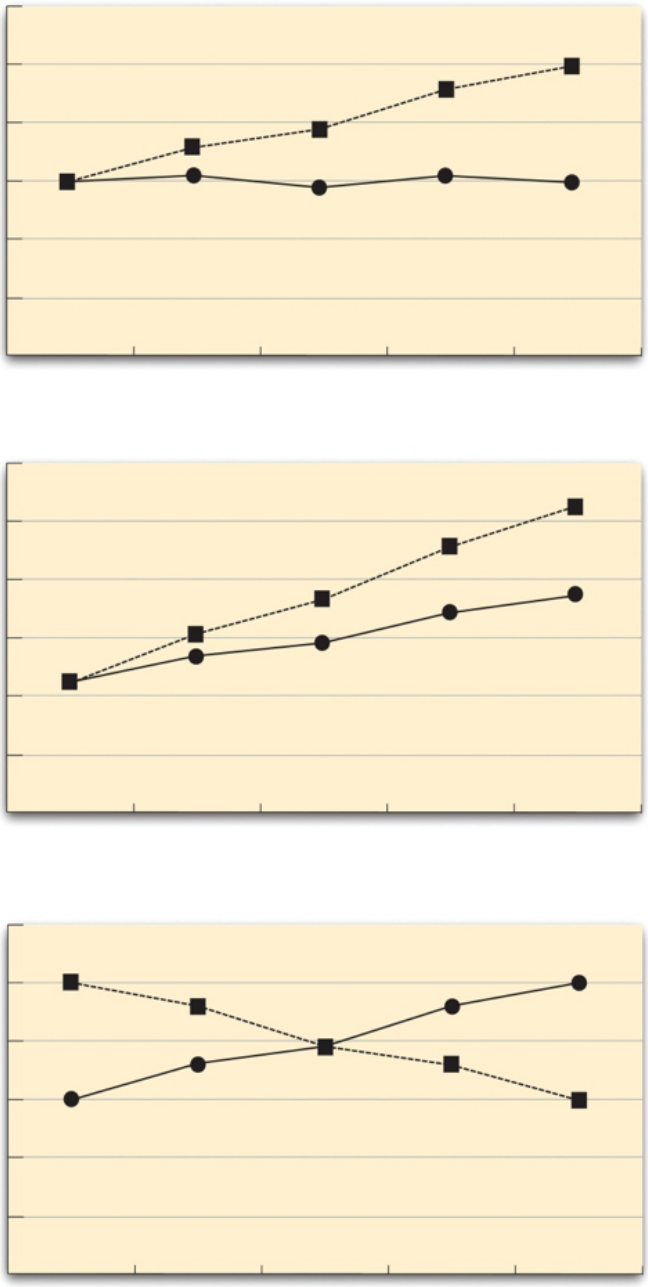


Figure 23.5: Three stacked line graphs illustrating interaction effects. The top graph shows an interaction at one level of a second independent variable but not the other. The middle graph shows one independent variable having a stronger effect at one level of the second variable. The bottom graph shows a crossover interaction, where the two lines cross, indicating opposite effects across levels of the second independent variable.

Figure 23.5 *Line Graphs Showing Different Types of Interactions*[6]

23.3.5 Simple Effects

When researchers find an interaction, it suggests that the main effects may be a bit misleading. Think of the example of a crossover interaction where introverts were found to perform better on a test of verbal ability than extraverts when they had not ingested any caffeine, but extraverts were found to perform better than introverts when they had ingested 4 mg of caffeine per kilogram of body weight. To examine the main effect of caffeine consumption, the researchers would have averaged across introversion and extraversion and simply looked at whether, overall, those who ingested caffeine had better or worse verbal ability. Because the positive effect of caffeine on extraverts would be wiped out by the negative effects of caffeine on the introverts, no main effect of caffeine consumption would have been found. Similarly, to examine the main effect of personality, the researchers would have averaged across the levels of the caffeine variable to look at the effects of personality (introversion vs. extraversion) independent of caffeine. In this case, the positive effects of extraversion in the caffeine condition would be wiped out by the negative effects of extraversion in the no caffeine condition. Does the absence of any main effects mean that there is no effect of caffeine and no effect of personality? No, of course not. The presence of the interaction indicates that the story is more complicated, that the effects of caffeine on verbal ability depend on personality. This is where simple effects come into play.

Simple effects are a way of breaking down the interaction to figure out precisely what is going on. An interaction simply informs us that the effects of at least one independent variable depend on the level of another independent variable. Whenever an interaction is detected, researchers need to conduct additional analyses to determine where that interaction is coming from. Of course, one may be able to visualize and interpret the interaction on a graph, but a simple effects analysis provides researchers with a more sophisticated means of breaking down the interaction. Specifically, a ***simple effects analysis*** allows researchers to determine the effects of each independent variable at each level of the other independent variable. So while the researchers would average across the two levels of the personality variable to examine the effects of caffeine on verbal test performance in a main effects analysis, for a simple effects analysis the researchers would examine the effects of caffeine in introverts and then examine the effects of caffeine in extraverts. The researchers also examined the effects of personality in the no caffeine condition and found that in this condition introverts performed better than extraverts. Finally, they examined the effects of personality in the caffeine condition and found that extraverts performed better than introverts in this condition. For a 2 x 2 design like this, there will be two main effects the researchers can explore, one interaction effect, and four simple effects (i.e., the effect of caffeine among introverts, the effect of caffeine among extraverts, the effect of personality in the no-caffeine condition, and the effect of personality in the caffeine condition).

Brown and colleagues found an interaction between type of words (health related or not health related) and health anxiety (high or low) on word recall. To break down this interaction using simple effects analyses, they examined the effect of health anxiety at each level of word type. Specifically, they examined the effect of health anxiety on recall of health-related words and then they subsequently examined the effect of health anxiety on recall of non-health related words. They found that people high in health anxiety were able to recall more health-related words than people low in health anxiety. In contrast, there was no effect of health anxiety on the recall of non-health related words.

Examining simple effects provides a way to clarify an interaction. Researchers often conduct these analyses after finding an interaction, but they may also plan simple-effects tests in advance when those comparisons follow directly from their hypotheses. When an interaction is not statistically significant, researchers generally focus on the main effects while recognizing that a nonsignificant result does not prove that the effect is identical across all levels of the other independent variable. To summarize, rather than averaging across the levels of the other independent variable, as is done in a main effects analysis, simple effects analyses are used to examine the effects of each independent variable at each level of the other independent variable(s). So a researcher using a 2×2 design with four conditions would need to look at 2 main effects, 1 interaction effect, and 4 simple effects. A researcher using a 2×3 design with six conditions would need to look at 2 main effects, 1 interaction effect, and 5 simple effects, while a researcher using a 3×3 design with nine conditions would need to look at 2 main effects, 1 interaction effect, and 6 simple effects. As you can see, while the number of main effects and interaction effects depends simply on the number of independent variables included (one main effect can be explored for each independent variable, one interaction effect can be explored for each combination of two or more independent variables), the number of simple effects analyses depends on the number of levels of the independent variables (because a separate analysis of each independent variable is conducted at each level of the other independent variable).

23.4 Advantages and Disadvantages of Factorial Designs

Factorial designs offer a powerful way to study two or more factors in a single study. Because the design includes every planned combination of factor levels, researchers can evaluate both ***main effects*** and ***interactions***. These strengths come with practical and interpretive costs.

23.4.1 Advantages

One major advantage is efficiency. A single factorial study can answer questions about each factor and about how the factors work together; studying the factors in separate experiments would not reveal whether the effect of one factor changes across levels of another. This makes factorial designs especially useful when a theory predicts an interaction or when researchers want to identify the conditions under which an effect is stronger, weaker, or reversed.

Factorial designs can also produce a more realistic and nuanced account of behavior. Psychological outcomes are often influenced by several characteristics of a person, task, or situation at once. Including those factors in one design can reveal boundary conditions that a simpler study would miss. When all factors are manipulated and alternative explanations are controlled, the design can support causal conclusions about main effects and interactions. If a factor is only measured, causal conclusions cannot be made about that factor.

23.4.2 Disadvantages

One major disadvantage is the rapid growth in the number of conditions. A 2×2 design has four conditions, but a $2 \times 2 \times 3$ design has 12. As conditions multiply, researchers usually need more participants, time, and resources to maintain enough observations in each cell and adequate statistical power.

Factorial designs can also be harder to conduct and interpret. Within-subjects or mixed designs may require careful counterbalancing to control order and carryover effects, and interactions often require graphs and follow-up tests of simple effects. With three or more factors, higher-order interactions can be especially difficult to explain. Researchers should therefore choose factors and levels that directly address their hypotheses and plan the sample size and analysis before collecting data.

23.5 Media Attributions

[1-6] Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

23.6 Text Attributions

Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by the current authors.

23.7 References

Brown, H. D., Kosslyn, S. M., Delamater, B., Fama, A., & Barsky, A. J. (1999). Perceptual and memory biases for health-related information in hypochondriacal individuals. *Journal of Psychosomatic Research, 47*(1), 67–78. https://doi.org/10.1016/S0022-3999(99)00011-2

Gilliland, K. (1980). The interactive effect of introversion-extraversion with caffeine-induced arousal on verbal performance. *Journal of Research in Personality, 14*(4), 482–492. https://doi.org/10.1016/0092-6566(80)90006-9

24 Chapter 24: Single-*N* Designs

Researcher Vance Hall and his colleagues were faced with the challenge of increasing the extent to which six disruptive elementary school students stayed focused on their schoolwork (Hall et al., 1968). For each of several days, the researchers carefully recorded whether or not each student was doing schoolwork every 10 seconds during a 30-minute period. Once they had established this baseline, they introduced a treatment. The treatment was that when the student was doing schoolwork, the teacher gave him or her positive attention in the form of a comment like "good work" or a pat on the shoulder. The result was that all of the students dramatically increased their time spent on schoolwork and decreased their disruptive behavior during this treatment phase. For example, a student named Robbie originally spent 25% of his time on schoolwork and the other 75% "snapping rubber bands, playing with toys from his pocket, and talking and laughing with peers" (p. 3). During the treatment phase, however, he spent 71% of his time on schoolwork and only 29% on other activities. Finally, when the researchers had the teacher stop giving positive attention, the students all decreased their studying and increased their disruptive behavior. This confirmed that it was, in fact, the positive attention that was responsible for the increase in studying. This was one of the first studies to show that attending to positive behavior—and ignoring negative behavior—could be a quick and effective way to deal with problem behavior in an applied setting.

Most of this textbook is about what can be called ***group research***, which typically involves studying a large number of participants and combining their data to draw general conclusions about human behavior. The study by Hall and his colleagues, in contrast, is an example of ***single-subject research***, which typically involves studying a small number of participants and focusing closely on each individual. In this chapter, we consider this alternative approach. We begin with an overview of single-subject research, including some assumptions on which it is based, who conducts it, and why they do. We then look at some basic single-subject research designs and how the data from those designs are analyzed. Finally, we consider some of the strengths and weaknesses of single-subject research as compared with group research and see how these two approaches can complement each other.

24.1 What Is Single-Subject Research?

Single-subject research is a type of quantitative research that involves studying a person or small number of people using methods that include an analysis of their interaction with their environment. Note that the term *single-subject* does not necessarily mean that only

one participant is studied; it is more typical for there to be somewhere between two and 10 participants. (This is why single-subject research designs are sometimes called ***small-n designs***, where n is the statistical symbol for the sample size.) Single-subject research can be contrasted with group research, which typically involves studying large numbers of participants and examining their behavior primarily in terms of group means, standard deviations, and so on. The majority of this textbook is devoted to understanding group research, which is the most common approach in psychology. But single-subject research is an important alternative, and it is the primary approach in some more applied areas of psychology, such as behavior analysis.

Before continuing, it is important to distinguish single-subject research from qualitative case studies and other more qualitative approaches that involve studying in detail a small number of participants (see Chapter 16). Qualitative case studies use in-depth analysis to interpret the meaning of what occurs during the case. More broadly speaking, qualitative research focuses on understanding people's subjective experience by observing behavior and collecting relatively unstructured data (e.g., detailed interviews) and analyzing those data using narrative rather than quantitative techniques. Single-subject research, in contrast, focuses on understanding objective behavior through experimental manipulation and control, collecting highly structured data, and analyzing those data quantitatively.

24.2 Assumptions of Single-Subject Research

Again, single-subject research involves studying a small number of participants and focusing intensively on the behavior of each one. But why take this approach instead of the group approach? There are several important assumptions underlying single-subject research, and it will help to consider them now.

First and foremost is the assumption that it is important to focus intensively on the *behavior* of individual participants. One reason for this is that group research can hide individual differences and generate results that do not represent the behavior of any individual. For example, a treatment that has a positive effect for half the people exposed to it but a negative effect for the other half would, on average, appear to have no effect at all. Single-subject research, however, would likely reveal these individual differences. A second reason to focus intensively on individuals is that sometimes it is the behavior of a particular individual that is primarily of interest. A school psychologist, for example, might be interested in changing the behavior of a particular disruptive student. Although previous published research (both single-subject and group research) is likely to provide some guidance on how to do this, conducting a study on this student would be more direct and probably more effective.

A second assumption of single-subject research is that it is important to *discover causal relationships* through the manipulation of an independent variable, the careful measurement of a dependent variable, and the control of extraneous variables. For this reason, single-subject research is often considered a type of experimental research with good internal validity. Recall,

for example, that Hall and his colleagues measured their dependent variable (studying) many times—first under a no-treatment control condition, then under a treatment condition (positive teacher attention), and then again under the control condition. Because there was a clear increase in studying when the treatment was introduced, a decrease when it was removed, and an increase when it was reintroduced, there is little doubt that the treatment was the cause of the improvement.

A third assumption of single-subject research is that it is important to *study strong and consistent effects* that have biological or social importance. Applied researchers, in particular, are interested in treatments that have substantial effects on important behaviors and that can be implemented reliably in the real-world contexts in which they occur. This is sometimes referred to as ***social validity*** (Wolf, 1976). The study by Hall and his colleagues, for example, had good social validity because it showed strong and consistent effects of positive teacher attention on a behavior that is of obvious importance to teachers, parents, and students. Furthermore, the teachers found the treatment easy to implement, even in their often-chaotic elementary school classrooms.

24.3 Who Uses Single-Subject Research?

Single-subject research has been around as long as the field of psychology itself. In the late 1800s, one of psychology's founders, Wilhelm Wundt, studied sensation and consciousness by focusing intensively on each of a small number of research participants. Herman Ebbinghaus's research on memory and Ivan Pavlov's research on classical conditioning are other early examples, both of which are still described in almost every introductory psychology textbook.

In the middle of the 20th century, B. F. Skinner clarified many of the assumptions underlying single-subject research and refined many of its techniques (Skinner, 1938). He and other researchers then used it to describe how rewards, punishments, and other external factors affect behavior over time. This work was carried out primarily using nonhuman subjects—mostly rats and pigeons. This approach, which Skinner called the ***experimental analysis of behavior***—remains an important subfield of psychology and continues to rely almost exclusively on single-subject research. For excellent examples of this work, look at any issue of the *Journal of the Experimental Analysis of Behavior*. By the 1960s, many researchers were interested in using this approach to conduct applied research primarily with humans—a subfield now called ***applied behavior analysis*** (Baer et al., 1968). Applied behavior analysis plays an especially important role in contemporary research on developmental disabilities, education, organizational behavior, and health, among many other areas. Excellent examples of this work (including the study by Hall and his colleagues) can be found in the *Journal of Applied Behavior Analysis*.

Although most contemporary single-subject research is conducted from the behavioral perspective, it can in principle be used to address questions framed in terms of any theoretical perspective. For example, a studying technique based on cognitive principles of learning

and memory could be evaluated by testing it on individual high school students using the single-subject approach. The single-subject approach can also be used by clinicians who take any theoretical perspective—behavioral, cognitive, psychodynamic, or humanistic—to study processes of therapeutic change with individual clients and to document their clients' improvement (Kazdin, 1982).

24.4 General Features of Single-Subject Designs

Before looking at any specific single-subject research designs, it will be helpful to consider some features that are common to most of them. Many of these features are illustrated in Figure 24.1, which shows the results of a generic single-subject study. First, the dependent variable (represented on the y-axis of the graph) is measured repeatedly over time (represented by the x-axis) at regular intervals. Second, the study is divided into distinct phases, and the participant is tested under one condition per phase. The conditions are often designated by capital letters: A, B, C, and so on. Thus, Figure 24.1 represents a design in which the participant was tested first in one condition (A), then tested in another condition (B), and finally retested in the original condition (A). (This is called a reversal design and will be discussed in more detail shortly.)

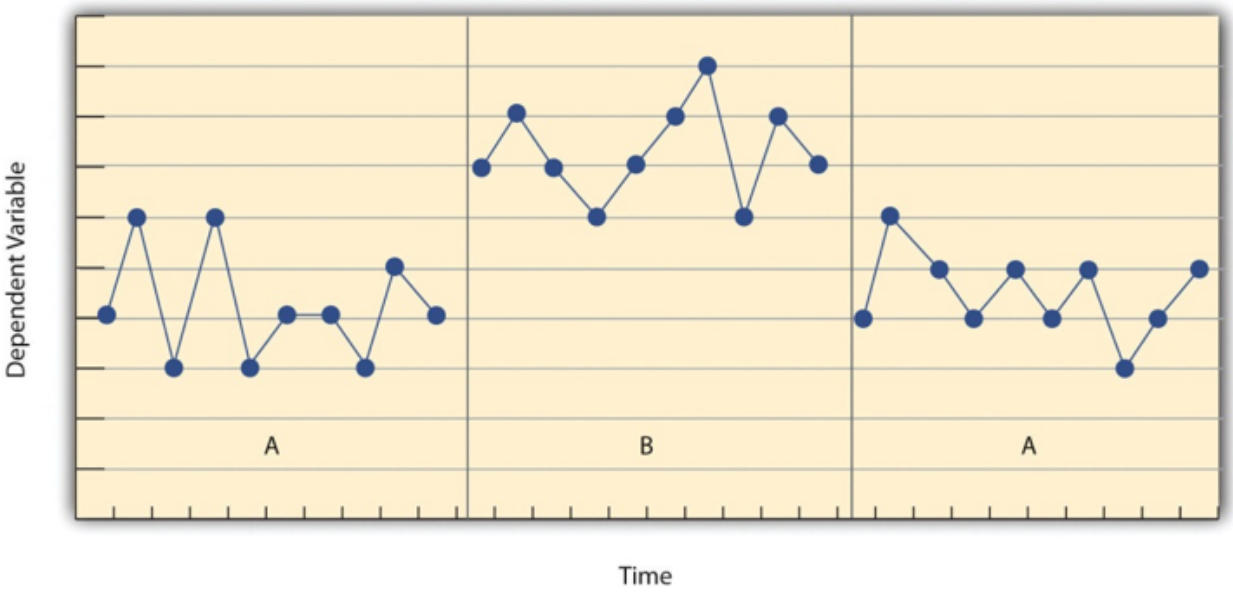


Figure 24.1: Time-series chart with Time on the x-axis and Dependent Variable on the y-axis. The plot is divided by two vertical lines into three segments labeled beneath as A, B, and A. In the first A (baseline), blue dots fluctuate at lower values. In B (intervention), the dots shift upward with higher averages and more variability. In the final A (withdrawal), values fall toward the original baseline range, with small ups and downs. The pattern suggests the intervention increased the dependent measure during phase B, with reduction after it was removed.

Figure 24.1 *Results of a Generic Single-Subject Study Illustrating Several Principles of Single-Subject Research*[1]

Another important aspect of single-subject research is that the change from one condition to the next does not usually occur after a fixed amount of time or number of observations. Instead, it depends on the participant's behavior. Specifically, the researcher waits until the participant's behavior in one condition becomes fairly consistent from observation to observation before changing conditions. This is sometimes referred to as the ***steady state strategy*** (Sidman, 1960). The idea is that when the dependent variable has reached a steady state, then any change across conditions will be relatively easy to detect. Recall that we encountered this same principle when discussing experimental research more generally. The effect of an independent variable is easier to detect when the "noise" in the data is minimized.

24.5 Reversal Designs

The most basic single-subject research design is the ***reversal design***, also called the ***ABA design***. During the first phase, A, a ***baseline*** is established for the dependent variable. This is the level of responding before any treatment is introduced, and therefore, the baseline phase is a kind of control condition. When steady state responding is reached, phase B begins as the researcher introduces the ***treatment***. There may be a period of adjustment to the treatment during which the behavior of interest becomes more variable and begins to increase or decrease. Again, the researcher waits until that dependent variable reaches a steady state so that it is clear whether and how much it has changed. Finally, the researcher removes the treatment and again waits until the dependent variable reaches a steady state. This basic reversal design can also be extended with the reintroduction of the treatment (ABAB), another return to baseline (ABABA), and so on.

The study by Hall and his colleagues employed an ABAB reversal design. Figure 24.2 approximates the data for Robbie. The percentage of time he spent studying (the dependent variable) was low during the first baseline phase, increased during the first treatment phase until it leveled off, decreased during the second baseline phase, and again increased during the second treatment phase.

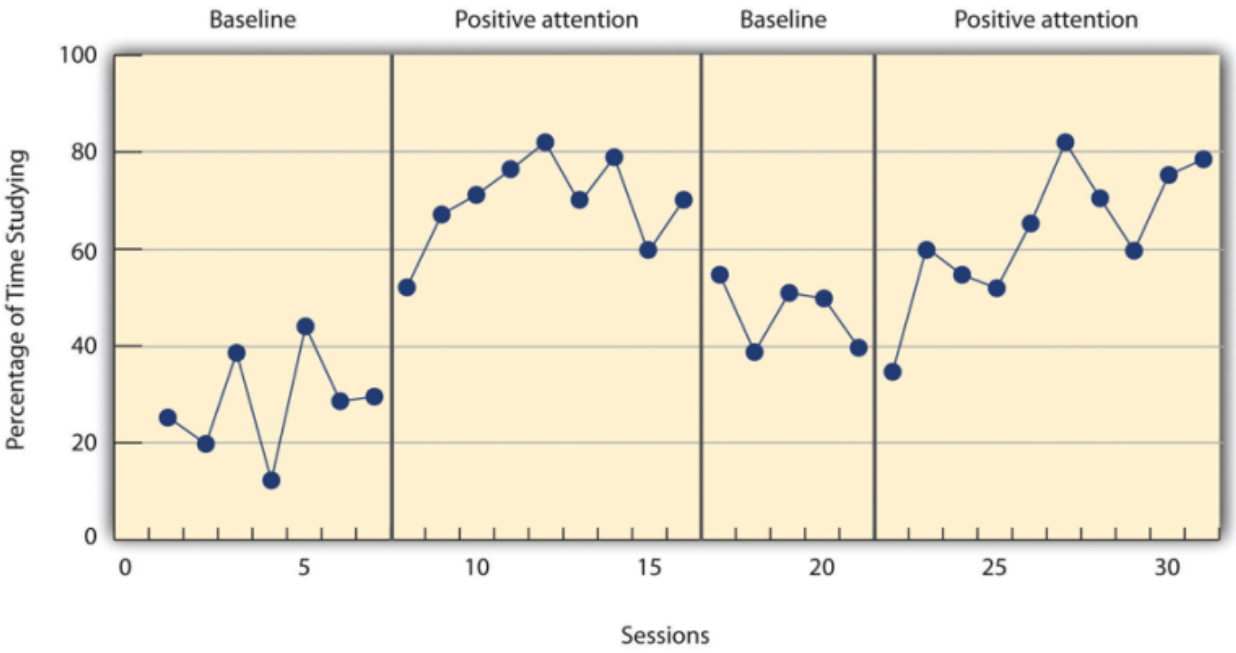


Figure 24.2: A time-series line chart with tan background plots Percentage of Time Studying (y-axis 0–100) by Sessions (x-axis). The figure is divided by vertical lines into four labeled segments across time: Baseline, Positive attention, Baseline, Positive attention. Within each of three parallel series (separated horizontally by the vertical dividers), blue points fluctuate at lower levels during Baseline (roughly 10–40%) and shift upward during Positive attention (generally 50–90%), with some variability. The repeated increases when the intervention is introduced—and decreases when baseline returns—illustrate an experimental effect of positive attention on study time.

Figure 24.2 *An Approximation of the Results for Hall and Colleagues' Participant Robbie in Their ABAB Reversal Design*[2]

Why is the reversal—the removal of the treatment—considered to be necessary in this type of design? Why use an ABA design, for example, rather than a simpler AB design? Notice that an AB design is essentially an interrupted time-series design applied to an individual participant (see Chapter 22). Recall that one problem with that design is that if the dependent variable changes after the treatment is introduced, it is not always clear that the treatment was responsible for the change. It is possible that something else changed at around the same time and that this extraneous variable is responsible for the change in the dependent variable. But if the dependent variable changes with the introduction of the treatment and then changes *back* with the removal of the treatment (assuming that the treatment does not create a permanent effect), it is much clearer that the treatment (and removal of the treatment) is the cause. In other words, the reversal greatly increases the internal validity of the study.

There are close relatives of the basic reversal design that allow for the evaluation of more than one treatment. In *a **multiple-treatment reversal design***, a baseline phase is followed by separate phases in which different treatments are introduced. For example, a researcher might establish a baseline of studying behavior for a disruptive student (A), then introduce a treatment involving positive attention from the teacher (B), and then switch to a treatment involving taking a break (escaping a non-preferred situation) (C). The participant could then be returned to a baseline phase before reintroducing each treatment—perhaps in the reverse order as a way of controlling for carryover effects. This particular multiple-treatment reversal design could also be referred to as an ABCACB design.

In an ***alternating treatments design***, two or more treatments are alternated relatively quickly on a regular schedule. For example, positive attention for studying could be used one day and receiving a break when requested the next, and so on. Or one treatment could be implemented in the morning and another in the afternoon. The alternating treatments design can be a quick and effective way of comparing treatments, but only when the treatments are fast-acting.

24.6 Multiple-Baseline Designs

There are two potential problems with the reversal design—both of which have to do with the removal of the treatment. One is that if a treatment is working, it may be unethical to remove it. For example, if a treatment seemed to reduce the incidence of self-injury in a child with an intellectual delay, it would be unethical to remove that treatment just to show that the incidence of self-injury increases. The second problem is that the dependent variable may not return to baseline when the treatment is removed. For example, when positive attention for studying is removed, a student might continue to study at an increased rate. This could mean that the positive attention had a lasting effect on the student's studying, which, of course, would be good. However, it could also mean that the positive attention was not really the cause of the increased studying in the first place. Perhaps something else happened at about the same time as the treatment—for example, the student's parents might have started rewarding him for good grades. One solution to these problems is to use a ***multiple-baseline design***, which is represented in Figure 24.3. There are three different types of multiple-baseline designs, which we will now consider.

24.6.1 Multiple-Baseline Design Across Participants

In one version of the design, a baseline is established for each of several participants, and the treatment is then introduced for each one. In essence, each participant is tested in an AB design. The key to this design is that the treatment is introduced at a different *time* for each participant. The idea is that if the dependent variable changes when the treatment is introduced for one participant, it might be a coincidence. If the dependent variable changes when the

treatment is introduced for multiple participants—especially when the treatment is introduced at different times for the different participants—then it is unlikely to be a coincidence.

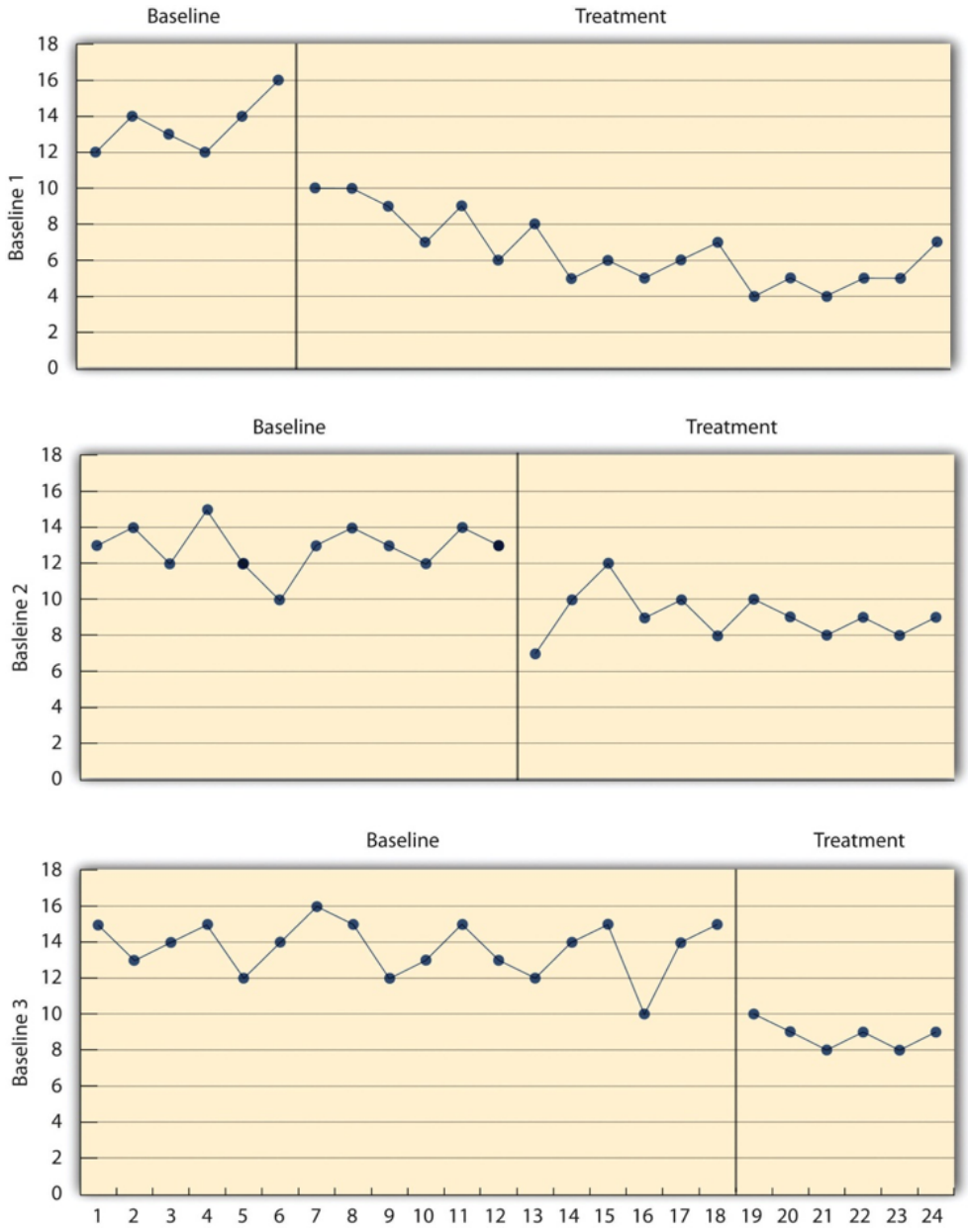


Figure 24.3: Three line graphs showing the results of a generic multiple-baseline study, in which different baselines are established and treatment is introduced to participants at different times. For Baseline 1, treatment is introduced one-quarter of the way into the study. The dependent variable ranges between 12 and 16 units during the baseline, but drops down to 10 units with treatment and mostly decreases until the end of the study, ranging between 4 and 10 units. For Baseline 2, treatment is introduced halfway through the study. The dependent variable ranges between 10 and 15 units during the baseline, then has a sharp decrease to 7 units when treatment is introduced. However, the dependent variable increases to 12 units soon after the drop and ranges between 8 and 10 units until the end of the study. For Baseline 3, treatment is introduced three-quarters of the way into the study. The dependent variable ranges between 12 and 16 units for the most part during the baseline, with one drop down to 10 units. When treatment is introduced, the dependent variable drops down to 10 units and then ranges between 8 and 9 units until the end of the study.

Figure 24.3 *Results of a Generic Multiple-Baseline Study. The multiple baselines can be for different participants, dependent variables, or settings. The treatment is introduced at a different time on each baseline*[3]

As an example, consider a study by Ross and Horner (2009). They were interested in how a school-wide bullying prevention program affected the bullying behavior of particular problem students. At each of the three different schools, the researchers studied two students who had regularly engaged in bullying. During the baseline phase, they observed the students for 10-minute periods each day during lunch recess and counted the number of aggressive behaviors they exhibited toward their peers. After 2 weeks, they implemented the program at one school. After 2 more weeks, they implemented it at the second school. And after 2 more weeks, they implemented it at the third school. They found that the number of aggressive behaviors exhibited by each student dropped shortly after the program was implemented at the student's school. Notice that if the researchers had only studied one school or if they had introduced the treatment at the same time at all three schools, then it would be unclear whether the reduction in aggressive behaviors was due to the bullying program or something else that happened at about the same time it was introduced (e.g., a holiday, a television program, a change in the weather). With their multiple-baseline design, this kind of coincidence would have to happen three separate times—a very unlikely occurrence—to explain their results.

24.6.2 Multiple-Baseline Design Across Behaviors

In another version of the multiple-baseline design, multiple baselines are established for the same participant but for different dependent variables, and the treatment is introduced at a different time for each dependent variable. Imagine, for example, a study on the effect of setting clear goals on the productivity of an office worker who has two primary tasks: making sales calls and writing reports. Baselines for both tasks could be established. For example, the researcher could measure the number of sales calls made and reports written by the worker each week for several weeks. Then the goal-setting treatment could be introduced for *one* of these tasks, and at a later time the same treatment could be introduced for the other task. The logic is the same as before. If productivity increases on one task after the treatment is introduced, it is unclear whether the treatment caused the increase. However, if productivity increases on both tasks after the treatment is introduced—especially when the treatment is introduced at two different times—then it seems much clearer that the treatment was responsible.

24.6.3 Multiple-Baseline Design Across Settings

In yet a third version of the multiple-baseline design, multiple baselines are established for the same participant but in different settings. For example, a baseline might be established for the amount of time a child spends reading during his free time at school and during his free time at home. Then, a treatment such as positive attention might be introduced first at school and later at home. Again, if the dependent variable changes after the treatment is introduced in

each setting, then this gives the researcher confidence that the treatment is, in fact, responsible for the change.

To see an overall brief review of many of the single-subject designs discussed so far, see the following video, "Single Subject Designs":

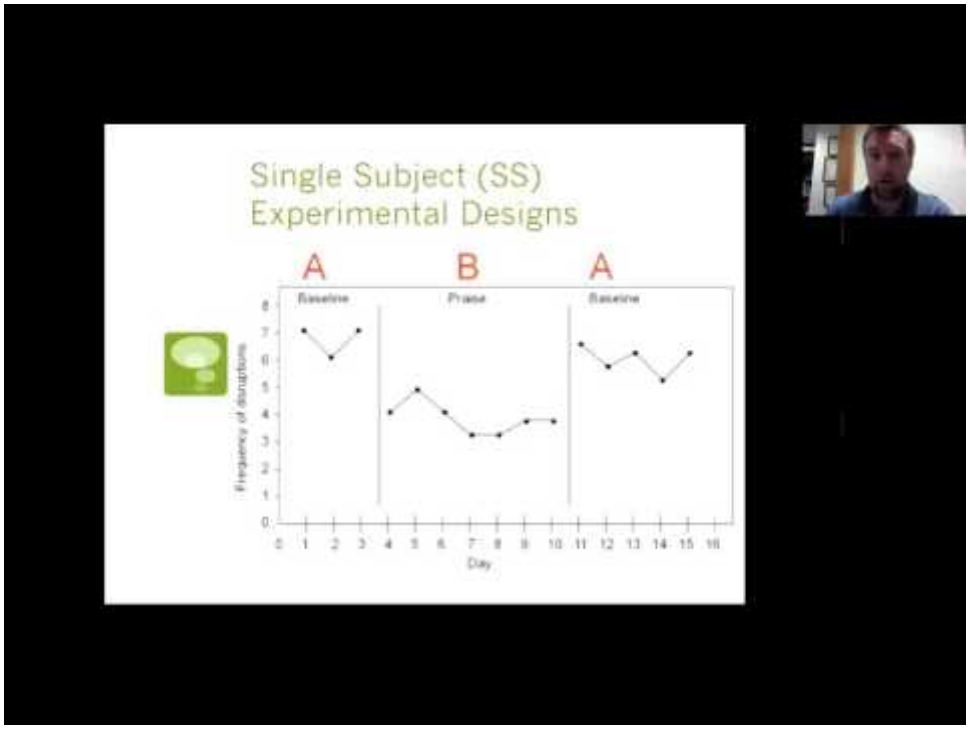


Figure 24.4: Single Subject Designs

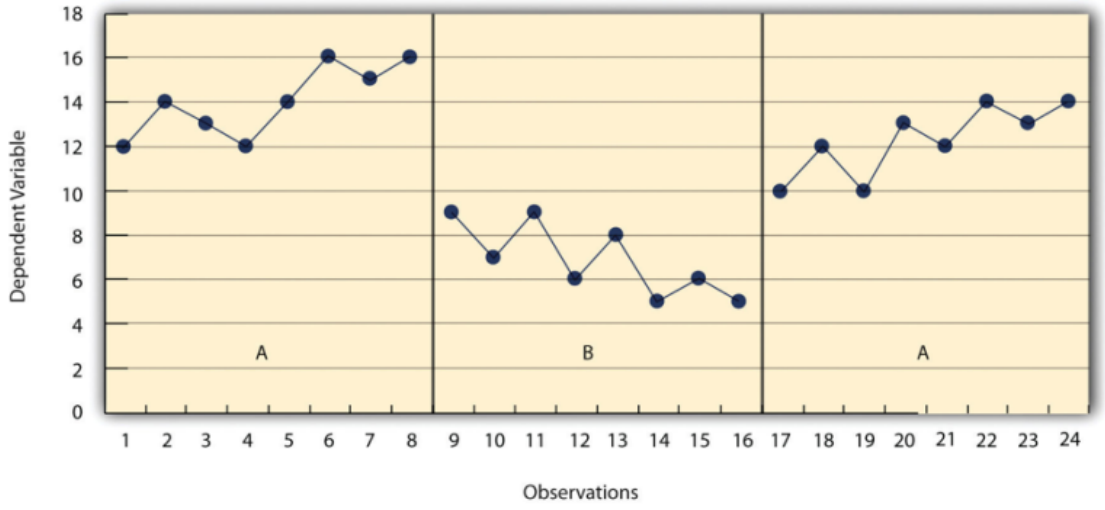
24.7 Data Analysis in Single-Subject Research

In addition to its focus on individual participants, single-subject research differs from group research in the way the data are typically analyzed. As we have seen throughout the book, group research involves combining data across participants. Group data are described using statistics such as means, standard deviations, correlation coefficients, and so on to detect general patterns. Finally, inferential statistics are used to help decide whether the result for the sample is likely to generalize to the population. Single-subject research, by contrast, relies heavily on a very different approach called ***visual inspection (charting)***. This means plotting individual participants' data as shown throughout this chapter, looking carefully at those data, and making judgments about whether and to what extent the independent variable had an effect on the dependent variable. Inferential statistics are typically not used.

In visually inspecting their data, single-case experimental researchers take several factors into account. One of them is changes in the ***level***, or values, of the dependent variable from condition to condition. In other words, like researchers using groups designs, single-case experimental researchers examine whether the values of the dependent variable change as the level of the independent variable is manipulated. If the dependent variable is much higher or much lower in one condition than another, this suggests that the treatment had an effect. A second factor is ***variability***, which refers to how consistent the dependent variable is from observation to observation. As noted above, variability adds noise to the data, making it difficult to detect whether a manipulation of the independent variable has affected the dependent variable. For this reason, researchers using single-case experimental designs often establish a ***stability criterion*** that specifies the maximum amount of variability permitted (a topic that is beyond the scope of this course but will be covered in more advanced research methods classes).

A third factor is ***trend***, which refers to gradual increases or decreases in the dependent variable across observations. If the dependent variable begins increasing or decreasing with a change in conditions, then again this suggests that the treatment had an effect. It can be especially telling when a trend changes directions—for example, when an unwanted behavior is increasing during baseline but then begins to decrease with the introduction of the treatment. A fourth factor is ***latency***, which in this context is the time it takes for the dependent variable to begin changing after a change in conditions. In general, if a change in the dependent variable begins shortly after a change in conditions, this suggests that the treatment was responsible.

In the top panel of Figure 24.4, there are fairly obvious changes in the level and trend of the dependent variable from condition to condition. Furthermore, the latencies of these changes are short; the change happens immediately. This pattern of results strongly suggests that the treatment was responsible for the changes in the dependent variable. In the bottom panel of Figure 24.4, however, the changes in level are fairly small. Although there appears to be an increasing trend in the treatment condition, it looks as though it might be a continuation of a trend that had already begun during baseline. This pattern of results strongly suggests that the treatment was not responsible for any changes in the dependent variable—at least not to the extent that single-subject researchers typically hope to see.



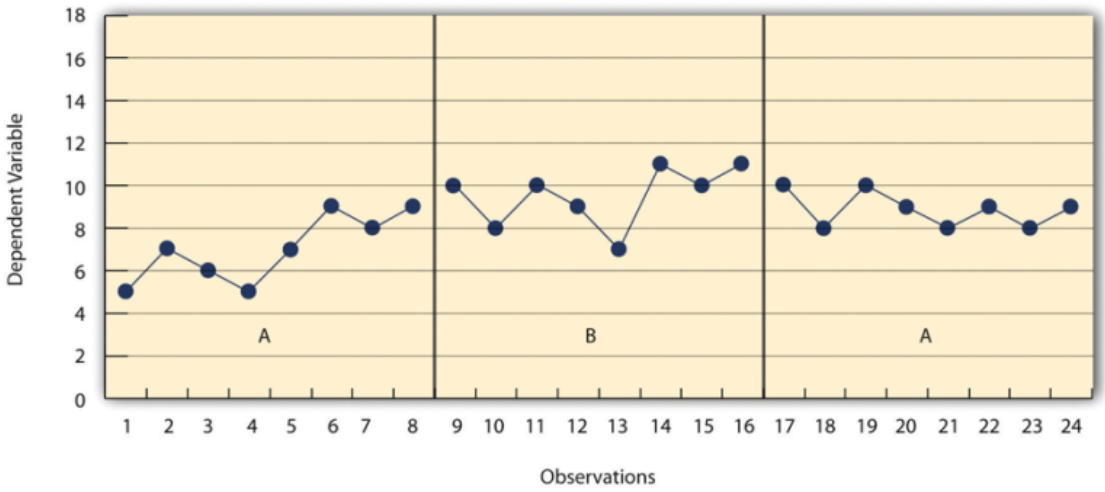


Figure 24.5: Two graphs showing the results of a generic single-subject study with an ABA design. In the first graph, under condition A, level is high and the trend is increasing. Under condition B, level is much lower than under condition A and the trend is decreasing. Under condition A again, level is about as high as the first time and the trend is increasing. For each change, latency is short, suggesting that the treatment is the reason for the change. In the second graph, under condition A, level is relatively low and the trend is increasing. Under condition B, level is a little higher than during condition A and the trend is increasing slightly. Under condition A again, level is a little lower than during condition B and the trend is decreasing slightly. It is difficult to determine the latency of these changes, since each change is rather minute, which suggests that the treatment is ineffective.

Figure 24.4 *Results of a Generic Single-Subject Study Illustrating Level, Trend, and Latency. Visual inspection of the data suggests an effective treatment in the top panel but an ineffective treatment in the bottom panel*[4]

The results of single-subject research can also be analyzed using statistical procedures—and this is becoming more common. There are many different approaches, and single-subject researchers continue to debate which are the most useful. One approach parallels what is typically done in group research. The mean and standard deviation of each participant's responses under each condition are computed and compared, and inferential statistical tests such as the *t*-test or analysis of variance (ANOVA) are applied (Fisch, 2001). (Note that averaging *across* participants is less common.) Another approach is to compute the ***percentage of non-overlapping data (PND)*** for each participant (Scruggs & Mastropieri, 2001). This is the percentage of responses in the treatment condition that are more extreme than the most extreme response in a relevant control condition. In the study of Hall and his colleagues, for example, all measures of Robbie's study time in the first treatment condition were greater than the highest measure in the first baseline, for a PND of 100%. The greater the percentage of non-overlapping data, the stronger the treatment effect. Still, formal statistical approaches to data analysis in single-subject research are generally considered a supplement to visual inspection, not a replacement for it.

24.8 Single-Subject versus Group Design "Debate"

Single-subject research is similar to group research—especially experimental group research—in many ways. They are both quantitative approaches that try to establish causal relationships by manipulating an independent variable, measuring a dependent variable, and controlling extraneous variables. However, there are important differences between these approaches too, and these differences sometimes lead to disagreements. It is worth addressing the most common points of disagreement between single-subject researchers and group researchers and how these disagreements can be resolved. As we will see, single-subject research and group research are probably best conceptualized as complementary approaches.

24.8.1 Debate Regarding How to Analyze Data

One set of disagreements revolves around the issue of data analysis. Some advocates of group research worry that visual inspection is inadequate for deciding whether and to what extent a treatment has affected a dependent variable. One specific concern is that visual inspection is not sensitive enough to detect weak effects. A second is that visual inspection can be unreliable, with different researchers reaching different conclusions about the same set of data (Danov & Symons, 2008). A third is that the results of visual inspection—an overall judgment of whether or not a treatment was effective—cannot be clearly and efficiently summarized or compared across studies (unlike the measures of relationship strength typically used in group research).

In general, single-subject researchers share these concerns. However, they also argue that their use of the steady state strategy, combined with their focus on strong and consistent effects, minimizes most of them. If the effect of a treatment is difficult to detect by visual inspection because the effect is weak or the data are noisy, then single-subject researchers look for ways to increase the strength of the effect or reduce the noise in the data by controlling extraneous variables (e.g., by administering the treatment more consistently). If the effect is still difficult to detect, then they are likely to consider it neither strong enough nor consistent enough to be of further interest. Many single-subject researchers also point out that statistical analysis is becoming increasingly common and that many of them are using this as a supplement to visual inspection—especially for the purpose of comparing results across studies (Scruggs & Mastropieri, 2001).

Turning the tables, some advocates of single-subject research worry about the way that group researchers analyze their data. Specifically, they point out that focusing on group means can be highly misleading. Again, imagine that a treatment has a strong positive effect on half the people exposed to it and an equally strong negative effect on the other half. In a traditional between-subjects experiment, the positive effect on half the participants in the treatment condition would be statistically cancelled out by the negative effect on the other half. The mean for the treatment group would then be the same as the mean for the control group, making it seem as though the treatment had no effect when in fact it had a strong effect on every single participant!

But again, group researchers share this concern. Although they do focus on group statistics, they also emphasize the importance of examining distributions of individual scores. For example, if some participants were positively affected by a treatment and others negatively affected by it, this would produce a bimodal distribution of scores and could be detected by looking at a histogram of the data. The use of within-subjects designs is another strategy that allows group researchers to observe effects at the individual level and even to specify what percentage of individuals exhibit strong, medium, weak, and even negative effects. Finally, factorial designs can be used to examine whether the effects of an independent variable on a dependent variable differ in different groups of participants (e.g., introverts vs. extraverts).

24.8.2 Discussion Regarding External Validity

The second issue about which single-subject and group researchers sometimes disagree has to do with ***external validity***—the ability to generalize the results of a study beyond the people and specific situation actually studied. In particular, advocates of group research point out the difficulty in knowing whether results for just a few participants are likely to generalize to others in the population. Imagine, for example, that in a single-subject study, a treatment has been shown to reduce self-injury for each of two children with intellectual disabilities. Even if the effect is strong for these two children, how can one know whether this treatment is likely to work for other children with intellectual delays?

Again, single-subject researchers share this concern. In response, they note that the strong and consistent effects they are typically interested in—even when observed in small samples—are likely to generalize to others in the population. Single-subject researchers also note that they place a strong emphasis on replicating their research results. When they observe an effect with a small sample of participants, they typically try to replicate it with another small sample—perhaps with a slightly different type of participant or under slightly different conditions. Each time they observe similar results, they rightfully become more confident in the generality of those results. Single-subject researchers can also point to the fact that the principles of classical and operant conditioning—most of which were discovered using the single-subject approach—have been successfully generalized across an incredibly wide range of species and situations.

And, once again, turning the tables, single-subject researchers have concerns of their own about the external validity of group research. One extremely important point they make is that studying large *groups* of participants does not entirely solve the problem of generalizing to other *individuals*. Imagine, for example, a treatment that has been shown to have a small positive effect on average in a large group study. It is likely that although many participants exhibited a small positive effect, others exhibited a large positive effect, and still others exhibited a small negative effect. When it comes to applying this treatment to another large *group*, we can be fairly sure that it will have a small effect on average. But when it comes to applying this treatment to another *individual*, we cannot be sure whether it will have a small, a large, or even a negative effect. Another point that single-subject researchers make is that group researchers also face a similar problem when they study a single situation and then generalize their results to other situations. For example, researchers who conduct a study on the effect of cell phone use on drivers on a closed oval track probably want to apply their results to drivers in many other real-world driving situations. However, notice that this requires generalizing from a single situation to a population of situations. Thus, the ability to generalize is based on much more than just the sheer number of participants one has studied. It requires a careful consideration of the similarity of the participants *and* situations studied to the population of participants and situations to which one wants to generalize (Shadish et al., 2002).

24.8.3 Single-Subject and Group Research as Complementary Methods

As with quantitative and qualitative research, it is probably best to conceptualize single-subject research and group research as ***complementary*** methods: They have different strengths and weaknesses and that are appropriate for answering different kinds of research questions (Kazdin, 1982). Single-subject research is particularly good for testing the effectiveness of treatments on individuals when the focus is on strong, consistent, and biologically or socially important effects. It is also especially useful when the behavior of particular individuals is of interest. Clinicians who work with only one individual at a time may find that it is their only option for doing systematic quantitative research.

Group research, on the other hand, is ideal for testing the effectiveness of treatments at the group level. Among the advantages of this approach is that it allows researchers to detect weak effects, which can be of interest for many reasons. For example, finding a weak treatment effect might lead to refinements of the treatment that eventually produce a larger and more meaningful effect. Group research is also good for studying interactions between treatments and participant characteristics. For example, if a treatment is effective for those who are high in motivation to change and ineffective for those who are low in motivation to change, then a group design can detect this much more efficiently than a single-subject design. Group research is also necessary to answer questions that cannot be addressed using the single-subject approach, including questions about independent variables that cannot be manipulated (e.g., number of siblings, extraversion, culture).

Finally, it is important to understand that the single-subject and group approaches represent different research traditions. This factor is probably the most important one affecting which approach a researcher uses. Researchers in the experimental analysis of behavior and applied behavior analysis learn to conceptualize their research questions in ways that are amenable to the single-subject approach. Researchers in most other areas of psychology learn to conceptualize their research questions in ways that are amenable to the group approach. At the same time, there are many topics in psychology in which research from the two traditions has informed each other and been successfully integrated. One example is research suggesting that both animals and humans have an innate "number sense"—an awareness of how many objects or events of a particular type they have experienced without actually having to count them (Dehaene, 2011). Single-subject research with rats and birds and group research with human infants have shown strikingly similar abilities in those populations to discriminate small numbers of objects and events. This number sense—which probably evolved long before humans did—may even be the foundation of humans' advanced mathematical abilities.

24.8.4 The Principle of Converging Evidence

Now that you have been introduced to many of the most commonly used research methods in psychology, it should be readily apparent that no design is perfect. Every research design has strengths and weaknesses. True experiments typically have high internal validity but may have problems with external validity, while non-experimental research (e.g., correlational research) often has good external validity but poor internal validity. Each study brings us closer to the truth, but no single study can ever be considered definitive. This is one reason why, in science, we say there is no such thing as scientific *proof*; there is only scientific *evidence*.

While the media will often try to reach strong conclusions on the basis of the findings of one study, scientists focus on evaluating a body of research. Scientists evaluate theories not by waiting for the perfect experiment but by looking at the overall trends in a number of partially flawed studies. The idea of ***converging evidence*** tells us to examine the pattern of flaws running through the research literature because the nature of this pattern can either support or undermine the conclusions we wish to draw. Suppose the findings from a number of different

studies were largely consistent in supporting a particular conclusion. If all of the studies were flawed in a similar way, for example, if all of the studies were correlational and contained the third variable problem and the directionality problem, this would undermine confidence in the conclusions drawn because the consistency of the outcome may simply have resulted from a particular flaw that all of the studies shared. On the other hand, if all of the studies were flawed in different ways and the weakness of some of the studies were the strength of others (the low external validity of a true experiment was balanced by the high external validity of a correlational study), then we could be more confident in our conclusions.

While there are fundamental tradeoffs in different research methods, the diverse set of approaches used by psychologists has complementary strengths that allow us to search for converging evidence. We can reach meaningful conclusions and come closer to understanding truth by examining a large number of different studies, each with different strengths and weaknesses. If the results of a large number of studies, all conducted using different designs, converge on the same conclusion, then our confidence in that conclusion can be increased dramatically. In science, we strive for progress, not perfection.

24.9 Additional Resources

- Video: "Single Subject Research: Visual Analysis of Trend" contains a theoretical discussion of data analysis, namely the visual analysis of trend.
- Video: "Single Subject Research: Visual Analysis of Level and Overlap" contains a theoretical discussion of data analysis, namely the visual analysis of level and percentage of non-overlapping data.
- Video: "Single Subject Research: Visual Analysis Practice Activity" contains a practical discussion of data analysis, looking at level, trend, stability, and one way to assess percentage of non-overlapping data.

Practice 1: Using a Single-Subject Design to Reduce Social Media Use

Research has examined how social media usage affects people, with global average daily social media use (SMU) being over 2 hours per day (Statista, 2025). Correlational research found that high SMU is associated with sleep disturbances, lower academic performance, and depression, among other issues, and experimental research has shown that SMU provides benefits (see Olgun et al., 2026 for a discussion). Olgun et al. (2026) conducted a similar study as a prior study (Stinson & Dallery, 2023) that employed an intervention where participants set limits on their SMU, received vouchers for reducing it (contingency management), and selected alternative activities to pursue. Four participants completed the study with this intervention, tracking how much time they engaged in social media and how much time they completed their alternative activities. The following graph tracks the 4 participants' daily social media usage prior to the study, during the baseline phase,

and during the intervention phase that included contingency management:

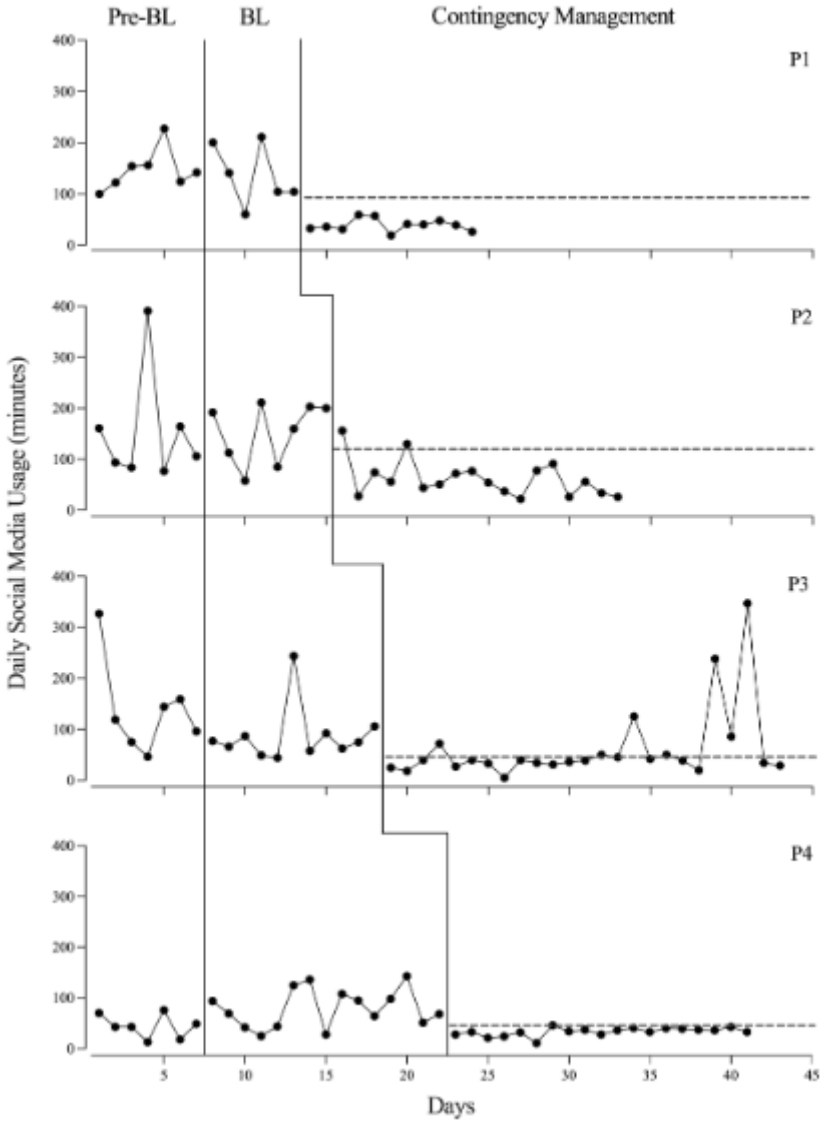


Figure 24.5 *Daily SMU for participants. The dotted horizontal lines indicate each participant's daily SMU goal.*[5]

1. What type of single-subject design is this?
2. Visually describe the data for P2 that you see.
3. Using this PND calculator [New Tab], what is the percentage of non-overlapping data from the Baseline to Contingency Management phase?

4. How do the researchers use this methodology to reduce social media usage? Could other designs have addressed this issue?

24.9.1 Practice 2: Designing Single-*N* Studies

Design a simple single-subject study (using either a reversal or multiple-baseline design) to answer the following questions. Be sure to specify the treatment, operationally define the dependent variable, decide when and where the observations will be made, and so on.

- Does positive attention from a parent increase a child's tooth-brushing behavior?
- Does self-testing while studying improve a student's performance on weekly spelling tests?
- Does regular exercise help relieve depression?

24.9.2 Practice 3: Creating a Graph for Single-Subject Studies

Create a graph that displays the hypothetical results for the study you designed in Practice 2. Write a paragraph in which you describe what the results show. Be sure to comment on level, trend, and latency.

24.9.3 Practice 4: Examining Group and Single-Subject Studies

Redesign as a group study the study by Hall et al. (1968) described at the beginning of this chapter, and list the strengths and weaknesses of your new study compared with the original study.

24.10 Media Attributions

[1-4] D'Costa, S., Ukeye, M., O'Neil, M., & Anguiano, R. (no date). Critical Research Methods in Psychology. "Figure 18.1" is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[5] Olgun, D., Deshais, M. A., Kahng, S., & LaRue, R. H. (2026). Reducing social media use via contingency management: A replication and extension. *Journal of Applied Behavior Analysis, 59*(2), e70061. "FIGURE 1" is licensed under a Creative Commons Attribution License. Modified by current authors.

24.11 Text Attributions

D'Costa, S., Ukeye, M., O'Neil, M., Anguiano, R. (n.d.). Critical Research Methods in Psychology. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Jhangiani, R.S., Chiang, I-C.A., Cuttler, C. & Leighton, D.C. (2019). Research methods in psychology. 4th edition. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Olgun, D., Deshais, M. A., Kahng, S., & LaRue, R. H. (2026). Reducing social media use via contingency management: A replication and extension. *Journal of Applied Behavior Analysis*, *59*(2), e70061. https://doi.org/10.1002/jaba.70061. Licensed under a Creative Commons Attribution License Modified by current authors.

Serdikoff, S. L. (n.d.). Research Methods for the Behavioral Sciences. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

24.12 References

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91–97. https://doi.org/10.1901/jaba.1968.1-91

Danov, S. E., & Symons, F. E. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behavior Modification, 32*, 828–839. https://doi.org/10.1177/0145445508318606

Dehaene, S. (2011). *The number sense: How the mind creates mathematics (2nd ed.).* Oxford.

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models. *Behavioral Processes, 54*, 137–154. https://doi.org/10.1016/S0376-6357(01)00155-3

Hall, R. V., Lund, D., & Jackson, D. (1968). Effects of teacher attention on study behavior. *Journal of Applied Behavior Analysis, 1*, 1–12. https://doi.org/10.1901/jaba.1968.1-1

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* Oxford University Press.

Olgun, D., Deshais, M. A., Kahng, S., & LaRue, R. H. (2026). Reducing social media use via contingency management: A replication and extension. *Journal of Applied Behavior Analysis*, *59*(2), e70061. https://doi.org/10.1002/jaba.70061

Ross, S. W., & Horner, R. H. (2009). Bully prevention in positive behavior support. *Journal of Applied Behavior Analysis, 42*, 747–759. https://doi.org/10.1901/jaba.2009.42-747

Scruggs, T. E., & Mastropieri, M. A. (2001). How to summarize single-participant research: Ideas and applications. *Exceptionality, 9*, 227–244.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* Authors Cooperative.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis.* Appleton-Century-Crofts.

Stinson, L., & Dallery, J. (2023). Reducing problematic social media use via a package intervention. *Journal of Applied Behavior Analysis*, *56*(2), 323-335.

Wolf, M. (1976). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11*, 203–214. https://doi.org/10.1002/jaba.975

25 Chapter 25: Replication Crisis, Replicability, and Science Practices

In science, replication is the process of repeating research to determine the extent to which findings generalize across time and across situations. Recently, the science of psychology has come under criticism because a number of research findings do not replicate. In this module we discuss reasons for non-replication, indicate the impact this phenomenon has on the field, and suggest solutions to the problem.

25.1 The Disturbing Problem

If you were driving down the road and you saw a pirate standing at an intersection, you might not believe your eyes. However, if you continued driving and saw a second, and then a third, you might become more confident in your observations. The more pirates you saw, the less likely the first sighting would be a false positive (you were driving fast and the person was just wearing an unusual hat and billowy shirt) and the more likely it would be the result of a logical reason (there is a pirate themed conference in town). This somewhat absurd example is a real-life illustration of ***replication***: the repeated findings of the same results.

Figure 25.1: Four pirates

Figure 25.1 *A pirate appears four times, as in replication*[1]

The replication of findings is one of the defining hallmarks of science. Scientists must be able to replicate the results of studies or their findings do not become part of scientific knowledge. Replication protects against ***false positives*** (seeing a result that is not really there) and also increases confidence that the result actually exists. If you collect satisfaction data among homeless people living in Kolkata, India, for example, it might seem strange that they would report fairly high satisfaction with their food (which is exactly what we found in Biswas-Diener & Diener, 2001). If you find the exact same result, but at a *different* time, and with a *different* sample of homeless people living in Kolkata, however, you can feel more confident that this result is true (as we did in Biswas-Diener & Diener, 2006).

In modern times, the science of psychology is facing a crisis. It turns out that many studies in psychology—including many highly cited studies—do not replicate. In an era where news is instantaneous, the failure to replicate research raises important questions about the scientific process in general and psychology specifically. People have the right to know if they can trust research evidence. For our part, psychologists also have a vested interest in ensuring that our methods and findings are as trustworthy as possible.

Psychology is not alone in coming up short on replication. There have been notable failures to replicate findings in other scientific fields as well. For instance, in 1989 scientists reported that they had produced "cold fusion," achieving nuclear fusion at room temperatures. This could have been an enormous breakthrough in the advancement of clean energy. However, other scientists were unable to replicate the findings. Thus, the potentially important results

did not become part of the scientific canon, and a new energy source did not materialize. In medical science as well, a number of findings have been found not to replicate—which is of vital concern to all of society. The non-reproducibility of medical findings suggests that some treatments for illness could be ineffective. One example of non-replication has emerged in the study of genetics and diseases: When replications were attempted to determine whether certain gene-disease findings held up, only about 4% of the findings consistently did so.

The non-reproducibility of findings is disturbing because it suggests the possibility that the original research was done sloppily. Even worse is the suspicion that the research may have been falsified. In science, faking results is *the biggest* of sins, the unforgivable sin, and for this reason the field of psychology has been thrown into an uproar. However, as we will discuss, there are a number of explanations for non-replication, and not all of them are bad.

While the vast majority of scientists are honest, unfortunately, some are not. Not only that, scientists are not immune to belief bias–it is easy for a researcher to end up deceiving themselves into believing the wrong thing, and this can lead them to conduct subtly flawed research and then hide those flaws when they write it up. So, you need to consider not only the (probably unlikely) possibility of outright fraud, but also the (probably quite common) possibility that the research is unintentionally "slanted."

25.2 Enormity of the Current Crisis

Recently, there has been growing concern as psychological research fails to replicate. To give you an idea of the extent of non-replicability of psychology findings, below are data reported in 2015 by the Open Science Collaboration project, led by University of Virginia psychologist Brian Nosek (Open Science Collaboration, 2015). Because these findings were reported in the prestigious journal, *Science,* they received widespread attention from the media. Here are the percentages of research that replicated—selected from several highly prestigious journals:

Table 25.1 *The Reproducibility of Psychological Science* [2]

Journal	% Findings Replicated
Journal of Personality and Social Psychology: Social	23
Journal of Experimental Psychology: Learning, Memory, and Cognition	48
Psychological Science, social articles	29
Psychological Science, cognitive articles	53
Overall	36

Figure 25.2: Percentage of findings published in prestigious journals which have replicated: (1) Journal of Personality and Social Psychology - Social, 23%, (2) Journal of Experimental Psychology - Learning, Memory, and Cognition, 48%, (3) Psychological Science - social articles, 29%, (4) Psychological Science - cognitive articles, 53%, (5) Overall, 36%

Clearly, there is a very large problem when only about 1/3 of the psychological studies in premier journals replicate! It appears that this problem is particularly pronounced for social psychology, but even the 53% replication level of cognitive psychology is cause for concern.

The situation in psychology has grown so worrisome that the Nobel Prize-winning psychologist Daniel Kahneman called on social psychologists to clean up their act (Kahneman, 2012). The Nobel laureate spoke bluntly of doubts about the integrity of psychology research, calling the current situation in the field a "mess." His missive was pointed primarily at researchers who study social "priming," but in light of the non-replication results that have since come out, it might be more aptly directed at the behavioral sciences in general.

25.3 What are the Different Types of Replication?

There are different types of replication. First, there is a type called "***exact replication***" (also called "direct replication"). In this form, a scientist attempts to exactly recreate the scientific methods used in conditions of an earlier study to determine whether the results come out the same. If, for instance, you wanted to exactly replicate Asch's (1956) classic findings on conformity, you would follow the original methodology: You would use only male participants,

you would use groups of 8, and you would present the same stimuli (lines of differing lengths) in the same order. The second type of replication is called "***conceptual replication***." This occurs when—instead of an exact replication, which reproduces the methods of the earlier study as closely as possible—a scientist tries to confirm the previous findings using a different set of specific methods that test the same idea. The same hypothesis is tested, but using a different set of methods and measures. A conceptual replication of Asch's research might involve both male and female ***confederates*** purposefully misidentifying types of fruit to investigate conformity—rather than only males misidentifying line lengths.

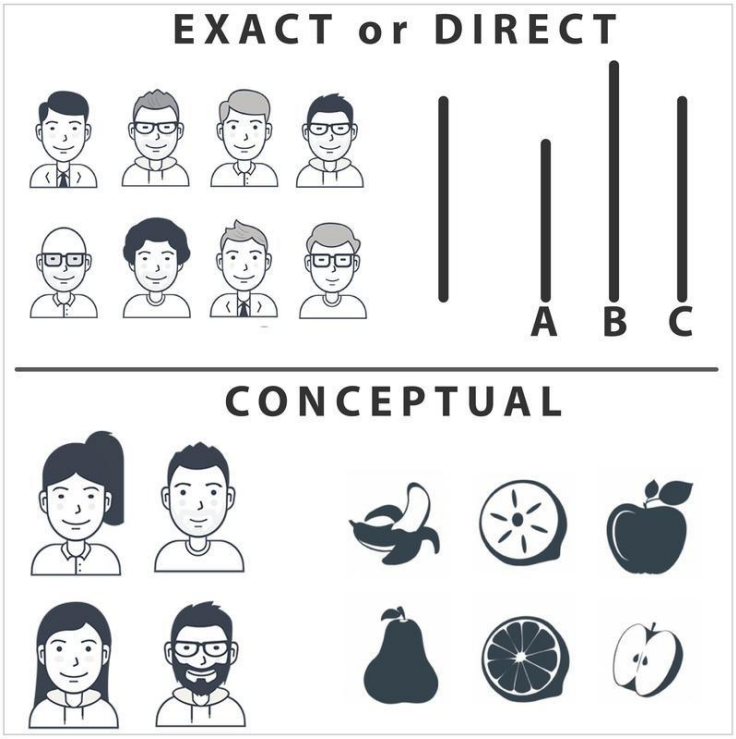


Figure 25.3: Top image - group of 8 men presented with lines of different lengths. Bottom image - group of two men and two women presented with different types of fruit.

Figure 25.2 *Example of direct replication and conceptual replication of Asch's conformity experiment*[3]

Both exact and conceptual replications are important because they each tell us something new. Exact replications tell us whether the original findings are true, at least under the exact conditions tested. Conceptual replications help confirm whether the theoretical idea behind the findings is true and under what conditions these findings will occur. In other words, conceptual replication offers insights into how generalizable the findings are.

25.4 Examples of Non-Replications in Psychology

A large number of scientists have attempted to replicate studies on what might be called "metaphorical priming," and more often than not these replications have failed. ***Priming*** is the process by which a recent reference (often a subtle, subconscious cue) can increase the accessibility of a trait. For example, if your instructor says, "Please put aside your books, take out a clean sheet of paper, and write your name at the top," you might find your pulse quickening. Over time, you have learned that this cue means you are about to be given a pop quiz. This phrase primes all the features associated with pop quizzes: They are anxiety-provoking, they are tricky, your performance matters.

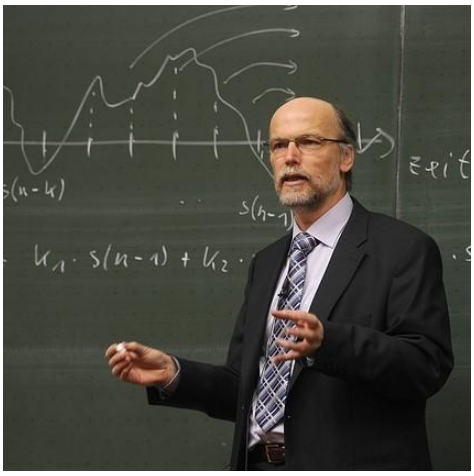


Figure 25.4: A stereotypical image of a professor - a white, middle-aged man with glasses and a beard, dressed in a coat and tie stands with chalk in hand in front of a blackboard which displays a mathematical formula.

Figure 25.3 *Priming with a stereotypical image of a professor*[4]

One example of a priming study that (at least in some cases) does not replicate is the priming of the idea of intelligence. In theory, it might be possible to prime people to actually become more intelligent (or perform better on tests, at least). For instance, in one study, priming students with the idea of a stereotypical professor versus soccer hooligans led participants in the "professor" condition to earn higher scores on a trivia game (Dijksterhuis & van Knippenberg, 1998). Unfortunately, in several follow-up instances this finding has not replicated (Shanks et al, 2013). This is unfortunate for all of us because it would be a very easy way to raise our test scores and general intelligence. If only it were true.

Another example of a finding that has not always replicated consistently is the finding that taking notes by hand is better than using a computer (Mueller & Oppenheimer, 2014). In theory, taking notes by hand may force note takers to reword the material, which would lead

to deeper processing and better learning. On the other hand, taking notes by computer may lead note takers to take notes verbatim, leading to shallower processing and worse learning. Although this finding showed us which methods of learning are most effective, direct replication and conceptual replication studies have shown that both strategies have similar effects on learning (Morehead et al., 2019; Urry et al., 2021). This is fortunate as it means that taking notes, no matter what the method, may help anyone!

As one can see from the examples, some of the studies that fail to replicate report extremely interesting findings—even counterintuitive findings that appear to offer new insights into the human mind. Critics claim that psychologists have become too enamored with such newsworthy, surprising "discoveries" that receive a lot of media attention. This raises the question of timing: Might the current crisis of non-replication be related to the modern, media-hungry context in which psychological research (indeed, all research) is conducted? Put another way: Is the non-replication crisis new or just more observable?

Nobody has tried to systematically replicate studies from the past, so we do not know if published studies are becoming less replicable over time. In 1990, however, Amir and Sharon were able to successfully replicate most of the main effects of six studies from another culture, though they did fail to replicate many of the interactions. This particular shortcoming in their overall replication may suggest that published studies are becoming less replicable over time, but we cannot be certain. What we can be sure of is that there is a significant problem with replication in psychology, and it's a trend the field needs to correct. Without replicable findings, nobody will be able to have confidence in scientific psychology.

25.5 In Defense of Replication Attempts

Failures in replication are not all bad and, in fact, some non-replication should be expected in science. Original studies are conducted when an answer to a question is uncertain. That is to say, scientists are venturing into new territory. In such cases, we should expect some answers to be uncovered that will not pan out in the long run. Furthermore, we hope that scientists take on challenging new topics that come with some amount of risk. After all, if scientists were only to publish safe results that were easy to replicate, we might have very boring studies that do not advance our knowledge very quickly. However, with such risks, some non-replication of results is to be expected.

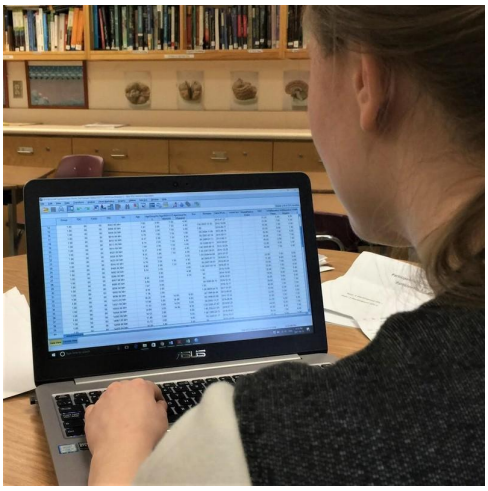


Figure 25.5: A woman analyzing data on a computer.

Figure 25.4. *Researchers use specialized statistical software to store, analyze, and share data*[5]

A recent example of risk-taking can be seen in the research of social psychologist Daryl Bem. In 2011, Bem published an article claiming he had found in a number of studies that future events could influence the past. His proposition turns the nature of time, which is assumed by virtually everyone except science fiction writers to run in one direction, on its head. Needless to say, attacks on Bem's article came fast and furious, including attacks on his statistics and methodology (Ritchie et al., 2012). There were attempts at replication and most of them failed, but not all. A year after Bem's article came out, the prestigious journal where it was published, *Journal of Personality and Social Psychology*, published another paper in which a scientist failed to replicate Bem's findings in a number of studies very similar to the originals (Galak et al., 2012). To learn more about this study and its impact, see the video "Is Most Published Research Wrong?" by Veritasium.

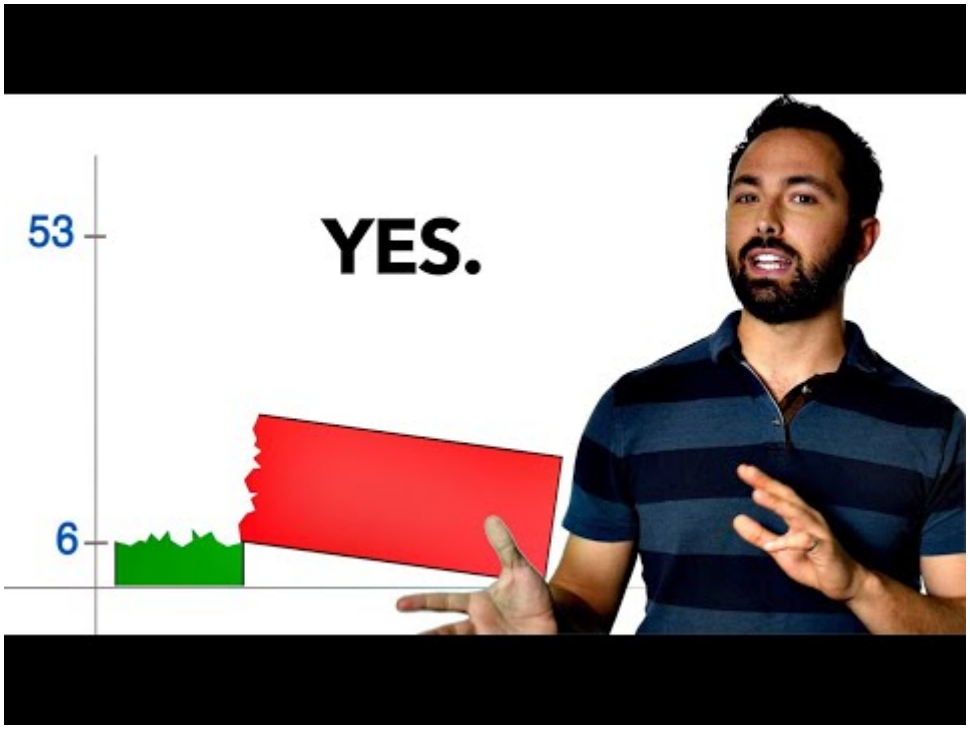


Figure 25.6: Is Most Published Research Wrong?

Some people viewed the publication of Bem's (2011) original study as a failure in the system of science. They argued that the paper should not have been published. However, the editor and reviewers of the article had moved forward with publication because, although they might have thought the findings provocative and unlikely, they did not see obvious flaws in the methodology. We see the publication of the Bem paper, and the ensuing debate, as a strength of science. We are willing to consider unusual ideas if there is evidence to support them; we are open-minded. At the same time, we are critical and believe in replication. Scientists should be willing to consider unusual or risky hypotheses but ultimately allow good evidence to have the final say, not people's opinions.

25.6 Reasons for Non-Replication

When findings do not replicate, the original scientists sometimes become indignant and defensive, offering reasons or excuses for non-replication of their findings—including, at times, attacking those attempting the replication. They sometimes claim that the scientists attempting the replication are unskilled or unsophisticated, or do not have sufficient experience to replicate the findings. This, of course, might be true, and it is one possible reason for non-replication.

25.6.1 Falsified Data

One reason for defensive responses is the unspoken implication that the original results might have been falsified. Faked results are only one reason studies may not replicate, but it is the most disturbing reason. We hope faking is rare, but a number of shocking cases have turned up. One possible early case is Cyril Burt, an educational psychologist. More recently, Diederik Stapel, a social psychologist in the Netherlands, admitted to faking the results of a number of studies. Marc Hauser, a popular professor at Harvard, apparently faked results on animal cognition (Couzin-Frankel, 2014). Karen Ruggiero at the University of Texas was also found to have **falsified** a number of her results (Price, 2010). Each of these psychologists—and there are quite a few more examples—was believed to have faked data. Subsequently, they all were disgraced and lost their jobs.

25.6.2 Small Sample Size and Chance

Another reason for non-replication is that, in studies with small **sample sizes**, statistically-significant results may often be the result of chance. For example, if you ask five people if they believe that aliens from other planets visit Earth and regularly abduct humans, you may get three people who agree with this notion—simply by chance. Their answers may, in fact, not be at all representative of the larger population, in which case you have inadvertently committed a Type I error. On the other hand, if you survey one thousand people, there is a higher probability that their belief in alien abductions reflects the actual attitudes of society. Now, consider this scenario in the context of replication: If you try to replicate the first study—the one in which you interviewed only five people—there is only a small chance that you will randomly draw five new people with exactly the same (or similar) attitudes. It's far more likely that you will be able to replicate the findings using another large sample, because it is simply more likely that the findings are accurate.

25.6.3 Changes Between Samples

Another reason for non-replication is that, while the findings in an original study may be true, they may only be accurate for some people in some circumstances and not necessarily universal or enduring. Imagine that a survey in the 1950s found a strong majority of respondents to have trust in government officials. Now, imagine the same survey administered today, with vastly different results. This example of non-replication does not necessarily invalidate the original results. Rather, it suggests that attitudes may have shifted over time.

25.6.4 Quality of the Original Study or the Replication

A final reason for non-replication relates to the quality of the original study or the replication study. For example, a researcher may "misdesign" a study that contains flaws that are never

reported in the paper. The data that are reported are completely real and are correctly analysed, but they are produced by a study that is actually quite wrongly put together. The researcher really wants to find a particular effect, and so the study is unintentionally set up in such a way as to make it "easy" to (artifactually) observe that effect.

Similarly, errors and misdesign decisions can occur in replications. For example, the newer investigation may not follow the original procedures closely enough. Similarly, the attempted replication study might, itself, have too small a sample size or insufficient statistical power to find significant results.

25.6.5 Questionable Research Practices

25.6.5.1 Data Misrepresentation

While fraud gets most of the headlines, it's much more common to see data being misrepresented– often the data don't actually say what the researchers think they say. It is possible that almost always this isn't the result of deliberate dishonesty, but instead is due to a lack of sophistication in the data analyses. It's very common to see people present "aggregated" data of some kind and sometimes, when you dig deeper and find the raw data yourself, you find that the aggregated data tell a different story to the disaggregated data.

In other cases, a researcher may selectively delete outliers in order to influence (usually by artificially inflating) statistical relationships among the measured variables. In yet other cases, a researcher may selectively report results, cherry-picking only those findings that support their hypotheses.

25.6.5.2 Data Mining and HARKing

Another way in which the authors of a study can more or less misrepresent the data is by engaging in what's referred to as "data mining" (see Gelman & Loken (2013) for a broader discussion of this as part of the "garden of forking paths" in statistical analysis). If one keeps trying to analyse their data in many different ways, they'll eventually find something that "looks" like a real effect but isn't. This is referred to as "data mining," and sometimes can be done without an *a priori* hypothesis. Data mining per se isn't "wrong", but the more that one does it, the bigger the risk being taken. The thing that is wrong and may be very common, is unacknowledged data mining. That is, the researcher runs every possible analysis known to humanity, finds the one that works, and then pretends that this was the only analysis that they ever conducted.

Relatedly, one may "invent" a hypothesis after looking at the data to cover up the data mining. This practice is called HARKing" or hypothesizing after the results are known (Kerr, 1998). Although it's not wrong to change your beliefs after looking at the data, and to reanalyse your data using your new "post hoc" hypotheses, what is wrong is failing to acknowledge what was

done. If you acknowledge that you did it, then other researchers are able to take your behaviour into account. If you don't, then they can't…and that makes your behaviour deceptive. Bad!

25.6.5.3 P-Hacking

Similar to data mining, a researcher may engage in a practice colloquially known as "**[*p*-hacking]**" (briefly discussed in the previous section; Head et al., 2015). This is when a researcher might perform inferential statistical calculations to see if a result was significant before deciding whether to recruit additional participants and collect more data (Head et al., 2015). As you have learned, the probability of finding a statistically significant result is influenced by the number of participants in the study. Additional information on *p*-hacking can be found in the below video, "The method that can "prove" almost anything - James A. Smith".

Figure 25.7: The method that can "prove" almost anything - James A. Smith

25.7 Solutions to the Problem

In addition to highlighting what [not] to do, the so-called "replication crisis" has also highlighted the importance of enhancing scientific rigor by:

- Designing and conducting studies that have **sufficient statistical power**, to increase the reliability of findings.
- Describing one's research designs in **sufficient detail** to enable other researchers to replicate your study using an identical or at least very similar procedure.
- Creating **systematic programs** of scientific research.
- **Publishing both null and significant findings**. This would counteract publication bias (where statistically significant results are published) and reducing the file drawer problem (where null results are not published).
- Conducting **high-quality replications** and publishing these results (Brandt et al., 2014).
- **Preregistering the planned methods** of a study and statistical analyses.
- Engaging in **open science practices**.

25.7.1 Creating Systematic Programs of Scientific Research

The reward structure in academia has served to discourage replication. Many psychologists—especially those who work full time at universities—are often rewarded at work—with promotions, pay raises, tenure, and prestige—through their published research. Replications of one's own earlier work, or the work of others, is typically discouraged because it does not represent original thinking. Instead, academics are rewarded for high numbers of publications, and flashy studies are often given prominence in media reports of published studies. Replicated studies - especially if they find no effects - just simply do not "count" towards researcher's livelihoods.

Psychological scientists need to carefully pursue programmatic research. Findings from a single study are rarely adequate and should be followed up by additional studies using varying methodologies. Thinking about research this way—as if it were a program rather than a single study—can help. We would recommend that laboratories conduct careful sets of interlocking studies, where important findings are followed up using various methods. It is not sufficient to find some surprising outcome, report it, and then move on. When findings are important enough to be published, they are often important enough to prompt further, more conclusive research. In this way scientists will discover whether their findings are replicable, and how broadly generalizable they are. If the findings do not always replicate, but do sometimes, we will learn the conditions in which the pattern does or doesn't hold. This is an important part of science—to discover how ***generalizable*** the findings are.

25.7.2 Dissemination of Replication Attempts

The fact that replications, including failed replication attempts, now have outlets where they can be communicated to other researchers is a very encouraging development, and should strengthen the science considerably. One problem for many decades has been the near-impossibility of publishing replication attempts, regardless of whether they've been positive or negative. There are several places where replication attempts can be published, such as the following:

- Center for Open Science: Psychologist Brian Nosek, a champion of replication in psychology, has created the Open Science Framework, where replications can be reported.
- Association of Psychological Science: Has registered replications of studies, with the overall results published in *Perspectives on Psychological Science.*
- PLOS One: Public Library of Science publishes a broad range of articles, including failed replications, and there are occasional summaries of replication attempts in specific areas.
- The Replication Index: Created in 2014 by Ulrich Schimmack, the so-called "R Index" is a statistical tool for estimating the replicability of studies, of journals, and even of specific researchers. Schimmack describes it as a "doping test".

25.7.3 Textbooks and Journals

Some psychologists blame the trend toward non-replication on specific journal policies, such as the policy of *Psychological Science* to publish short single studies. When single studies are published, we do not know whether even the authors themselves can replicate their findings. The journal *Psychological Science* has come under perhaps the harshest criticism. Others blame the rash of nonreplicable studies on a tendency of some fields for surprising and counterintuitive findings that grab the public interest. The irony here is that such counterintuitive findings are in fact less likely to be true precisely because they are so strange—so they should perhaps warrant *more* scrutiny and further analysis.

The criticism of journals extends to textbooks as well. In our opinion, psychology textbooks should stress true science, based on findings that have been demonstrated to be replicable. There are a number of inaccuracies that persist across common psychology textbooks, including small mistakes in common coverage of the most famous studies, such as the Stanford Prison Experiment (Griggs & Whitehead, 2014) and the Milgram studies (Griggs & Whitehead, 2015). To some extent, the inclusion of non-replicated studies in textbooks is the product of market forces. Textbook publishers are under pressure to release new editions of their books, often far more frequently than advances in psychological science truly justify. As a result, there is pressure to include "sexier" topics such as controversial studies.

25.7.4 Open Science Practices

One particularly promising response to the replicability crisis has been the emergence of open science practices that increase the transparency and openness of the scientific enterprise. Many of these practices include being transparent about how the study was conducted, sharing all materials used for conducting the study, and sharing how data were analyzed. Other practices focus on ensuring that everyone, scientist or non-scientist, can freely learn how science works (as in open educational resources) and can read scientific articles for themselves (as in open access articles). See Figure 25.5 for a list of these principles and video "What is Open Science?" from the National Library of Medicine to learn more.

Figure 25.8: The six principles of open science: open data, open source, open access, open methodology, open peer review, open educational resources.

Figure 25.5 *Principles of Open Science*[6]

Figure 25.9: What is Open Science?

Journals have created incentives rewarding such open science practices. For example, *[Psychological Science]* (the flagship journal of the Association for Psychological Science) and other journals now issue digital badges to researchers who pre-registered their hypotheses and data analysis plans, openly shared their research materials with other researchers (e.g., to enable attempts at replication), or made available their raw data with other researchers (see Figure 25.6).

Figure 25.10: Three open science badges: Open Data, Preregistered, and Open Materials

Figure 25.6 *Open Science Badges - Center for Open Science*[7]

These initiatives, which have been spearheaded by the Center for Open Science, have led to the development of "Transparency and Openness Promotion guidelines" (Grant et al., 2025) that have since been formally adopted by more than 500 journals and 50 organizations, a list that grows each week. When you add to this the requirements recently imposed by federal funding agencies in Canada (the Tri-Council) and the United States (National Science Foundation) concerning the publication of publicly-funded research in open access journals, it certainly appears that the future of science and psychology will be one that embraces greater "openness" (Nosek et al., 2015).

Ultimately, people also need to learn to be intelligent consumers of science. Instead of getting overly excited by findings from a single study, it's wise to wait for replications. When a body of studies is built on a phenomenon, we can begin to trust the findings. Journalists must be educated about this too and learn not to readily broadcast and promote findings from single flashy studies. If the results of a study seem too good to be true, maybe they are. Everyone needs to take a more skeptical view of scientific findings, until they have been replicated.

25.8 Additional Resources

- Article: New Yorker article on the "replication crisis". Article link.
- Blog: The Replication Index estimates the replicability of existing, published studies. Blog link.

- Web: Open Science Framework, created by the Center for Open Science, where replications and preregistration plans can be created. Web link.
- Web: Collaborative Replications and Education Project - This is a replication project where students are encouraged to conduct replications as part of their courses. Web link.
- Web: Commentary on what makes for a convincing replication. Web link.
- Web: The Association for Psychological Science, which publishes *Psychological Science* and *Perspectives in Psychological Science*, two journals describing studies that were published with open science practices such as registered replications, preregistered studies, and open materials. Web link.
- Web: PLOS One by the Public Library of Science publishes a broad range of original research, replication attempts, and summaries of replication attempts in specific areas of science. Web link.

25.9 Discussion Questions

1. Why do scientists see replication by other laboratories as being so crucial to advances in science?
2. Do the failures of replication shake your faith in what you have learned about psychology? Why or why not?
3. Can you think of any psychological findings that you think might not replicate?
4. What findings are so important that you think they should be replicated?
5. Why do you think quite a few studies do not replicate?
6. How frequently do you think faking results occurs? Why? How might we prevent that?

25.10 Vocabulary

- **Conceptual Replication** - A scientific attempt to copy the scientific hypothesis used in an earlier study in an effort to determine whether the results will generalize to different samples, times, or situations. The same—or similar—results are an indication that the findings are generalizable.
- **Confederate** - An actor working with the researcher. Most often, this individual is used to deceive unsuspecting research participants. Also known as a "stooge."

- **Exact Replication** (also called **Direct Replication**) - A scientific attempt to exactly copy the scientific methods used in an earlier study in an effort to determine whether the results are consistent. The same—or similar—results are an indication that the findings are accurate.
- **Falsified data** (or **faked data**) - Data that are fabricated, or made up, by researchers intentionally trying to pass off research results that are inaccurate. This is a serious ethical breach and can even be a criminal offense.
- **Priming** - The process by which exposing people to one stimulus makes certain thoughts, feelings or behaviors more salient.
- **Sample Size** - The number of participants in a study. Sample size is important because it can influence the confidence scientists have in the accuracy and generalizability of their results.

25.11 Media Attributions

[1-6] Diener, E. & Biswas-Diener, R. (2026). The replication crisis in psychology. In R. Biswas-Diener & E. Diener (Eds), *Noba textbook series: Psychology.* Champaign, IL: DEF publishers. http://noba.to/q4cvydeh. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

[7] Center for Open Science, https://www.cos.io/initiatives/badges. Licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) License.

25.12 Text Attributions

Diener, E. & Biswas-Diener, R. (2026). The replication crisis in psychology. In R. Biswas-Diener & E. Diener (Eds), *Noba textbook series: Psychology.* Champaign, IL: DEF publishers. http://noba.to/q4cvydeh. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Modified by current authors.

Jhangiani, R.S., Chiang, I-C.A., Cuttler, C., & Leighton, D.C. (2019). Research methods in psychology. 4th edition. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted. Modified by current authors.

Navarro, D., & Foxcroft, D. (2025). Learning Statistics with JAMOVI: a tutorial for beginners in statistical analysis. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Modified by current authors.

25.13 References

Amir, Y., & Sharon, I. (1990). Replication research: A "must" for the scientific advancement of psychology. *Journal of Social Behavior and Personality, 5*(4), 51–69.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9), 1–70. https://doi.org/10.1037/h0093718

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407–25. https://doi.org/10.1037/a0021524

Biswas-Diener, R., & Diener, E. (2006). Subjective well-being of the homeless, and lessons for happiness. *Social Indicators Research. 76*, 185-205. https://doi.org/10.1007/s11205-005-8671-9

Biswas-Diener, R., & Diener, E. (2001). Making the best of a bad situation: Satisfaction in the slums of Calcutta. *Social Indicators Research, 55*, 329-352. https://doi.org/10.1023/A:1010905029386

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … can't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217-224. https://doi.org/10.1016/j.jesp.2013.10.005

Couzin-Frankel, J. (2014, May 30). Harvard misconduct investigation of psychologist released. *Science News*. https://www.sciencemag.org/news/2014/05/harvardmisconduct-investigation-psychologist-released

Dijksterhuis, A., & van Knippenberg, A. (1998). The relation between perception and behavior or how to win a game of Trivial Pursuit. *Journal of Personality and Social Psychology, 74*, (4), 865–877. https://doi.org/10.1037/0022-3514.74.4.865

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the Past: Failures to Replicate Psi. *Journal of Personality and Social Psychology, 103*(6), 933–948. https://doi.org/10.1037/a0029709

Gelman, A. & Loken E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.* Retrieved from https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf

Grant, S., Corker, K. S., Mellor, D., Stewart, S. L. K., Cashin, A. G., Lagisz, M., ... & Nosek, B. A. (2025). *TOP 2025: An update to the transparency and openness promotion guidelines.* MetaArXiv. https://doi.org/10.31222/osf.io/nmfs6_v2

Griggs & Whitehead (2015). Coverage of Milgram's obedience experiments in social psychology textbooks: Where have all the criticisms gone? *Teaching of Psychology, 42*, 315-322. https://doi.org/10.1177/0098628315603065

Griggs, R. A. & Whitehead, G. I. (2014). Coverage of the Stanford Prison Experiment in Introductory Social Psychology textbooks. *Teaching of Psychology, 41*, 318-324. https://doi.org/10.1177/0098628314549703

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Kahneman, D. (2012). A proposal to deal with questions about priming effects. An open letter to the scientific community. https://www.nature.com/news/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Morehead, K., Dunlosky, J., & Rawson, K. A. (2019). How much mightier is the pen than the keyboard for note-taking? A replication and extension of Mueller and Oppenheimer (2014). *Educational Psychology Review*, *31*(3), 753-780. https://doi.org/10.1007/s10648-019-09468-2

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425. https://doi.org/10.1126/science.aab2374

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). https://doi.org/10.1126/science.aac4716

Price, M. (2010, July/August). Sins against science. *Monitor on Psychology, 41*(7), page 44. https://www.apa.org/monitor/2010/07-08/misconduct

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS One, 7*(3). https://doi.org/10.1371/journal.pone.0033423

Shanks, D. R., Newell, B., Lee, E. H., Balikrishnan, D., Ekelund, L., Cenac, Z., Kavvadia, F. & Moore, C. (2013). Priming intelligent behavior: Elusive phenomenon. *PLoS One, 8*(4). https://doi.org/10.1371/journal.pone.0056515

Urry, H. L., Crittle, C. S., Floerke, V. A., Leonard, M. Z., Perry III, C. S., Akdilek, N., ... & Zarrow, J. E. (2021). Don't ditch the laptop just yet: A direct replication of Mueller and Oppenheimer's (2014) study 1 plus mini meta-analyses across similar studies. *Psychological Science*, *32*(3), 326-339. https://doi.org/10.1177/0956797620965541

Williams, L. E., & Bargh, J. A. (2008). Keeping one's distance: The influence of spatial distance cues on affect and evaluation. *Psychological Science, 19*, 302-308. https://doi.org/10.1111/j.1467-9280.2008.02084

Part VI

Module 6: Communicating Your Outcomes & Growth

26 Chapter 26: Presenting & Publishing Research

26.1 Presenting Scientific Findings

Across the semester, we have focused on how to do a research project from start to finish, including how to select an appropriate research design using various research methods like surveys, observations, experiments, and quasi-experiments; how to conduct a research study well (and ethically); and how to collect and analyze data to test hypotheses. Once you have done all the hard work to design and conduct a study for a course or research lab, now the fun (we hope) part comes in – sharing what you have found!

There are two primary ways that research scientists in many academic disciplines share their work. The first is ***presenting***, which is sharing the work with other people in a live or recorded session, and the other is publishing. In science, ***publishing*** is typically providing an audience with your work in a written form (although those in the arts may publish works of song, art, or other kinds of performances or deliverables). In addition, it is also beneficial to help science inform life by ***sharing*** what we have learned in everyday ways. This can include websites, blog posts, vlogs, social media – and even around the dinner table!

Let's begin with the more formal ways of sharing one's work and close out our online book where we began: Recognizing the importance of science in everyday life, even if research has not become your new life's passion.

26.2 Conferences

When scientists present at conferences, they share their research and other kinds of scholarship with other people in their field. Notice we do not restrict presentations to just "oral presentations," as presenting could also include mechanisms like sign language or text-to-speech language (think famous physicist, Stephen Hawking) (Gernsbacher, 2026).

Figure 26.1: **Figure 26.1** Stephen Hawking being presented by his daughter Lucy Hawking at the lecture he gave for NASA's 50th anniversary[1]

If you are considering going to graduate school, attending or (better yet) presenting at research or teaching conferences is a fantastic way to show graduate school faculty and admissions committees that you are interested in research ideas and trained in research practices. Demonstrating your experience with conducting research (not just being a research subject or participant) is something that a lot of graduate schools value, especially at the doctoral (PsyD or PhD) level.

As examples, Georgia Southern Psychology Department faculty have mentored students who have presented at a wide variety of research conferences, including campus, local, regional, and national venues. Here are just a few examples of recent conferences where Georgia Southern psychology students have presented in recent years or could present. Talk to your psychology statistics and research methods course professor or your research mentor for more information!

- Campus/student conferences:
  - Psychology Department's annual Experiential Learning Showcase
  - Psychology Department's annual Visionary Fair
  - CEPO: Committee for Equality and Professional Opportunity (CEPO)/Psi Chi Undergraduate Research Program
  - GS4: Georgia Southern Student Scholars Symposium
  - GSPS: Georgia Students in Psychological Science (Kennesaw, GA)
  - NCUR: National Conference on Undergraduate Research
  - PURC: Psychology Undergraduate Research Conference (Atlanta, GA)
- Local:

  - FABA: Florida Association for Behavior Analysis (Orlando, FL)
  - GABA: Georgia Association for Behavior Analysis (Atlanta, GA)
  - GPA: Georgia Psychological Association (Atlanta, GA)
  - SETOP: Southeastern Teaching of Psychology (Atlanta, GA)
  - SoTL (Scholarship of Teaching and Learning) Commons (Savannah, GA)
- Regional
  - EPA: Eastern Psychological Association
  - MPA: Midwestern Psychological Association
  - NEPA: New England Psychological Association
  - RMPA: Rocky Mountain Psychological Association
  - SCS: Standard Celeration Society
  - SEABA: Southeastern Association for Behavior Analysis
  - SEPA: Southeastern Psychological Association
  - SSPP: Southern Society for Philosophy and Psychology
  - SSSP: Society of Southeastern Social Psychologists
  - SWPA: Southwestern Psychological Association
  - WPA: Western Psychological Association
- National
  - ABAI: Association for Behavior Analysis International
  - APA: American Psychological Association (also virtual)
  - AP-LS: American Psychology-Law Society
  - APS: Association for Psychological Science
  - IDSP: International Society for Developmental Psychobiology
  - NITOP: National Institute for Teaching of Psychology
  - SPR: Society for Psychophysiological Research
  - SPSP: The Society for Personality and Social Psychology
  - SRCD: Society for Research in Child Development
  - SRA: Society for Research on Adolescence

  - STP-ACT: Society for Teaching of Psychology's Annual Conference on Teaching

- Virtual
  - APA: American Psychological Association (also in-person)
  - OpenEd: Open Education Conference
  - IOCBS: International Online Conference on Behavioral Sciences
  - STP-VCT: Society for Teaching of Psychology's Virtual Conference on Teaching

Although there are many variations of presentation types, at science conferences, researchers primarily present in one of three ways: Posters, symposia, and oral sessions. We will discuss each of these in turn.

26.2.1 Poster Sessions

Poster sessions are a hallmark of research and teaching conferences. They give people a chance to share about their work one-on-one with a colleague, rather than presenting to a room full of people all at once. Presenters may share their work in-person, virtually, or pre-recorded on video. For an in-person poster session, there will be 50 or more posters all gathered in one big space on poster boards. Think elementary or middle school science fair, but without the baking soda volcanos and lights made from potatoes. For the examples below, please follow the links from the presentation images so you can zoom in and actually read the information, rather than just see the big picture.

Introducing the Online Career Exploration Resource (OCER) 2.0: Helping Students Search and Learning About Psychology-Related Careers

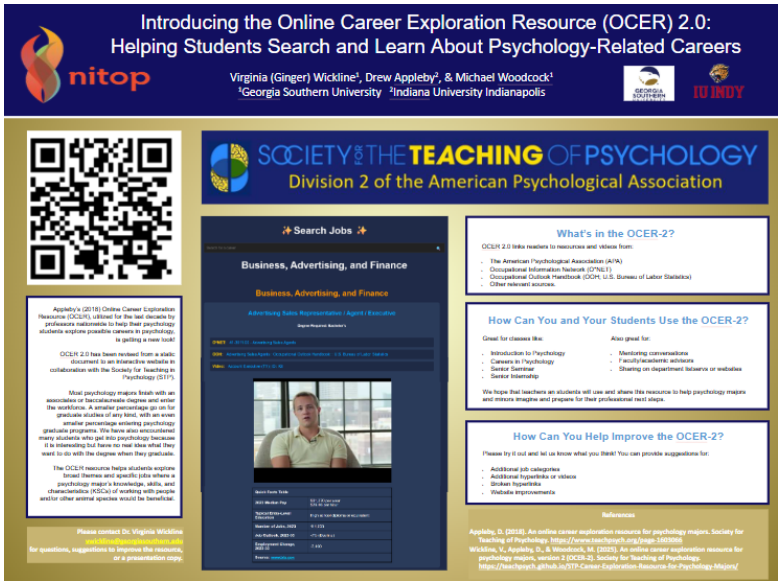


Figure 26.2: **Figure 26.2** The Online Career Exploration Resource (OCER) 2.0 to help psychology students explore careers was introduced by Virginia Wickline, Drew Appleby & Michael Woodcock at the National Institute for Teaching of Psychology (NITOP) conference[2]

A ***research poster*** will be one big slide (about 3 x 4 feet or 4 x 6 feet, depending on the conference's requirements) that includes key highlights about the background literature and hypotheses, the method (participants, measures or materials, procedure), a good bit of information about the study results, and a few closing points about the discussion (take-away messages, limitations, next steps). This kind of poster will have more detail than other types of posters, so people could use it as a resource (view it or cite it) after the conference.

Crossing Borders: Impacts of a Collaborative Online International Learning (COIL) Project

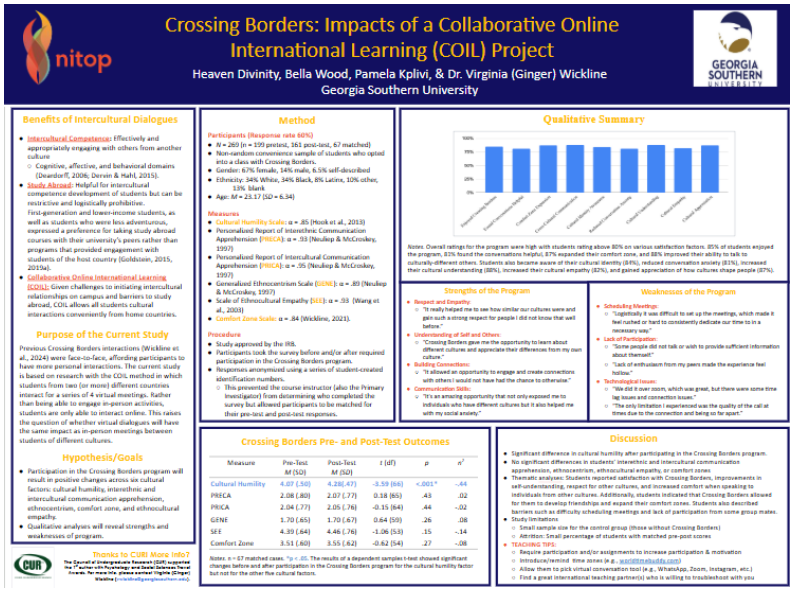


Figure 26.3: **Figure 26.3** Example of a research poster presented by psychology alumni Heaven Divinity, Bella Wood, and Pamela Kplivi for the National Institute for Teaching of Psychology (NITOP), demonstrating that conversing with people in other cultures increases cultural humility[3]

Wooster-Wickline College Adjustment Test (WOWCAT): A Reliable and Valid Scale

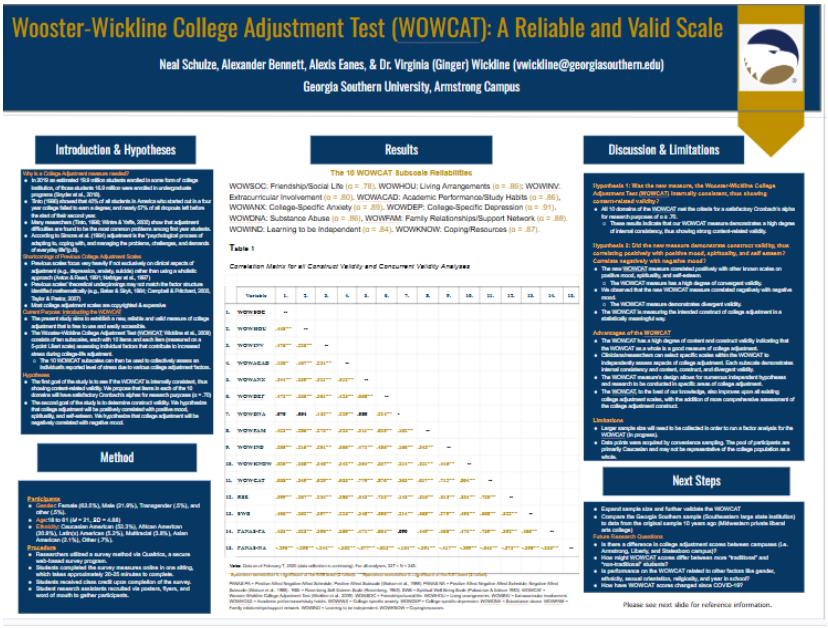


Figure 26.4: **Figure 26.4** Georgia Southern alumni Neal Schulze, Alexander Bennett, and Alexis Eanes display a correlation matrix validating the factor structure of the Wooster-Wickline College Adjustment Test (WOWCAT), which is frequently used as the basis for research projects in Dr. Wickline's research design & analysis lab course[4]

Greater Gains in Intercultural Competence: Study Abroad or at Home?

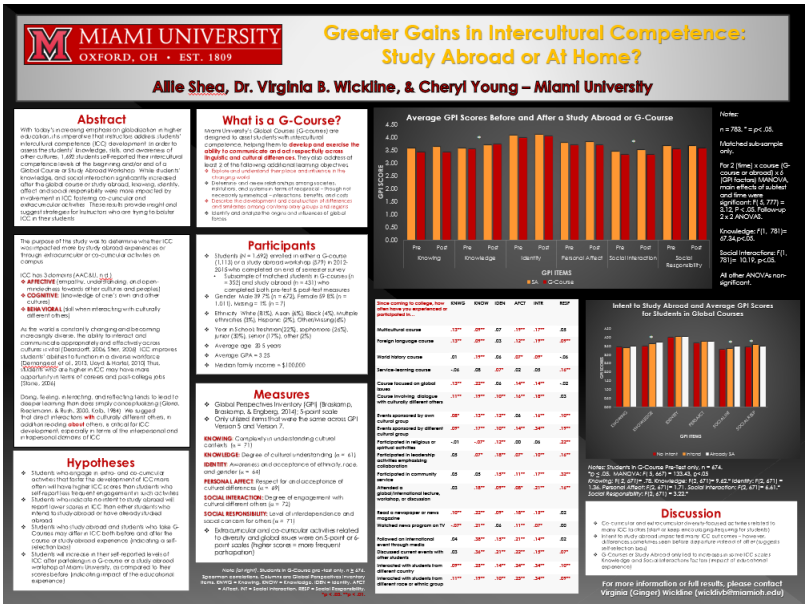


Figure 26.5: **Figure 26.5** Dr. Wickline's former student, Allie Shea, presented her honor's thesis on the benefits of study abroad for the NAFSA (Association of International Educators) international conference[5]

"So…tell me what you did with your project." At a ***poster session***, people walk around and engage by reading the poster and/or talking to the presenters about their work for a 60-90 minute session. Sometimes the presenters answer people's specific questions after they read the poster, and sometimes they give a brief synopsis (overview) of the project and then answer follow-up questions. There are also virtual programs like Gather that let you (through your little self-designed avatar) wander around a virtual space to view posters and then chat with the real presenter in a virtual meeting room before moving on to the next poster of interest.

Figure 26.6: **Figure 26.6** Kaylee McCook (undergraduate), Shea Hall (Experimental Psychology Master's program), Dr. Viriginia Wickline, and Ryan Lavrisa (undergraduate) present their research on mask-wearing attitudes by political party for the Southeastern Psychological Association (SEPA) Conference[6]

A good poster is one that is concise but still has the information needed to help the audience get the key points and findings. The design and delivery are two important aspects you will want to consider for your poster presentation. The first goal of a poster presentation is to get people to want to stop and see your work. Thus, the design should be simple and eye-catching, not dull, distracting, or big and flashy. It should not be too crowded or messy, nor should the font be too small or too difficult to read. It should not have any typographical errors (typos), slang, or grammatical errors. The second design goal is not to have big, dense, text-heavy paragraphs but rather ***bullets*** (lists) of information. Keep it short and precise; have room for some white or blank space around your words. The third design goal is to use colors, graphs, or imagery to make it interesting. Try your best to make sure you are using images that are not copyrighted, and give credit for the images you use. Look for Creative Commons images where possible. Wikimedia Commons is a great source for images with Creative Commons licenses. If you are using Google images to search, use the Tools tab, then Usage Rights to search for images with Creative Commons licenses.

In terms of delivery, do not read verbatim (word-for-word) off the slides (it is OK to have a page or notes or some notecards if you are a new or nervous presenter). The second delivery goal is to have a conversation. Poster presentations are not a lecture – they are a base to jump-start a dialogue. They are a chance to share ideas, take and ask questions, and perhaps learn more about the person who has come to see your poster. Perhaps your poster viewer is not just a professor who is being kind by giving nervous student presenters a chance to practice. If not, then people who stop at a poster usually have some specific interest in the topic or also conduct work in the same research area. Thus, the third delivery goal is to use your poster as a potential chance to ***network***, to build connections with other professionals in your field or

area of interest. You never know who you might meet at your poster – maybe a friendly face, an eager undergraduate attending their first conference, a potentially new research colleague, someone who is hiring in your research area, or a professor who teaches at the graduate school you hope to attend. Finally, the fourth delivery goal, as cheesy as it sounds, is to ***have fun***! You have worked hard on your research, and now it is time for show-and-tell! You might also get some good questions that help you reconsider your findings or address concerns before pursuing publication of your work in a manuscript form (see Publications: Books, Book Chapters & Journal Articles, below).

26.2.2 Symposia

A ***symposium*** (the plural of symposium is symposia) is an hour-long block that is typically broken up with 2-3 presenters (solo or group). Each presenter has about 15-20 minutes to share their research project, and some time is left at the end for questions from the audience. Research presenters will typically use a slide deck, whether that be Google Slides, PowerPoint, Canva, Prezzi, or a similar design program. A symposium presentation has the same sections as a poster (see above) or written research manuscript (see below), with a few slides about the background literature and hypotheses, then the method (participants, measures/materials, procedure), a good bit of slides about the results, and a few closing slides about the discussion. Depending on the conference, symposium presenters may share their work in-person, virtually, or pre-recorded.

As an example, please check out this slide deck presentation about predictors and consequences of food insecurity (not having enough money for food) for Georgia Southern students. This project, which is currently under review for publication (as of 2026), was conducted by a research team from the Armstrong Campus as part of PSYC 3729 (Service-Learning in Psychology) and PSYC 3900 (Independent Study in Research).

Examining Food Insecurity on College Campuses

Figure 26.7: **Figure 26.7** Opening slide of a symposium slide deck by alumni Rachel Baker, Destiny Truth, Rashuna Middleton & Sharon Hathway presented for the Southeastern Psychological Association (SEPA) Conference[7]

Below is what a pre-recorded symposium presentation would look like for a virtual conference. If the conference is in person, the presentation is given in front of a live audience instead.

The Eyes Don't Have It: Masks Make it Very Hard to Read Emotion in Most Facial Expressions

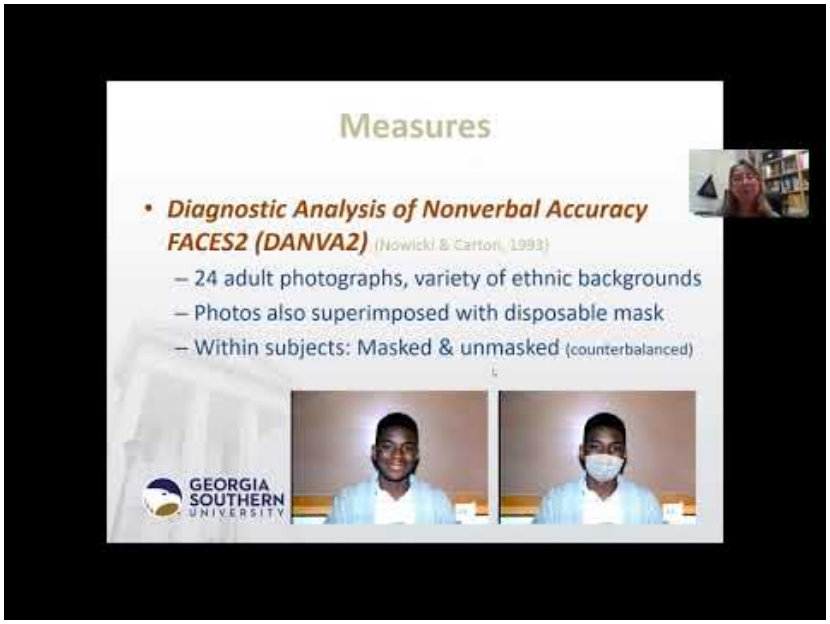


Figure 26.8: The Eyes Don't Have It: Masks Make it Very Hard to Read Emotion in Most Facial Expressions

Symposium presentations can also take the form of a panel of speakers. In this case, instead of

presenting research, the presenters have a common topic, and a moderator (coordinator) asks predetermined questions of the panel members or collects questions from the audience.

26.2.3 Oral Sessions

Once scientists have been doing research for a while, they are sometimes invited to do a longer talk of 30-60 minutes for a larger audience, sometimes called an ***oral session, plenary session, or keynote talk***. Oral sessions will usually be a collection of broader ideas or studies, rather than focusing on one specific research study. Here is an example from one of our faculty members, Dr. Virginia Wickline, who was asked to speak on the similarities and differences between counseling (therapy) and research mentoring for the Council on Undergraduate Research (CUR). Keep in mind that if you are recording a video, it should have a written transcript or ***captions***, which are text-based descriptions that people can read while watching the video. Captions are not just for people with auditory impairments, like people who are deaf or hard of hearing. They help comprehension for second language speakers, older adults, neurodivergent individuals (those with conditions like autism or ADHD), and…well…just about everyone (Gernsbacher, 2015).

Talk to Me: How Undergraduate Research Mentoring Parallels and Diverges from Counseling Interventions

Figure 26.9: Talk to Me: A Webinar with the 2025 CUR Psychology Division Mid-Career Mentor Awardee

26.3 Creating and Delivering a Good Oral or Video Presentation

Desjardins (n.d.) has some wonderful suggestions for making a presentation more engaging to the audience. Before you even begin to speak, a big part of presenting is visual design. Take the time and effort to make it look nice. There are lots of design tools and templates out there these days, whether in PowerPoint, Canva, or Prezzi.

Please check out: Steal This Presentation! You can review or download Desjardins' slides for future reference.

Figure 26.10: **Figure 26.8** Opening image for "Steal This Presentation" slide deck by Jesse Desjardins on presenting effective PowerPoint slides[8]

A good symposium or oral presentation is about making the audience comfortable and ready to hear your work, plus delivering clear talking points that engage the audience and keep their attention. This is true whether or not you will be seen by the audience (as some oral presentations are voiceovers or screencasts of your computer screen). Keeping your listeners awake and not bored is important too! Gernsbacher (2026) notes the difference between a presentation that is designed to be read when the author is not present, which will have more detail, and a presentation made to deliver to the audience, which will have more key words and ideas but fewer details. In either case, presenting is story-telling. Good stories have a beginning, middle, and end. In a research talk, this is the background and hypotheses (beginning), what you did and found with your method and results (middle), and a summary of why we should care (end).

Please watch: How to Give an Awesome (PowerPoint) Presentation

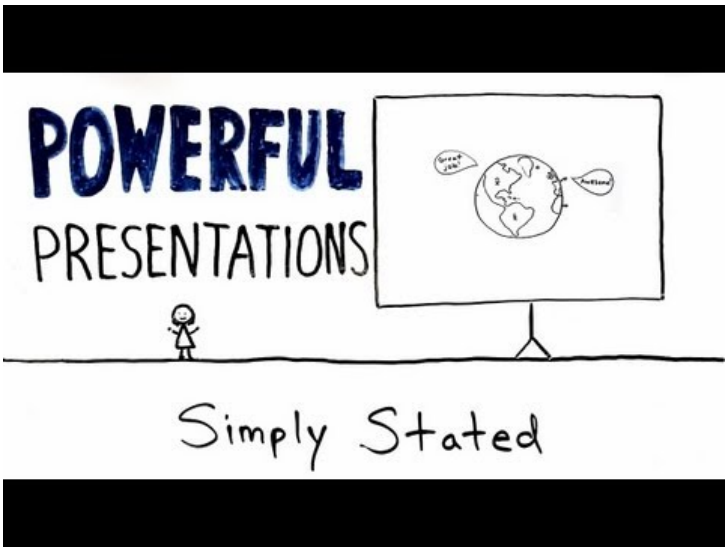


Figure 26.11: How to Give an Awesome (PowerPoint) Presentation (Whiteboard Animation Explainer Video).

Please watch: The Do's and Don'ts of Effective Presentation Slides

Figure 26.12: The Do's and Don'ts of Effective Presentation Slides

Once your presentation is ready, it is time to share it! If the thought of speaking in front of an audience makes your knees weak and your heart pound, you are not alone! Ballard (2026) suggests that 25-75% of people have mild, moderate, or severe ***glossophobia*** (speech anxiety).

However, in the professional world, learning to share one's ideas both formally and informally is a very important skill to develop. Good public speaking skills can help you land a job with a polished interview. They can also help you keep a job if you are working in teams or if sharing ideas with customers or clients is part of your work responsibilities.

What kinds of behaviors make for an engaging presentation? Here are a few ideas to get you started (this list is by no means comprehensive):

- ***Practice makes perfect!*** Make sure you have run through your presentation out loud at least once, whether you speak to an empty room, watch yourself in a mirror, grab a friend or loved one to listen to you, or deliver your presentation to your fur baby. This will help you see typos in your slide deck, find those spots you get stuck on when talking, stay within the time limits, and work through your nerves.
- ***Take a deep breath.*** It is OK and fairly normal to be nervous. Over 100 years ago, researchers established the well-replicated ***Yerkes-Dodson law***, which suggests that there is a positive relation between anxiety (arousal) and performance…but only up to the point where it gets overwhelming, which is where we can crash and burn (Yerkes & Dodson, 1908). So keep in mind: Some nerves are good because it means you care, and it will motivate you to try hard enough so you do not look goofy. Stay in the optimal zone! Try something like box breathing, diaphragmatic breathing, mindfulness, or closing your eyes for a moment before you begin to settle your nervous system.

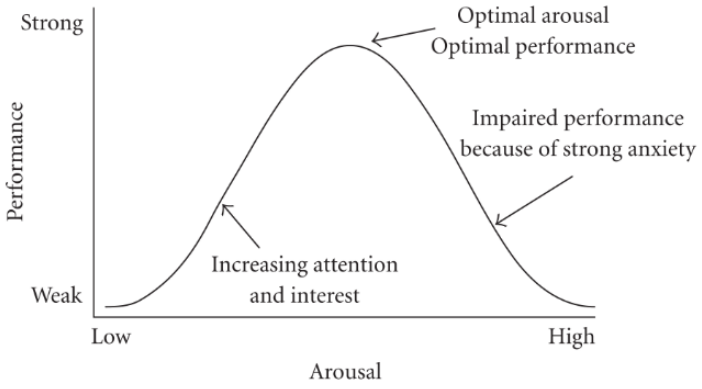


Figure 26.13: **Figure 26.9** Hebbian version of the Yerkes Dodson curve[9]

- ***Make eye contact with your audience.*** You should know your presentation content well enough that you are not just reading off the slides. It is OK to have notes in case you get nervous, but do not just straight read off those either.
- ***Project your voice.*** Like being on stage (if you were acting), be sure to speak loud enough that people in the back of the room can hear you. In fact, perhaps even think of

public speaking like acting. If you are not feeling confident, act like you are confident anyway.

- ***Speak a little slower than you think you need to***. Most beginning level speakers rush through their slides at a pace that makes it hard for an audience to keep up.
- ***Smile. Smile again. Smile some more.*** This is both to engage your audience and to settle your nerves. It is hard to pass out or cry if you are smiling. If you smile, it will also affect your voice in a positive way. Fake it 'til you make it and are actually having fun!
- ***Remember that your audience chose to be there.*** Well, unless y'all are required to be presenting to each other for a class assignment. Even then, everyone is expected to be there, and they are probably as nervous or more nervous than you. If you are at a conference, people paid money to be there, so they are most likely going to be supportive (or at least not rude) when you share. They want to hear you and want it to go well.

Take a look at one or more of these engaging presentations (which do not even use a slide deck). Keep in mind while you are watching to look for the beginning, middle, and end, even in a brief presentation.

The Mystery of Asparagus Pee

Figure 26.14: The Mystery Of Asparagus Pee

Do Fetuses Poop?

Figure 26.15: Do Fetuses Poop?

How Different are Different Kinds of Dogs?

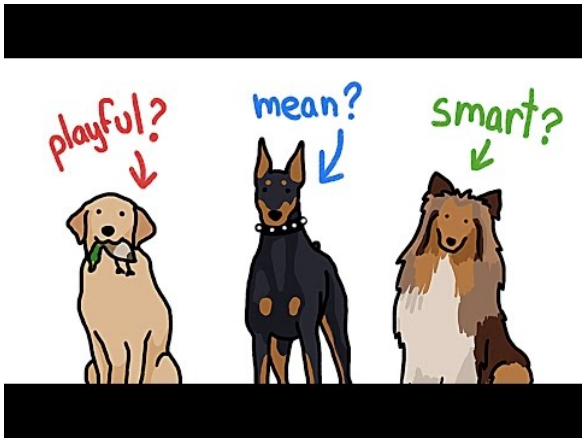


Figure 26.16: How Different Are Different Types of Dogs?

Jamila Lyiscott: 3 Ways to Speak English

Figure 26.17: Jamila Lyiscott: 3 ways to speak English | TED

Malcolm Gladwell: Choice, Happiness & Spaghetti Sauce

Figure 26.18: Choice, happiness and spaghetti sauce | Malcolm Gladwell

Are you going to be a Malcolm Gladwell or Jamila Lyiscott with your first presentation? Heck, naw…but everyone has to start somewhere. Also, a straight-up research presentation may not be nearly as exciting as some of these more conversational examples…but all the same, learning how to present your ideas is good for your personal and professional growth and important for getting your work out there in the world. Billionaire Warren Buffet said public speaking is the single-biggest skill you can have to boost your career (Gallo, 2017). So, take a deep breath and embrace it! Remember: We have the chance to grow the most in the spaces where we are uncomfortable.

26.4 Publications: Books, Book Chapters, and Journal Articles

After or instead of presenting research at a conference, scientists also try to share their written ideas and findings in ***publications*** like books, book chapters, and journal articles. Mainly, these kinds of works are for other academics – not the kind of light reading you would typically take with you to the beach, backyard hammock, or campsite. When theories and hypotheses are captured in written form, this is otherwise known as ***scholarship***. Although scholarship is more often done by those with a graduate degree, you might get the opportunity to try to publish research as an undergraduate if you did an honor's thesis or worked in a professor's research lab – this would be looked upon very favorably by most graduate school programs. Scholarship can be produced in traditional venues via publishers who produce books and journals for profit. In the case of books, the author gets some royalties (payment) when the book sells so many copies. For journal article publications, the author does not usually receive profits, but they may benefit professionally from gaining ***citations*** to their work when other people refer to the author's work in their own. Most authors find it exciting that others are reading what they wrote. If they have stayed in ***academia*** (becoming a psychology professor), these citations are an important metric of the quality of their work, because it gives an idea of how many people find their ideas useful and worth discussing. In today's world there are also options for open access scientific publishing (usually at a cost to the author), open educational resources like this textbook (no publisher, no profits), and self-publishing.

26.5 Popular Press, Social Media, and the Dinner Table

Another important skill to learn is how to translate scientific findings in an easy-to-digest fashion so that they can be used in everyday life. Your friends and family may think you are awesome, but we're sorry to say that most people still do not want to read your research paper or talk about your poster in depth, especially at family dinner, during game night, or in your social media postings. In those contexts, short explanations in simple, everyday language are better. Here is what sharing a research finding might look like in four different contexts.

> *Research Manuscript or Symposium: "Research shows that infants who are securely attached to their primary caregiver(s) will use them as a secure base from which to explore their environment and a safe haven in times of duress and potential danger."*
>
> *Poster: "Infant attachment to caregiver serves two purposes: Secure base, safe haven"*
>
> *Dinner Table: "My project showed that babies love parents who make them feel secure enough to go explore new things but safe when they are scared."*
>
> *Instagram: "Loving mom (dad) = baby feels safe, explores."*

I sometimes explain to my research design and analysis students that when they finish summarizing their data, they should "translate math to English" or "explain it to my nana, my nephew, my neighbor who does not speak math." In a world of political echo chambers, plus social media and trackers that feed us certain algorithms to see what we want to see and convince us to buy all kinds of stuff, it is very important to keep in mind your audience and their communication preferences and adapt accordingly when presenting scientific findings. Make it digestible and appropriate for those with whom you are sharing your information. Don't bore your barber with voluminous research details while you are getting your haircut; give him the highlights instead. In other words, you might need to ***code-switch*** (see again Ms. Lyiscott's TED talk), to adapt how you speak to fit a particular context or audience.

Figure 26.19: **Figure 26.10** Barber asleep while at work – don't bore your barber with your voluminous research details[10]

26.6 Getting Additional Research Experience

If you are at Georgia Southern University, when you are finished with your research design and analysis courses, you could develop additional research experience through PSYC 3900 (Independent Research) where you work on a faculty member's projects in their lab, with the McNair Scholars Program (open to students who are first-generation, have low income, or have a disability), or while completing an undergraduate thesis as an Honors College student. Getting these kinds of research opportunities as an undergraduate requires initiative and hustle – putting yourself out there. These are competitive and somewhat limited opportunities. Faculty might have hundreds of students in their courses in a semester, but just two in their research lab. You most often need to take the lead in reaching out to professors to see how their lab works and whether they are taking students. Although it is possible to do research for just a semester, most faculty members like for students to join their research lab for at least a year, perhaps longer. If you want to design your own project as a senior, for example, you would typically join someone's lab in your first-year, sophomore, or junior year to get training

on research, finish your research design and analysis course sequence, and then have time to design and conduct a study of your own (which almost always is a 2-3 semester process).

If you are reading this text at another institution of higher education, be sure to talk to your professors, department leadership, and academic advisors about the research opportunities available to you in your program and at your college or university.

26.7 Summary

Once all the hard work of research design and analysis is done, hopefully the exciting part is getting to share what you found with others, even if you did not find what you expected. Whether this is done with a poster, symposium, or written publication—or even at the dinner table—try to enjoy the sharing phase of research work.

26.8 Additional Video Helps

Giang, V. (2026, February 27). *In 7 minutes, you'll be 170% better at presentations.* https://www.youtube.com/watch?v=Sh-se-i0afA

Practical Psychology (2017, January 16). *How to give a great presentation – 7 presentation skills and tips to leave an impression.* https://www.youtube.com/watch?v=MnIPpUiTcRc

TED-Ed (2025, July 15). *How to communicate clearly.* https://www.youtube.com/watch?v=btWlBHE0pe4

26.9 Additional Resources

Borup, J. (2021, February 3). *Putting your best self forward: 6 keys for filming quality videos.* Educause Review: The Voice of the Higher Education Technology Community. https://er.educause.edu/blogs/2021/2/putting-your-best-self-forward-6-keys-for-filming-quality-videos

Feldman, D. B., & Silvia, P. J. (2011). Don't do that! Five things to avoid when planning your first conference speech. *Eye on Psi Chi, 15*(4), 23-25. https://doi.org/10.24839/1092-0803.eye15.4.25

Gernsbacher, M. A. (n.d.). *PSY 225 (Professor Gernsbacher): How to use Google image search to find non-copyrighted images.* https://online225.psych.wisc.edu/wp-content/uploads/225-Master/225-UnitPages/Unit-13/PSY-225_GoogleImageSearch.pdf

Kapterev, A. (n.d.). *Death by PowerPoint (and how to fight it).* https://www.slideshare.net/slideshow/death-by-powerpoint/85551

Psi Chi: The International Honor Society in Psychology (n.d.). *Resource: Attending and presenting at conferences*. https://www.psichi.org/page/RES_ConvPresent

Psi Chi: The International Honor Society in Psychology (n.d.). *Resource: Paper presentations*. https://www.psichi.org/page/RES_ConvPapers

Psi Chi: The International Honor Society in Psychology (n.d.). *Resource: Poster checklist, templates, and samples*. https://www.psichi.org/page/RES_ConvPosCheck

Psi Chi: The International Honor Society in Psychology (n.d.). *Resource: Tips for attending conventions*. https://www.psichi.org/page/RES_ConvTips

26.10 Media Attributions

[1] (Public Domain via The Commons by NASA via Wikimedia Commons)

[2-7] Figure 26.2-26.7. (Picture by Dr. Virginia Wickline, Georgia Southern, licensed under CC BY-NC-SA 4.0.)

[8] ("Steal This Presentation" slide deck by Jesse Desjardins available on slideshare)

[9] (CC0 1.0 Universal by Quibik via Wikimedia Commons)

[10] (CC BY 2.0 by McZustaz via Wikimedia Commons)

26.11 Text Attributions

26.12 References (including Video Attributions)

Ballard, J. (2026). Do you have a fear of public speaking? *Psychology Today*. https://www.psychologytoday.com/us/blog/expanding-our-potential/202601/do-you-have-a-fear-of-public-speaking

Council on Undergraduate Research (2026, March 25). *Talk to me: How undergraduate research mentoring parallels and diverges from counseling interventions*. A webinar with the 2025 CUR Psychology Division Mid-Career Mentor Awardee. https://www.youtube.com/watch?v=J0yIjpupXCI&t=330s

Desjardin, J. (2010, August 23). *Steal this presentation!* https://www.slideshare.net/slideshow/steal-this-presentation-5038209/5038209

Gallo, C. (2017, January 5). Billionaire Warren Buffet says this 1 skill will boost your career value by 50 percent. *Inc.Com.* https://www.inc.com/carmine-gallo/the-one-skill-warren-buffett-says-will-raise-your-value-by-50.html

Gernsbacher, M. A. (2026). *PSY225: Research Methods. Unit 13: How to communicate through presentations.* https://online225.psych.wisc.edu/unit-13/

Gernsbacher, M.A. (2015). Video captions benefit everyone. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 195-202. https://doi.org/10.1177/2372732215602130

Gladwell, M. (2007, January 16). *Choice, happiness & spaghetti sauce.* TED. https://www.youtube.com/watch?v=iIiAAhUeR6Y

Kwan, C. (2019, November 19). *The do's and don'ts of effective presentation slides*. https://www.youtube.com/watch?v=tEF2vNP3S9A

Lyiscott, J. (2014, June 19). *Jamila Lyiscott: 3 ways to speak English.* TED. https://www.youtube.com/watch?v=k9fmJ5xQ_mc

MinuteEarth (2016, April 14). *The mystery of asparagus pee*. https://www.youtube.com/watch?v=8ySf9jn6cmM

MinuteEarth (2013, June 28). *Do fetuses poop?* https://www.youtube.com/watch?v=qc-244lKjjM

MinuteEarth (2016, October 10). *How different are different kinds of dogs?* https://www.youtube.com/watch?v=c7puPXZVsFQ

Wickline, V. (2021, June 8). *The eyes don't have it: Masks make it very hard to read emotion in most facial expressions.* Oral presentation for the 2021 Southeastern Psychological Association. https://www.youtube.com/watch?v=h1DKnH6XoEI&t=20s

Wienot Films (2011, May 9). *How to give an awesome (PowerPoint) presentation*. https://www.youtube.com/watch?v=i68a6M5FFBc

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18*(5), 459–482. https://doi.org/10.1002/cne.920180503

27 Chapter 27: Conclusion: Putting Your Research & Professional Skills Into Practice

27.1 Can You Think Like a Psychological Scientist?

So, to recap our work together, why do psychology majors need to understand statistics and research? Way back when we began the course, remember that the primary goal of learning research designs and statistical methods was to help you think like a psychological scientist, whether or not you develop an interest in becoming involved with research in your future career(s). Being able to "speak math" (otherwise known as ***quantitative literacy***) is central to what psychology students and professionals do. Being able to think critically, ground your decisions in data and evidence, and test ideas for their accuracy and relevance is as important (or more important!) in everyday life and work as it is in research.

Now that you can think like a researcher, a review of why this is important would seem to be in order. First, we will look at why the American Psychological Association (APA)—the guiding force of psychology education in the United States—thinks these kinds of skills are important. Then, we will specify more of the skills that employers and graduate schools are looking for from psychology majors. Lastly, we will look at how you can market yourself to employers and graduate schools by describing effectively, both in writing and orally, the skills they want to see from their applicants.

27.2 The American Psychological Association (APA) Guidelines for Undergraduates

The ***American Psychological Association (APA)*** is the governing body for psychology professionals in the United States of America. Whether we are using psychology in teaching, research, or mental health service provision (e.g., therapy), psychologists and psychology students follow the APA guiding policies and principles for how we should go about our work (APA, 2019). For example:

- How should we conduct ourselves as psychology professionals and treat people at our work or in our care? See the APA Ethical Principles and Code of Conduct (APA 2017a).

- How do we consider diversity and multicultural practices in professional psychology? See the Multicultural Guidelines: An Ecological Approach to Context, Identity, and Intersectionality (APA, 2017b).
- What words should we use to describe people in our classwork, research, and therapy settings? See the APA Inclusive Language Guide (APA, 2023a).
- What should we teach undergraduate psychology majors? See the APA Guidelines for the Undergraduate Psychology Major, Version 3.0 (APA, 2023b).

Of prime importance here is the APA Guidelines for the Undergraduate Psychology Major, Version 3.0 (APA, 2023b), which has specific guidelines for what universities should be teaching undergraduate psychology students. These guidelines include five primary goals:

- [Content Knowledge and Applications] – Psychology students should understand major themes and concepts that are relevant to psychology and use this knowledge to address personal and societal-level problems.
- [Scientific Inquiry and Critical Thinking] – Students should develop proficiency in research methods and statistics to use scientific reasoning and investigation as a way to draw conclusions.
- [Values in Psychological Science] – Psychology students should understand how individual and group differences influence people's behavior. They should also exhibit ethical behavior and intercultural sensitivity to act in socially responsible ways.
- [Communication, Psychological Literacy, and Technology Skills] – Psychology students should be able to use written, oral, and technology communication skills to build relationships, discuss scientific understandings, and share their work in a wide variety of contexts.
- [Personal and Professional Development] – Psychology students should be able to self-regulate their work so they are responsible and self-aware, can work well with others in professional settings, and can develop plans for their professional life after college.

As you can hopefully see, your research design and analysis courses were intended to address most if not all of these APA goals for undergraduates. But even if we (your Psychology Department) do well at planning opportunities for your growth as a psychology major, it will be up to ***YOU*** to convince a future employer or graduate school that you have what it takes to be a successful student or employee. Simply being a psychology major is not necessarily going to be convincing. Neither is wanting to help people, figure out your family, or really loving crime shows. And to be honest…while getting good grades is a good starting point, it won't get you hired either. You need to be doing more than going to class and studying if you want to make a living from your psychology degree.

By now, we hope you have figured out that psychology majors go into a lot of different fields. The APA's (2011) "Careers in Psychology" publication is well worth a deeper look if you are trying to

figure out what to do with your psychology degree once you graduate. It notes how psychologists do a wide variety of things, to include promoting well-being, advising policymakers and business, teaching and studying learning, applying science, conducting research, and providing healthcare. Yes, of course, some psychology majors end up being therapists or counselors, but not all psychology students end up in the helping professions. In fact, U.S. Bureau of Labor Statistics Occupational Outlook Handbook notes that the top five occupational groups that psychology degree holders wind up in are management occupations (15%), community and social services (13%), teaching & library occupations (12%), healthcare and technical occupations (11%), and business and finance (10%), with 39% of psychology degree holders working in other fields. You might be surprised to learn that only 6% of psychology majors end up as therapists or counselors of some kind, with another 3% becoming social workers. Notice that the other 90% of psychology majors ended up doing something besides counseling for their careers.

Recognize also that a psychology degree is really popular. The National Center for Education Statistics (NCES, 2024) showed that in 2021-2022, of the two million bachelor's degrees awarded in the United States, 6% of them (nearly 130,000) were psychology degrees, landing as the fifth most popular degree granted. According to the APA's Center for Workforce Studies (CWS, 2021), 3.7 million people held a bachelor's degree in psychology in 2019. Half a million of these (about 14%) obtained a higher degree in psychology (master's or doctoral degree), over a million people (about 28%) obtained higher education degrees in other fields, and roughly two million people (about 57%) entered the workforce. The U.S. Bureau of Labor Statistics Occupational Outlook Handbook (2026) suggests the 2023 median annual wage for those 3.7 million people who work in the field of psychology was about $60,000.

How are you going to stand out from the crowd? How are you going to convince a graduate school that you are ready for a more intensive program? How are you going to persuade an employer to hire you and not the next person on their interview schedule?

Let's break down these five important goals that the APA has for you in some more specific ways.

27.3 The Skillful Psychology Student

In 2018, the APA's Committee on Associate and Baccalaureate Education (CABE) commissioned a task force to put together "The Skillful Psychology Student," a list of the transferable skills that psychology students need to be successful in the 21st Century (Naufel et al., 2018). The five broad skill domains captured in their list also include more specific subskills:

1. thinking, creativity, information management, judgment and decision-making
2. communication
3. self-regulation (time management)

4. service orientation
5. familiarity with hardware and software

Again, your psychology degree is designed with these skills and your personal and professional growth in mind.

27.4 Why Skills Matter: Describing Your Transferable Skills to Graduate Schools and Employers

Perhaps no other courses in the major capture as many of these skills as the research design and analysis course sequence. These skills will hopefully help you professionally (perform at your job more successfully) and personally (live a better life). The challenge for you, though, is just taking the courses is not enough to prove that you have developed these skills. As Appleby et al. (2019) note, several ways exist to demonstrate your transferable skills. The first is to earn high grades in your courses that tap into these skills (say, an A or a B in your research design and analysis courses). You could then use a manuscript you produce from a final project in the lab portion of the course as a writing sample for an interested employer or graduate school admissions committee. Combined in an electronic portfolio or dossier with strong letters of recommendation from several of your professors who speak to your skills, along with a copy of any presentations (poster or slide deck, including where the presentation was delivered) or publications, would make an even more convincing pitch.

27.5 Professional Documents

Be ready to weave these transferable skills into your descriptions in your own professional documents for employment possibilities or graduate school admission. To do this well, you need to know the skills that are valued and then explain how your coursework helped you refine those skill sets, not just list them. What professional documents? So glad you asked!

27.5.1 Professional Documents for Jobs

For jobs, you will need two professional documents: A résumé and a cover letter.

Résumé: A résumé is a brief document that summarizes your contact information, education, awards and honors, work and volunteer experiences, and skills or certifications. Most résumés are written only on one side, but some job fields will allow front and back, especially if you have had an extensive career. A résumé should be uniquely tailored to each individual position or type of position. Your résumé for a hotel concierge position might contain different things

than your résumé as a paraprofessional in an elementary school or your résumé for a mental health technician opening.

- ***NOTE***: When you are in college, you can include high school information and extracurriculars, but be ready to jettison all but exceptional highlights from high school on your résumé once you complete college. Thus, it is important to keep building your experiences across college in preparation for this transition.
- ***NOTE***: Do not use a template for making your résumé. Type it yourself, especially if you are submitting your document online. Many businesses now use Applicant Tracking Systems (ATS), an AI-generated screener before it gets to a person, and they tend to kick out template-based résumés (Purcell, 2026).

Cover letter: A cover letter is a formal introduction and expression of interest in a given position. It should supplement your résumé, not duplicate it, and entice a reader with some extended examples or highlights so they want to look more into your qualifications on your résumé.

27.5.2 Professional Documents for Graduate School

Graduate school is more rigorous than undergraduate, shifting you from consumer to producer of information and focusing more on skills, rather than primarily knowledge (Freis & Kraha, 2016). If you want your application to graduate school to be considered, you will need two primary professional documents: A curriculum vitae and a personal statement.

Curriculum vitae (CV): A curriculum vitae is an academic list of qualifications. It is used only in academically related positions, like graduate school, college teaching, or possibly research jobs or grants. A CV typically has sections like: contact information; educational history; professional experience; presentations and publications; honors, awards, and memberships; and *references* (professors and/or work supervisors who can speak for you) (Landrum, 2005). Unlike a résumé, a CV has no page limit and grows with time and experience. For example, after working in academia since 1999, the author's CV is now over 50 pages long! A CV usually only includes jobs or volunteer experiences that are professionally or academically relevant. For example, working at a summer camp for people with disabilities, a homeless shelter, a crisis hotline, or in a nursery school would probably be relevant and included for psychology majors, but work at a restaurant, retail store, or doing hair would not. Other things would be maybes depending on your program of interest. If you were applying to an animal learning lab, dog-sitting would be relevant, but babysitting would not. If you were applying to a developmental or child behavior program, babysitting would be relevant but dog-sitting would not. CVs should include educational history (college or technical school, but not high school), awards and honors, relevant coursework, psychology-related work experience (whether paid or volunteer), internship and practicum, research projects, research conference presentations and publications, and contact information for references.

Personal statement: These are often the trickiest documents for students. Sounds like they want to know about you as a person, right? Sort of. Your personality, experiences, approaches to work, and research or clinical interests – yes. Your hobbies, pets, and favorite binging TV series – no. Sometimes personal statements will have a specific question to answer, e.g., "Please describe your academic interests and professional goals" or "Please describe a moment or personal experience that has influenced your professional aspirations." Other times, the prompt will just say to provide a personal statement in 1,000 words or less. Keep in mind that for PhD programs, faculty review applications to specifically recruit students for their research labs, so a personal statement for PhD programs should also include some idea of your research interests and who, in particular, you want to work with during your graduate training. Applications for PsyD and master's programs do not tend to have this kind of research specificity, as you are assigned to or coordinate your faculty mentor after program acceptance.

You can schedule a professional documents review in-person or virtually with a professional in your Office of Career and Professional Development. For example, Georgia Southern's Office of Career and Professional Development (OCPD) (n.d.) allows current students and alumni to get some feedback on professional documents before or while on the job market or graduate school circuit. OCPD also utilizes Handshake, where you can email your documents for feedback by a professional and/or get AI-generated feedback. There are benefits for each of these valuable resources, as a person can catch some things that a computer cannot and vice versa, so a savvy student would be sure to check them both out.

"I get the idea, but could I see some examples to model my documents after?" What a wise question! Observing others is an excellent way to learn. The author (Dr. Wickline) has a storehouse of past résumés, cover letters, CVs, and personal statements shared with permission from past students right after they graduated who were successful in landing jobs and getting into graduate school. Please feel free to email me (vwickline@georgiasouthern.edu) if you would like to have access to this resource.

27.5.3 Professional Documents: A Practice Activity

Here is a brief professional documents preparation activity you can try out!

> Activity: Reflecting on Skills Learned
>
> In this lab, you have learned a variety of skills that are transferrable to different careers and jobs. Now is a great time to reflect on those skills so you can communicate them to a potential employer.
> First, list at least three specific skills you have developed in your research design and analysis courses:
> 1.

2.
3.
Next, practice writing for a professional context. Imagine or rewrite each of these statements into a professional tone that you could include in a cover letter or interview.

- I used SPSS to look at the data and figure out what it meant.
- I wrote a research report and fixed it based on feedback.
- I did a lot of data stuff in my R&A II lab class.
- I read through research articles and pulled out the important information for the literature review.
- I practiced citing sources in APA style so my writing was good.
- I gave feedback to my classmates on their writing and helped them fix issues.
- I checked my writing to make sure I wasn't accidentally plagiarizing.

27.6 Interviews

OK, so imagine you got through the ATS screening, got a real person to take a look at your documents, and are now excited to have an interview for your dream position (or any job that will help you pay the bills and make a living). Hooray – good on you! Consider that you should then be ready to describe your transferable skills in any interviews you secure for employment possibilities or graduate school admission. Most places of employment will not hire someone until they have done a phone, virtual, or in-person interview…or maybe all three. The same is true of graduate schools.

As extended examples, imagine with me that a recently graduated college student is asked a question during a job interview about their preparations in college. Notice how the descriptions are more persuasive when specific experience details are mentioned and when skills are visible through those examples, rather than being listed separately.

Architecture: "*Tell me how your college experiences made you ready for an entry-level architecture position.*"

> [Weak answer]: "*I got mostly As and Bs in my architecture courses. We made both cardboard models and computer drafts of blueprint designs. I really like drawing*

and being artistic. I am a good team player and communicator, with good time management skills."

[Persuasive answer]: "*I secured a 3.7 grade point average while working as a paid teaching assistant for the first-year architecture classes and completing a summer internship with Cogdell and Mendrala in Savannah, where I sat in on staff meetings and site visits. I am trained in Computer Aided Drafting (CAD) and Revit for three-dimensional architectural design and structural engineering. In my Buildings & Community course, we partnered with a local homeless shelter to redesign their sleeping area to maximize privacy, comfort, and safety for the residents, where I took the team lead position in designing the rooms for women and children. I really enjoyed presenting our final design to our community partner at the department's end-of-year showcase, where we received second place for the department's annual design competition. If you are interested, I can provide you a flash drive with my design portfolio, letters of recommendation from my internship supervisor and the professors with whom I served as a teaching assistant, and an article from the Atlanta Journal Constitution showing the ribbon-cutting ceremony at the shelter's grand re-opening."*

Nursing: "*Tell me how your college experiences made you ready for this position as an emergency room shift nurse."*

[Weak answer]: "*I got mostly As and Bs in my nursing courses, including anatomy and physiology. The simulation labs and clinical rotations helped me put classroom information into practice. I passed my NCLEX exam on the second try. I am a good team player and communicator, with good time management skills."*

[Persuasive answer]: "*I maintained a 3.7 grade point average in my nursing courses, with strong performance in anatomy and physiology and med-surg nursing. I passed my NCLEX exam on the second try, showing persistence in the face of adversity and a willingness to work hard to achieve my goals. The simulation labs and clinical rotations helped me put classroom information into practice. I especially enjoyed my rotation in maternal and pediatric nursing, which prompted me to secure a summer shadowing opportunity with my local pediatrician's office. I also joined a missions team with my church to provide medical care for orphans and street children in Calcutta, where I assisted with surgical procedures, delivering babies, and a nutrition deficiency outreach program. If you are interested, I can provide you a flash drive with a copy on my paper regarding the benefits of calcium supplements for slowing the effects of scoliosis in street children and letters of recommendation from my pediatric office supervisor, the nursing professor who supervised my clinical rotations, and the director of the Calcutta orphanage."*

Psychology: "*Tell me how your college experiences made you ready for this entry-level position as a behavioral health technician."*

[Weak answer]: "*I got mostly As and Bs in all of my psychology courses, including research design and analysis. I really enjoyed learning about abnormal psychology and personality disorders. I am a good team player and communicator, with good time management skills.*"

[Persuasive answer]: "*I maintained a 3.7 grade point average in my psychology courses while waiting tables to pay the bills and working overnights at the youth shelter, where I learned conflict deescalation skills and ways to get youth with challenging home and life situations to open up and trust me. Although I needed to take a second attempt at my stats course, I passed the second time with a B, showing persistence in the face of adversity and a willingness to work hard to achieve my goals. In my research and analysis II course, we helped to recruit for a study about college adjustment. I looked at potential differences in anxiety and substance use by gender, ethnic group, and sexual orientation. I was the team lead for our group project, where we found that women are more likely to report anxiety than men, LGBTQ students report more anxiety and substance use than heterosexual individuals, and White people report more substance use than other ethnic groups. We were excited to present our project as part of a symposium for the Southeastern Psychological Association(SEPA) conference, and I joined my professor's research lab for the next year to continue expanding my research skills. We are currently working on writing up the project for publication. I also completed the Counseling Center's Question, Persuade, Refer (QPR) training to recognize warning signs of suicide and self-harm and assist people to helping resources. I especially enjoyed my senior internship with Tharros Place, a non-profit that provides advocacy and shelter for human trafficking victims, with a residential facility for teenage girls. At Tharros Place, I assisted with outreach efforts, case management, and clinical intake interviews. If you are interested, I can provide you a flash drive with a copy on my QPR training certificate, SEPA symposium presentation, and letters of recommendation from my internship supervisor, research supervisor, and the professor of my psychological disorders and health psychology courses.*"

Check with your institution's Office of Career and Professional Development to see what valuable resources they may have for you. For example, Georgia Southern's Office of Career and Professional Development (OCPD) video storehouse (n.d.), powered by CandidCareer+, includes videos for how to do various kinds of interviews (e.g., in-person, reverse interview, virtual interviewing, informational interview, call-back interview). You can schedule a mock interview with an OCPD professional to get some practice and feedback, before you are nervous and really wanting the opportunity in front of you. OCPD (n.d.) also affords the opportunity to practice virtual interviews with VMOCK, which will provide instantaneous, AI-generated feedback. Again, different benefits exist for each of these valuable resources. Especially if you are getting close to graduation, please give them a look!

27.7 Summary

Thinking like a psychological scientist is the primary reason why the research design and analysis course sequence is something all psychology majors take. It is that important, like organic chemistry for chem majors, anatomy and physiology for pre-med majors, physics for engineering majors, or drawing for architecture majors. Like you, we want you to learn a whole lot from your psychology major, things that will help inform and improve your personal life, secure your career, and find your passion and purpose. Knowing how to run statistical analyses to test ideas is good. Knowing how to think logically and support claims with data and evidence so you avoid fake news and social media algorithms and make healthy and safe choices for you and those you care about in your everyday life is even better. We hope that your time in research design and analysis has prepared you for whatever life throws at you next.

27.8 Additional Video Helps & Resources

Puglia, M. (n.d.). *Un-hidden curriculum: A podcast providing tools, tutorials, and truth-telling for the next generation of scientists.* YouTube. https://www.youtube.com/channel/UCaDP1cjf4NbZljpW87O1K-Q

27.9 Text Attributions

This text was written by Dr. Virginia Wickline for the current manuscript. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, unless otherwise noted.

27.10 References

American Psychological Association. (2023a). *Inclusive language guide* (2nd ed.). https://www.apa.org/about/apa/equity-diversity-inclusion/language-guidelines

American Psychological Association. (2023b). APA guidelines for the undergraduate psychology major, version 3.0: Empowering people to make a difference in their lives and communities. https://www.apa.org/about/policy/undergraduate-psychology-major

American Psychological Association (2019). Council policy manual. https://www.apa.org/about/policy

American Psychological Association (2017a). *Ethical principles of psychologists and code of conduct.* https://www.apa.org/ethics/code

American Psychological Association (2017b). *Multicultural guidelines: An ecological approach to context, identity, and intersectionality, 2017.* https://www.apa.org/about/policy/multicultural-guidelines

American Psychological Association (2011). *Careers in psychology.* https://www.apa.org/education-career/guide/careers

Bureau of Labor Statistics (2026). *Occupational outlook handbook. Field of degree: Psychology.* U.S. Department of Labor. https://www.bls.gov/ooh/field-of-degree/psychology/psychology-field-of-degree.htm

Center for Workforce Studies (2021). *CWS data tool: Degree pathways in psychology.* American Psychological Association. https://www.apa.org/workforce/data-tools/degrees-pathways

Freis, S. D., & Kraha, A. (2016). You're not in Kansas anymore: How grad school is different from undergrad. *Eye on Psi Chi, 21*(1), 4-5. https://doi.org/10.24839/1092-0803.eye21.1.4

Landrum, R. E. (2005). The curriculum vita: A student's guide to preparation. *Eye on Psi Chi, 9*(2). https://doi.org/10.24839/1092-0803.Eye9.2.28

National Center for Education Statistics (2024, May). Undergraduate degree fields. https://nces.ed.gov/programs/coe/indicator/cta

Naufel, K. Z., Appleby, D. C., Young, J., Van Kirk, J. F., Spencer, S. M., Rudmann, J., Carducci, B., Hettich, P., & Richmond, A. S. (2018). *The skillful psychology student: Prepared for success in the 21st century workplace.* https://www.apa.org/careers/resources/guides/transferable-skills.pdf

Office of Career and Professional Development (n.d.). *Explore career journeys.* https://ocpd.georgiasouthern.edu/videos/

Office of Career and Professional Development (n.d.). *Resumes and CVs.* https://ocpd.georgiasouthern.edu/channels/resumes-cvs/

Office of Career and Professional Development (n.d.). *VMock.* https://ocpd.georgiasouthern.edu/resources/vmock/

Purcell, K. (2026). *The anatomy of an ATS friendly resume format.* Jobscan. https://www.jobscan.co/blog/20-ats-friendly-resume-templates

28 Chapter 28: Appendix

28.1 Critical Values Tables for Hypothesis Testing

If you are conducting hypothesis tests, you will need to use ***critical values tables*** to find the correct critical value or values that go along with your pre-determined probability level (). This information helps you decide whether to reject or retain the null hypothesis. Below is a list of links to critical values tables for different kinds of hypothesis tests mentioned in this text:

- Unit Normal Table (Z-scores)
- [Pearson's] *[r]* (Correlation)
- [Spearman's] [] (Correlation)
- t-tests
- F-tests (ANOVA)
- Chi-square
- Additional critical value tables

28.2 Text Attributions

This chapter was compiled by Dr. Virginia Wickline for the current manuscript. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

28.3 References

Cote, L. R., Gordon, R. G. Randell, C. E., Schmitt, J., & Marvin, H. (2021). *[Introduction to statistics in the psychological sciences]*. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

Crafton Hills College Tutoring Center (2017). Critical values for Spearman's rank-order correlation coefficient. https://www.craftonhills.edu/current-students/tutoring-center/mathematics-tutoring/correlation_coefficient.pdf

Weathington, B. L., Cunningham, J. L., & Pittenger, D. J. (2012). Understanding business research (1st Ed). https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118342978.app2