RESEARCH METHODS FOR THE SOCIAL AND BEHAVIORAL SCIENCES



Michelle Oja
Taft College

LibreTexts™

Research Methods for the Social and Behavioral Sciences

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

This text was compiled on 04/28/2026

TABLE OF CONTENTS

Acknowledgements

Licensing

Licensing

For Faculty

1: Introduction to Research Methods

- 1.1: Prelude to Research Methods
- 1.2: Methods of Knowing
- 1.3: Why are you taking this course?
- 1.4: Summary of Introduction to Research Methods of the Social Sciences

2: Overview of the Scientific Method

- 2.1: Prelude to the Scientific Method
- 2.2: Goals of Science
- 2.3: A Model of Scientific Research
  - 2.3.1: Research Literature
  - 2.3.2: Research Questions
  - 2.3.3: Empirical Study
  - 2.3.4: Data Analysis
  - 2.3.5: Conclusions
  - 2.3.6: Sharing Your Research
- 2.4: Populations and Samples
  - 2.4.1: Sampling
- 2.5: Summary of the Scientific Method

3: Ethics in Research

- 3.1: Prelude to Research Ethics
- 3.2: History of Ethics Codes
- 3.3: APA Ethics Code Standard 8
- 3.4: Principles of Ethical Research
- 3.5: Summary of Research Ethics

4: Measuring Variables

- 4.1: Prelude to Psychological Measurement
- 4.2: Introduction to Measuring Variables
  - 4.2.1: Types of Validity
- 4.3: Introduction to Surveys
  - 4.3.1: Constructing Surveys
  - 4.3.2: Reliability and Validity of Measurement
- 4.4: Interviews and Focus Groups

- 4.5: Measuring Behavior through Biology
- 4.6: Observations
- 4.7: Unobtrusive Research
  - 4.7.1: What Is Left Behind (Physical Trace)
  - 4.7.2: Archival Data
- 4.8: Summary of Measuring Variables

5: Literature Review

- 5.1: Prelude to Reviewing Literature
- 5.2: How to do a Literature Review
- 5.3: How to Read a Research Article
- 5.4: Interpreting Results (p-values and NHST)
- 5.5: Summary of Reviewing the Literature

6: Analyzing

- 6.1: Prelude to Analyzing
- 6.2: Describing Variables
- 6.3: Graphs
  - 6.3.1: Graphing Qualitative Variables- Bar Graphs
  - 6.3.2: Describing the Shape of Graphs of Quantitative Data
    - 6.3.2.1: Graphing Quantitative Data- Line Graphs
    - 6.3.2.2: Graphing Quantitative Data- Histograms
  - 6.3.3: Graphing Quantitative Data- Scatterplots
  - 6.3.4: Honesty in Graphing
- 6.4: Understanding Null Hypothesis Significance Testing
- 6.5: Comparing Means
  - 6.5.1: Effect Sizes
- 6.6: Introduction to Correlations
  - 6.6.1: Correlation versus Causation
  - 6.6.2: Factor Analysis
- 6.7: The Qualitative Approach
  - 6.7.1: Qualitative Analysis- Preparing and Managing Qualitative Data
  - 6.7.2: Qualitative Coding
  - 6.7.3: From Qualitative Data to Findings
- 6.8: Summary of Analyzing

7: Formatting in APA Style

- 7.1: Prelude to Presenting Your Research
- 7.2: Introduction to American Psychological Association (APA) Style
- 7.3: General Paper Format
- 7.4: Formatting by Section
- 7.5: Presenting Results
  - 7.5.1: Tables and Figures
- 7.6: Research Proposals
- 7.7: Summary and APA Style Resources

8: Experimental Research

- 8.1: Prelude to Experimental Research
- 8.2: Initial Equivalence
  - 8.2.1: Natural Groups Research
- 8.3: Between Groups Designs and Matched Group Designs
- 8.4: Repeated Measures Design
- 8.5: Practical Considerations
- 8.6: Summary of Experimental Research

9: Quasi-Experimental Research

- 9.1: Prelude to Quasi-Experimental Research
- 9.2: Quasi-Experimental Designs with One Group
- 9.3: Quasi-Experimental Designs with Two or More Groups
- 9.4: Threats to Internal Validity
- 9.5: Summary of Quasi-Experimental Research

10: Factorial Designs

- 10.1: Prelude to Factorial Designs
- 10.2: Setting Up a Factorial Experiment
- 10.3: Interpreting the Results of a Factorial Experiment
  - 10.3.1: Practice Interpreting Main Effects and Interactions
  - 10.3.2: Interpreting Main Effects and Interactions in Graphs
  - 10.3.3: Interpreting Beyond 2x2 in Graphs
- 10.4: Interpreting Interactions- Do Main Effects Matter?
- 10.5: Summary of Factorial Designs

11: Small-N Designs

- 11.1: Prelude to Single-Subject Research
- 11.2: Introduction to Small-N Research
- 11.3: Small-N Research Designs and Analysis
- 11.4: When should you use Small-N Designs?
- 11.5: Summary of Small-N Design

12: Nonexperimental Research

- 12.1: Prelude to Nonexperimental Research
- 12.2: Introduction to Nonexperimental Research
- 12.3: Longitudinal and Similar Designs
- 12.4: Qualitative Research
- 12.5: Case Studies
- 12.6: Summary of Nonexperimental Research

Index

Glossary

Detailed Licensing

Common Formulas

Detailed Licensing

Acknowledgements

Dr. MO would like to acknowledge the assistance of Taft College and the Open Educational Resources Initiative (OERI) of the Academic Senate for California Community Colleges (ASCCC) for their support in developing this textbook. In particular, Dean Jaime Lopez was critical in procuring the grant money from the California Community College Chancellor's Office's (CCCCO) Zero Textbook Cost (ZTC) program to support this work.

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

For Faculty

Licensing

The licensing for this textbook changes based on the webpage (sub-section). As most of the textbook was modified from Jhangiani et al. (2019) (direct link to Authors page), much of the textbook is CC-BY-NC-SA (a Creative Commons license that allows anyone to use the information for non-commercial purposes and the user must license the modified work in the same way). Some chapters were also modified from Oja (2022), which has a variety of licenses but is mostly CC-BY-SA or CC-BY-NC-SA Most of the other authors' work is licensed as CC-BY-SA (a Creative Commons license allowing use that requires that any modifications of the current work is licensed in the same way).

Reference

For referencing this textbook, use the following:

Oja, M. (2025). *Research Methods for the Social and Behavioral Sciences*. LibreTexts. https://socialsci.libretexts.org/Cou...ioral_Sciences

Ancillaries

There are currently no slide decks, course shells, lecture videos, or quiz banks for this textbook. If you make your own for previous version of this textbooks or this version, and are willing to share with an open license, please let Dr. MO know at the bottom of any page!

Articulation

This book was compiled to fulfill the requirements of the Course Identification descriptor (C-ID) for PSYC 200 and PSYC 205B and for the California Community Colleges.

Additional Textbooks

This textbook is a compilation of Dr. MO's own work with a lot of support from several CC-BY-SA or CC-BY-NC-SA textbooks on LibreTexts website, including:

- Jhangiani, R. S., Chiang, I-C., A., Cuttler, C., & Leighton, D. C. (2019). *Research Methods in Psychology*. Retrieved April 2, 2025, from Research Methods in Psychology (Jhangiani, Chiang, Cuttler, and Leighton)
- Price, P. C., Jhangiani, R. S., Chiang, I-C., A., Cuttler, C., & Leighton, D. C. (2019). *Research Methods in Psychology*. Retrieved April 2, 2025, from https://opentext.wsu.edu/carriecuttler/
- Oja, M. (2022). *Behavioral Statistics, 1*e. Retrieved April 2, 2025 from Behavioral Statistics 1e

Enjoy!

CHAPTER OVERVIEW

1: Introduction to Research Methods

1.1: Prelude to Research Methods

1.2: Methods of Knowing

1.3: Why are you taking this course?

1.4: Summary of Introduction to Research Methods of the Social Sciences

---

1: Introduction to Research Methods is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

1.1: Prelude to Research Methods

Many people believe that women tend to talk more than men—with some even suggesting that this difference has a biological basis. One widely cited estimate is that women speak 20,000 words per day on average and men speak only 7,000. This claim seems plausible, but is it *true*? A group of psychologists led by Matthias Mehl decided to find out. They checked to see if anyone had actually tried to count the daily number of words spoken by women and men. No one had. So these researchers conducted a study in which female and male college students (369 in all) wore audio recorders while they went about their lives. The result? The women spoke an average of 16,215 words per day and the men spoke an average of 15,669—an extremely small difference that could easily be explained by chance. In an article in the journal *Science*, these researchers summed up their findings as follows: "We therefore conclude, on the basis of available empirical evidence, that the widespread and highly publicized stereotype about female talkativeness is unfounded" (Mehl, Vazire, Ramirez-Esparza, Slatcher, & Pennebaker, 2007, p. 82)[1].

1. Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, 317, 82. ↵

1.1: Prelude to Research Methods is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

- **1.1: Prelude to the Science of Psychology** has no license indicated.

1.2: Methods of Knowing

Learning Objectives

1. Describe the methods of acquiring knowledge
2. Understand the benefits and problems with each.

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, or perhaps you know that your friend is lying to you because she is acting strange and won't look you in the eye. But should we trust knowledge from these sources? The methods of acquiring knowledge can be broken down into the following categories each with its own strengths and weaknesses. These are listed in a general order of how trustworthy the knowledge gained may be, but that does not mean that one method is inherently better than another in all circumstances. The method that you rely on may change based on the purpose of the knowledge.

Intuition

The first method of knowing that we will discuss 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.

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. While the strange behavior of your friend may lead you to think s/he is lying to you it may just be that s/he is holding in a bit of gas or is preoccupied with some other issue that is irrelevant to you.

The strength of this approach is that sometimes decisions based on intuition are actually superior to those based on analysis (people interested in this idea should read Malcolm Gladwell's (2005) book Blink). Additionally, weighing alternatives and thinking of all the different possibilities may be paralyzing for some people, but it can definitely be a hassle for everyone!

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, religious authorities, the government, and even professors.

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 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.

Rationalism

Rationalism involves using logic and reasoning to acquire new knowledge. If you"ve taken a philosophy or critical thinking class, then yo'uve encountered rationalism. 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.

Despite this problem, if the premises are correct and logical rules are followed appropriately then this is sound means of acquiring knowledge. [Notice that this statement is an If/Then statement, a common logical rule used in rationalism!]

Empiricism

Empiricism involves acquiring knowledge through observation and experience. Another way to put it is that empiricism is about collecting data, but in a non-systemic way; you are going through your life observing information and experiencing interactions with the world, but you are not systematically collecting data.

Once again many of you may have believed that all swans are white because you have only ever seen white swans. The data that you have collected through your observations is that all swans are white. These examples illustrate the problems with relying on empiricism alone to derive knowledge. We are limited in what we can experience and observe. Moreover, our prior experiences can alter the way we perceive events.

The strength of this approach is that you are actually observing reality. Knowledge through empiricism is not based on a critical analysis or your gut feeling, but on information collected from the environment.

The Scientific Method

Empiricism is at the heart of the scientific method because science relies on observations. But not just any observations, science relies on structured observations which is known as systematic empiricism. 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 don't stop there. Scientists go a step further by using systematic empiricism to make careful observations under various controlled conditions in order to test their ideas and they use rationalism to arrive at valid conclusions.

Like all methods of acquiring knowledge it also has its drawbacks. One major problem is that it is not always feasible to use the scientific method; this method can require considerable time and resources. Another problem with the scientific method is that it cannot be used to answer all questions; the scientific method can only be used to address empirical questions (questions that can be answered by collecting data).

The scientific method is believed to be the most likely of all of the methods to produce verifiable knowledge. This book and your research methods course are designed to provide you with an in-depth examination of how psychologists use the scientific method to advance our understanding of human behavior and the mind.

Indigenous Ways of Knowing

Traditional ways of of learning about the world are practiced by all indigenous cultures. Many researchers from WEIRD (Western, Educated, Industrialized, Rich, Democratic; FMI on WEIRD in Psychology at SimplyPutPsych.co.uk) cultures have never learned about traditional research practices, but they exist, are utilized, and indigenous researchers are innovating new practices. In Simonds and Christopher's (2013) discussion of merging the scientific method with Indigenous ways of knowing, there are several important points from their discussion. First, Indigenous knowledge is both specific to individual cultures, but also to individual locations. The knowledge itself is based on and includes the local environment. Another important consideration is that the community is probably still dealing with the effects of colonization. If researchers focus only on a few current variables, then they are missing critical context for behavior.

Finally, researchers should be have high levels of integrity to collect data and use of the knowledge gained in ethical ways (Simonds & Christopher, 2013). Many Indigenous communities have had their cultural and physical bodies stolen or desecrated by previous researchers and colonizers. Simonds and Christopher (2013, 2189) share that "In Native American communities, knowledge is sacred and access to it must be earned"; current and future researchers must respect the community. Building trust is a critical, but long, process. Researchers should consider community-based participatory research (CBPR) when working with Indigenous communities (Simonds & Christopher, 2013). In one such partnership, it became clear that how data is interpreted differed between the researcher and the Indigenous co-researchers (Simonds and Christopher, 2013). The qualitative researcher was trying to find themes by pulling out quotations, but that method did not make sense to the co-researchers, who suggested that such analysis separated the information from the context of the participants' stories and lacked information about the tribal role of the speaker. A team made up of the academic researcher and the Indigenous co-researcher eventually found a symbol (Crow tipi, a

unique type of design with four base poles that is a symbol for home for that culture) that the worked for all. The team was able to accomplish this through respect, a willingness for both sides to learn and work together, and trust through the CBPR process.

While the next subsection will discuss science more, and the scientific method is the main focus of this entire textbook, it's important to consider what the best method might be to answer your research question. When you are working for indigenous communities, the best research methods may not be the scientific method alone.

References

Gladwell, M. E. (2005). *Blink: The power of thinking without thinking.* (9th ed.). New York: Little, Brown & Co. ↵

Simonds, V. W. & Christopher, S. (2013). Adapting western research methods to Indigenous ways of knowing. *American Journal of Public Health, 103*(12), 2185-2192.

This page titled 1.2: Methods of Knowing is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **1.2: Methods of Knowing** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

1.3: Why are you taking this course?

Learning Objectives

1. Describe what topics this course might cover.
2. List 2-3 reasons why what you learn in this course might be helpful in your future career.

What

You probably signed up for this course because your counselor said that you had to, but didn't really know what it was about. The very simplist way to describe what this class is about is that you'll learn how to understand, evaluate, and interpret other people's research, plus conduct high quality research yourself. To do this, you'll learn about different ways to measure behavior and mental process, to read and interpret research articles to identify if variables *cause* changes in outcomes, and learn about the ethics of conducting research.

But, why is this required for your future career?

Why

Why must you take a class to learn about research when your major is in the social sciences (probably), and you just want to help people? There are several answers to that. The one that is probably most compelling to you right now is:

> *Understanding research will help make you better in your future career.*

Whether you become a therapist, a social worker, a police officer, a nurse, or run a business, you should read research from your field, and understand the results. To make decisions in your career, you'll need to understand and apply the results of research by others in your field. Understanding the research process and being able to interpret research results will help you make decisions based on evidence. What you will learn in this class can help you show that your program (or your decision) is better than the alternatives.

What you might not know right now, but will know when you start in a career field in the social sciences, is that much of what you will be doing is documenting what you did, and how it worked. In other words, you will be writing technical reports to describe your clients based on their different groups. That leads us to a second major reason that you have to take this course:

> *You will need to report evidence to show that what you are doing helps your clients/patients/customers.*

This relates to the first reason because knowing what works and what doesn't will help you figure out how to be more successful in the future. It also turns out that the organizations that fund workers in many helping fields only want to continue funding programs and services that are effectives. It will be much easier to get funding if you understand the best way to organize and present your results. This evidence comes from **collecting**, *analyzing*, and reporting data from your program.

Finally, some of you will fall in love with research, and will become a researcher for a living! This isn't a reason why you have to take the course, but it is a happy accident for those of you who want to help people, but tend to be more logical rather than emotional (like Dr. MO!).

This page titled 1.3: Why are you taking this course? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **1.6: Experimental and Clinical Psychologists** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.
- **1.1: On the Psychology of Statistics** by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.

1.4: Summary of Introduction to Research Methods of the Social Sciences

Key Takeaways

- There are different ways to gather knowledge, and some are better than others.
- You were introduced to this textbook, your course, and research methods in the social sciences.

What's next?

Now that you have a sense of what research methods might cover and why it's important, we'll get more in-depth information on the scientific method.

Exercises

- Practice: Consider three things you know and determine how you acquired that knowledge (authority, intuition, rationalism, empiricism, the scientific method, Indigenous ways of knowing).
- Practice: Try to generate different research questions to describe, predict, and explain a phenomenon that interests you.
- Practice: List three empirical questions about human behavior. List three nonempirical questions about human behavior.
- Practice: For each of the following intuitive beliefs about human behavior, list three reasons that it might be true and three reasons that it might not be true:
  - You cannot truly love another person unless you love yourself.
  - People who receive crisis counseling immediately after experiencing a traumatic event are better able to cope with that trauma in the long term.
  - Studying is most effective when it is always done in the same location.
- Discussion: Consider the following psychological claim. “People’s choice of spouse is strongly influenced by their perception of their own parents. Some choose a spouse who is similar in some way to one of their parents. Others choose a spouse who is different from one of their parents.” Is this claim falsifiable? Why or why not?
- Discussion: People sometimes suggest that psychology cannot be a science because either (a) human behavior cannot be predicted with perfect accuracy or (b) much of its subject matter (e.g., thoughts and feelings) cannot be observed directly. Do you agree or disagree with each of these ideas? Why?
- Watch the following video by PHD Comics for an overview of open access publishing and why it matters:



Open Access Explained! https://youtu.be/L5rVH1KGBCY

- Discussion: Some clinicians argue that what they do is an “art form” based on intuition and personal experience and therefore cannot be evaluated scientifically. Write a paragraph about how satisfied you would be with such a clinician and why from each of three perspectives:
  - a potential client of the clinician
  - a judge who must decide whether to allow the clinician to testify as an expert witness in a child abuse case
  - an insurance company representative who must decide whether to reimburse the clinician for their services

- Practice: Create a short list of questions that a client could ask a clinician to determine whether they pay sufficient attention to scientific research.

References

Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, *317*, 82.

Gladwell, M. E. (2005). *Blink: The power of thinking without thinking* (9th ed.). New York: Little, Brown & Co.

Simonds, V. W. & Christopher, S. (2013). Adapting western research methods to Indigenous ways of knowing. *American Journal of Public Health*, *103*(12), 2185-2192.

This page titled 1.4: Summary of Introduction to Research Methods of the Social Sciences is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **1.7: The Science of Psychology (Summary)** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

2: Overview of the Scientific Method

2.1: Prelude to the Scientific Method
2.2: Goals of Science
2.3: A Model of Scientific Research
2.3.1: Research Literature
2.3.2: Research Questions
2.3.3: Empirical Study
2.3.4: Data Analysis
2.3.5: Conclusions
2.3.6: Sharing Your Research
2.4: Populations and Samples
2.4.1: Sampling
2.5: Summary of the Scientific Method

2: Overview of the Scientific Method is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

2.1: Prelude to the Scientific Method

Here is the abstract of a 2014 article in the journal *Psychological Science.*

> Taking notes on laptops rather than in longhand is increasingly common. Many researchers have suggested that laptop note taking is less effective than longhand note taking for learning. Prior studies have primarily focused on students' capacity for multitasking and distraction when using laptops. The present research suggests that even when laptops are used solely to take notes, they may still be impairing learning because their use results in shallower processing. In three studies, we found that students who took notes on laptops performed worse on conceptual questions than students who took notes longhand. We show that whereas taking more notes can be beneficial, laptop note takers' tendency to transcribe lectures verbatim rather than processing information and reframing it in their own words is detrimental to learning. (Mueler & Oppenheimer, 2014, p. 1159)[1]

In this abstract, the researcher has identified a research question—about the effect of taking notes on a laptop on learning—and identified why it is worthy of investigation—because the practice is ubiquitous and may be harmful to learning.

References

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, *25*(6), 1159-1168. ↵

2.1: Prelude to the Scientific Method is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

- **2.1: Prelude to the Scientific Method** has no license indicated.

2.2: Goals of Science

Learning Objectives

1. Describe the three goals of science and give an example for each.
2. Distinguish between basic research and applied research.

The Broader Purposes of Social Scientific Research

People have always been curious about the natural world, including themselves and their behavior (in fact, this is probably why you are studying psychology in the first place). Science grew out of this natural curiosity and has become the best way to achieve detailed and accurate knowledge. Keep in mind that most of the phenomena and theories that fill psychology textbooks are the products of scientific research. In a typical introductory psychology textbook, for example, one can learn about specific cortical areas for language and perception, principles of classical and operant conditioning, biases in reasoning and judgment, and people's surprising tendency to obey those in positions of authority. And scientific research continues because what we know right now only scratches the surface of what we *can* know.

The Three Goals of Science

The first and most basic goal of science is to describe. This goal is achieved by making careful observations. As an example, perhaps I am interested in better understanding the medical conditions that medical marijuana patients use marijuana to treat. In this case, I could try to access records at several large medical marijuana licensing centers to see which conditions people are getting licensed to use medical marijuana. Or I could survey a large sample of medical marijuana patients and ask them to report which medical conditions they use marijuana to treat or manage. Indeed, research involving surveys of medical marijuana patients has been conducted and has found that the primary symptom medical marijuana patients use marijuana to treat is pain, followed by anxiety and depression (Sexton, Cuttler, Finnell, & Mischley, 2016).

The second goal of science is to predict. Once we have observed with some regularity that two behaviors or events are systematically related to one another we can use that information to predict whether an event or behavior will occur in a certain situation. Once I know that most medical marijuana patients use marijuana to treat pain I can use that information to predict that an individual who uses medical marijuana likely experiences pain, and that those who use medical marijuana more frequently experience more pain. Of course, my predictions will not be 100% accurate, but if the relationship between medical marijuana use and pain is strong then my predictions will have greater than chance accuracy.

The third and ultimate goal of science is to intervene. This goal involves determining the causes of behavior in order to improve outcomes. For example, researchers might try to understand the mechanisms through which marijuana reduces pain. Does marijuana reduce inflammation which in turn reduces pain? Or does marijuana simply reduce the distress associated with pain rather than reducing pain itself? Knowing the answer to why medical marijuana helps with pain may lead to improving pain management with additional treatments. We could compare medical marijuana, anti-inflammatory medication, and distraction (as a psychological coping mechanism to alleviate distress) alone and in combination to see which treatment works best. Sometimes it is enough to be able to predict relationships, but sometimes it is important to know causal relationships (cause and effect) to more effectivement make improvements.

Basic versus Applied Research

Scientific research is used to often be classified as being either basic or applied. Basic research is conducted primarily for the sake of achieving a more detailed and accurate understanding of human behavior, without necessarily trying to address any particular practical problem. The research of Mehl et al. (2007) falls into this category. Applied research is conducted primarily to address some practical problem. Research on the effects of cell phone use on driving, for example, was prompted by safety concerns and has led to the enactment of laws to limit this practice. Although the distinction between basic and applied research is convenient, it is not always clear-cut. For example, basic research on sex differences in talkativeness could eventually have an effect on how marriage therapy is practiced, and applied research on the effect of cell phone use on driving could produce new insights into basic processes of perception, attention, and action.

References

Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, 317, 82

Sexton, M., Cuttler, C., Finnell, J., & Mischley, L (2016). A cross-sectional survey of medical cannabis users: Patterns of use and perceived efficacy. *Cannabis and Cannabinoid Research, 1*, 131-138. doi: 10.1089/can.2016.0007. ↵

This page titled 2.2: Goals of Science is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **1.4: Goals of Science** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

2.3: A Model of Scientific Research

Learning Objectives

1. Review a general model of scientific research in psychology.

Figure **2.3.1** presents a simple model of scientific research in psychology. The researchers formulate a research question, conduct an empirical study designed to answer the question, analyze the resulting data, draw conclusions about the answer to the question, and publishes the results so that they become part of the research literature (i.e., all the published research in that field). Because the research literature is one of the primary sources of new research questions, this process can be thought of as a cycle. New research leads to new questions, which lead to new research, and so on. Figure **2.3.1** also indicates that research questions can originate outside of this cycle either with informal observations or with practical problems that need to be solved. But even in these cases, the researcher would start by checking the research literature to see if the question had already been answered and to refine it based on what previous research had already found.**2.3.1**

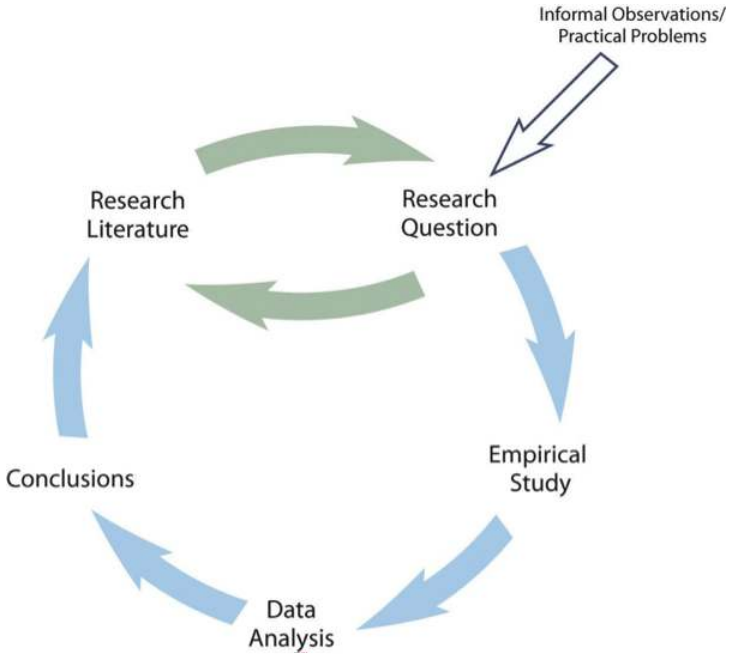
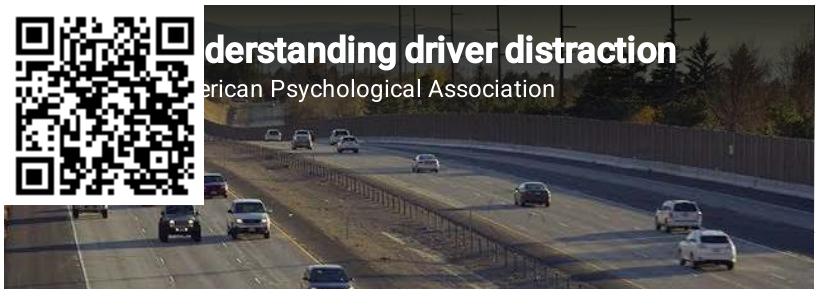
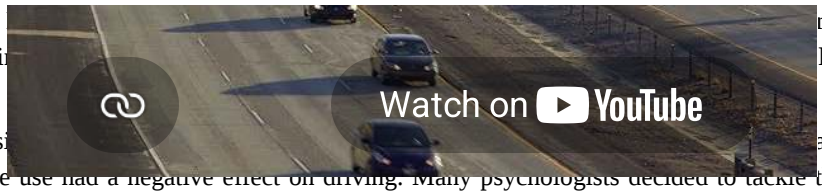


Figure **2.3.1**: A Simple Model of Scientific Research in Psychology

The research by Mehl and his colleagues is described nicely by this model. Their research question—whether women are more talkative than men—was suggested to them both by people's stereotypes and by claims published in the research literature about the relative talkativeness of women and men. When they checked the research literature, however, they found that this question had not been adequately addressed in scientific studies. They then conducted a careful empirical study, analyzed the results (finding very little difference between women and men), formed their conclusions, and published their work so that it became part of the research literature. The publication of their article is not the end of the story, however, because their work suggests many new questions (about the reliability of the result, about potential cultural differences, etc.) that will likely be taken up by them and by other researchers inspired by their work.



Cognitive neuroscientist [redacted] rstand what causes driver distraction. By determi[redacted] licies that keep us safer



As another example, consi[redacted] an to wonder whether, and to what extent, cell phone use had a negative effect on driving. Many psychologists decided to tackle this question scientifically (e.g., Collet, Guillot, & Petit, 2010)[1]. It was clear from previously published research that engaging in a simple verbal task impairs performance on a perceptual or motor task carried out at the same time, but no one had studied the effect specifically of cell phone use on driving. Under carefully controlled conditions, these researchers compared people's driving performance while using a cell phone with their performance while not using a cell phone, both in the lab and on the road. They found that people's ability to detect road hazards, reaction time, and maintain control of the vehicle were all impaired by cell phone use. Each new study was published and became part of the growing research literature on this topic. For instance, other research teams subsequently demonstrated that cell phone conversations carry a greater risk than conversations with a passenger who is aware of driving conditions, which often become a point of conversation (e.g., Drews, Pasupathi, & Strayer, 2004)[2].

References

1. Collet, C., Guillot, A., & Petit, C. (2010). Phoning while driving I: A review of epidemiological, psychological, behavioral and physiological studies. *Ergonomics*, *53*, 589–601. ↵
2. Drews, F. A., Pasupathi, M., & Strayer, D. L. (2004). Passenger and cell-phone conversations in simulated driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*, 2210–2212. ↵

This page titled 2.3: A Model of Scientific Research is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **2.2: A Model of Scientific Research in Psychology** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

2.3.1: Research Literature

Learning Objectives

1. Describe the importance of a literature review is for any research study.
2. Provide examples of sources that are part of the research literature and sources that are not.

You may have many questions about the world and human behavior, but how do you go about testing them? The scientific method model (Figure 2.3.1) doesn't really have a starting point since it's circular, but a good starting point is to take any question you have about human behavior, and then use previous research to help define your terms and make predictions. Finding and reading previous research on your topic is called a literature review.

A later chapter will go more in-depth about conducting a literature review, and will provide suggestions for reading these technical articles. For now, you are just tipping your toe into the waters of research literature to understand its place in the scientific method.

Literature Review Overview

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. 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.

The academic literature in any discipline is enormous—including millions of scholarly articles, essays, and books dating to the beginning of the field—and it continues to grow. Most of what you will read will be from professional journals. There are thousands of professional journals that publish research in social science fields. They are usually published monthly or quarterly in individual issues, each of which contains several articles. The issues are organized into volumes, which usually consist of all the issues for a calendar year. Some journals are published in hard copy only, others in both hard copy and electronic form, and still others in electronic form only. Most articles in professional journals are one of two basic types: empirical research reports (sometimes called primary sources) and review articles (which are a type of secondary sources). Let's start with these empirical studies.

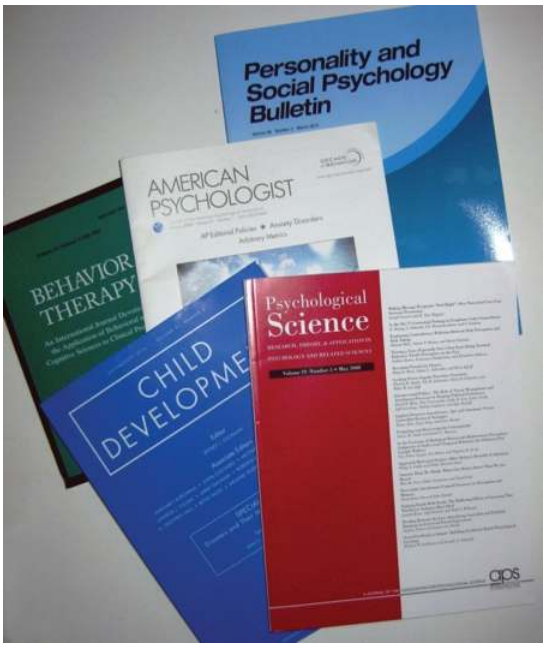


Figure **2.3.1.1**: Small Sample of the Thousands of Professional Journals That Publish Research in Psychology and Related Fields
Covers of psychology journals.

Most professional journals in psychology undergo a process of 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 newer open access journals (e.g., Frontiers in Psychology) 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.

Empirical Research Articles

When you conduct a literature review, you'll focus on articles in professional journals that describe studies in which the authors of article collected and analyzed data. These are sometimes called empirical research reports or primary sources. They are empirical because data is collected, and they are primary because the article authors collected the data. If you are reading an article and it doesn't describe the methods of data collection, especially the number of participants, then you are probably not reading an empirical article.

Other Academic Literature

As described above, empirical research reports describe one or more new empirical studies conducted by the authors. Empirical research reports introduce a research question, explain why it is interesting, review previous research, describe their method and results, and draw their conclusions. But empirical research articles aren't the only academic literature that can help you develop your research study. Review articles and scholarly books provide summaries of research. This may save you time on reading every single research article on the topic, as well as help you narrow your focus and find the right terminology.

What can be confusing is that other types of academic literature, like review articles, are also published in professional journals. Empirical research reports are where you'll learn about individual studies and their results, while 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. Review articles are considered secondary

sources, in comparison to the primary source of empirical research articles. When a review article provides a statistical summary of all of the previous results it is referred to as a meta-analysis. These types of articles are sometimes considered primary sources because the authors choose the articles and (re-)analyze the data, but they could also be considered secondary sources because they did not collect their own data.

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. Although edited volumes can also give a coherent presentation of the topic, it is not unusual for each chapter to take a different perspective or even for the authors of different chapters to openly disagree with each other. In general, scholarly books undergo a peer review process similar to that used by professional journals.

Non-Academic Literature

There are other writings that you may find useful, especially when you are first starting to research a topic. Its boundaries are somewhat fuzzy, but academic literature definitely does not include dictionary, websites, non-academic books, and similar sources that are intended mainly for the general public. These are considered unreliable because they are not reviewed by other researchers and are often not based on empirical research studies. Wikipedia contains much valuable information, but because its authors are anonymous and may not have any formal training or expertise in that subject area, and its content continually changes it is unsuitable as a basis of sound scientific research. These readings can provide background and help you determine a good research question, but they should quickly be abandoned when starting on academic research.

Research Literature to Research Question

Hopefully you have a sense that reading academic research, especially empirical research reports, will provide a background as you begin your own academic research. This background knowledge will help lead you to develop research questions. Review the video from the University Libraries at North Carolina State University to see how you must use previous research to tweak your research question to something that you can find academic research on and still answer your research question (University Libraries, 2013).

References

Universities Libraries. (2013, August 1). *Picking your topic of research!* North Carolina State University. https://www.lib.ncsu.edu/videos/pick...ic-is-research

---

This page titled 2.3.1: Research Literature is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **2.3: Finding a Research Topic** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

2.3.2: Research Questions

Learning Objectives

1. Describe some techniques for turning research ideas into empirical research questions and use those techniques to generate questions.
2. Explain what makes a research question interesting and evaluate research questions in terms of their interestingness.

Research Questions

We started reviewing the scientific method with reviewing research literature, the research process could easily start with research questions. As the model suggestions, informal observations or practical problems can lead to a research question, which would lead to a review of the research literature on those topics.

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. One way to do this is to look closely at the discussion section in a recent research article on the topic. This is the last major section of the article, in which the researchers summarize their results, interpret them in the context of past research, and suggest directions for future research. These suggestions often take the form of specific research questions, which you can then try to answer with additional research. This can be a good strategy because it is likely that the suggested questions have already been identified as interesting and important by experienced researchers.

But you may also want to generate your own research questions. How can you do this? First, if you have a particular behavior or psychological characteristic in mind, you can simply conceptualize it as a variable and ask how frequent or intense it is. How many words on average do people speak per day? How accurate are our memories of traumatic events? What percentage of people have sought professional help for depression? If the question has never been studied scientifically—which is something that you will learn when you conduct your literature review—then it might be interesting and worth pursuing.

If scientific research has already answered the question of how frequent or intense the behavior or characteristic is, then you should consider turning it into a question about a relationship between that behavior or characteristic and some other variable. One way to do this is to ask yourself the following series of more general questions and write down all the answers you can think of.

- What are some possible causes of the behavior or characteristic?
- What are some possible effects of the behavior or characteristic?
- What types of people might exhibit more or less of the behavior or characteristic?
- What types of situations might elicit more or less of the behavior or characteristic?

In general, each answer you write down can be conceptualized as a second variable, suggesting a question about a relationship. If you were interested in talkativeness, for example, it might occur to you that a possible cause of this psychological characteristic is family size. Is there a relationship between family size and talkativeness? Or it might occur to you that people seem to be more talkative in same-sex groups than mixed-sex groups. Is there a difference in the average level of talkativeness of people in same-sex groups and people in mixed-sex groups? This approach should allow you to generate many different empirically testable questions about almost any behavior or psychological characteristic.

If through this process you generate a question that has never been studied scientifically—which again is something that you will learn in your literature review—then it might be interesting and worth pursuing. But what if you find that it has been studied scientifically? Although novice researchers often want to give up and move on to a new question at this point, this is not necessarily a good strategy. For one thing, the fact that the question has been studied scientifically and the research published suggests that it is of interest to the scientific community. For another, the question can almost certainly be refined so that its answer will still contribute something new to the research literature. Again, asking yourself a series of more general questions about the relationship is a good strategy.

- Are there other ways to define and measure the variables?
- Are there types of people for whom the relationship might be stronger or weaker?

- Are there situations in which the relationship might be stronger or weaker—including situations with practical importance?

For example, research has shown that women and men speak about the same number of words per day—but this was when talkativeness was measured in terms of the number of words spoken per day among university students in the United States and Mexico. We can still ask whether other ways of measuring talkativeness—perhaps the number of different people spoken to each day—produce the same result. Or we can ask whether studying elderly people or people from other cultures produces the same result. Again, this approach should help you generate many different research questions about almost any relationship.

Another way to derive research questions is from theories. A theory is a coherent explanation or interpretation of one or more phenomena. Although theories can take a variety of forms, one thing they have in common is that they go beyond the phenomena they explain by including variables, structures, processes, functions, or organizing principles that have not been observed directly. Asking if a particular theory is accurate in specific situations is a research question.

Evaluating Research Questions

Researchers usually generate many more research questions than they ever attempt to answer. This means they must have some way of evaluating the research questions they generate so that they can choose which ones to pursue. In this section, we consider two criteria for evaluating research questions: the interestingness of the question and the feasibility of answering it.

Interestingness

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 absolutely no 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. According to Emerald Publishing Group's guide to the peer review process, an important criteria for your research question should be "Does the article or case study say something original? Does it add to the body of knowledge?" So you could start with a question about how to help community members, and start reading academic literature to understand the issue better. You can use this literature review to determine whether your original research question has already been answered. This means that the question has not already been fully answered by scientific research.

A final factor to consider when deciding whether a research question is interesting is whether its answer has important practical implications. Again, the question of whether taking notes by hand improves learning has important implications for education, including classroom policies concerning technology use. The question of whether cell phone use impairs driving is interesting because it is relevant to the personal safety of everyone who travels by car and to the debate over whether cell phone use should be restricted by law.

Feasibility

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, 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. However, as a newcoming to research methodology, you are encouraged to not drop a research question when you can't immediately think of an easy way to answer the question. It takes years of practice plus creativity to design high-quality research studies!

Looking through a sample of professional journals during your literature review will help you see how other researchers have tested similar questions. Keep in mind, too, that research tends to be carried out by teams of highly trained researchers whose work is often supported in part by government and private grants. Also, keep in mind 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 university 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.

Research Hypotheses

Once you have a general idea for a research question, you can start identifying the variables that you'd like to focus on and how they might related. This becomes your research hypothesis. A research hypothesis is a specific prediction about how variables interact with each other that can be tested. Research hypotheses are developed by considering existing evidence and using reasoning to infer what will happen in the specific context of interest. Hypotheses are usually expressed as statements, but these predictions can be rephrased from research questions. You could start with a broad research question, such as "What study skills will help me pass my class?" or a more specific research question, such as "Will organizing my notes help me pass my class?". To turn this into a research hypothesis, you could answer this question (with a comparison group) by stating: "Students who organize their notes regularly will earn higher grades than students who do not organize their notes."

Good Research Hypotheses

There are three general characteristics of a good hypothesis. First, a good hypothesis must be testable and falsifiable. We must be able to test the hypothesis using the methods of science and it must be possible to gather evidence that will disconfirm the hypothesis if it is indeed false (Popper's falsifiability criterion). Second, a good hypothesis must be logical. As described above, hypotheses are more than just a random guess. Hypotheses should be informed by theories or observations and logical reasoning. Finally, the hypothesis should be positive. That is, the hypothesis should make a positive statement about the existence of a relationship or effect, rather than a statement that a relationship or effect does not exist. As scientists, we don't set out to show that relationships do not exist or that effects do not occur so our hypotheses should not be worded in a way to suggest that an effect or relationship does not exist. The nature of science is to assume that something does not exist and then seek to find evidence to prove this wrong, to show that it really does exist. That may seem backward to you but that is the nature of the scientific method. The underlying reason for this is beyond the scope of this chapter, but it has to do with statistical theory. You may remember that a null hypothesis always says that nothing is happening, and that is what inferential statistical analyses are testing. Then, if our data align, we can reject the null hypothesis that nothing is happening and hopefully support the research hypothesis (that something is happening, and it's happening in the direction that we predicted). What null hypothesis significance testing can't have is to retain the null hypothesis (that nothing is happening) and support the research hypothesis, so all research hypotheses must predict that the variables do interact in some way.

? Exercise 2.3.2.1

What were the three Goals of Science (see section 2.2)?

Answer

The three goals of science listed in section 2.2 were:

1. To describe,
2. To predict, or
3. To intervene.

We can create research hypotheses based on the three goals of science. Starting with the easiest (and least interesting) goal, a research hypothesis that predicts a description would predict the state of the topic. For Imada et al.'s (2024) study of ingroup favoritism, this could be to show that ingroup favoritism exists, or that each of the options to explain ingroup favoritism (reputation management or expectation of cooperation) do happen. For example, "We hypothesize that people treat ingroup members better than outgroup members." The second goal of science, to predict, is often a research hypothesis that compares to groups or two variables to show that they are related. An example related to ingroup favoritism could be: "We hypothesize that higher expectations of cooperation from ingroup members are related to higher ingroup favoritism." The third goal of science, intervention, may have a similar research hypothesis as the second goal, but the research design would differ. To make predictions, we could ask all sorts of people questions about their reputation management, expectations of ingroup cooperation, and ingroup favoritism. But to intervene, we would have to somehow nudge participants to either focus on reputation management or to focus on expectations of ingroup cooperation, then measure which foci was most impactful on measures of ingroup favoritism. Students usually recognize the simplicity of the first goal of science (description), but often get stuck on the second goal (prediction). To actually improve outcomes for your clients, customers, students, or patients, professionals need to focus on the third goal of science (intervention). It is not enough to know that medical marijuana use is related to pain reduction; we want to ensure that the treatment that we prescribe *causes* pain reduction. If medical marijuana works because of its anti-inflammation properties, then pain researchers could focus on the best anti-inflammatory treatments.

This page titled 2.3.2: Research Questions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja via source content that was edited to the style and standards of the LibreTexts platform.

- **2.4: Generating Good Research Questions** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.
- **2.5: Developing a Hypothesis** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

2.3.3: Empirical Study

Learning Objectives

1. Define the concept of a variable, and distinguish between quantitative from categorical variables.
2. Explain the difference between a population and a sample.
3. Distinguish between experimental and non-experimental research.
4. Distinguish between lab studies, field studies, and field experiments.

Identifying and Defining the Variables and Population

Variables and Operational Definitions

Part of generating a hypothesis involves identifying the variables that you want to study and operationally defining those variables so that they can be measured. Research questions in psychology are about variables. A variable is a quantity or quality that varies across people or situations. For example, the height of the students enrolled in a university course is a variable because it varies from student to student. The chosen major of the students is also a variable as long as not everyone in the class has declared the same major. Almost everything in our world varies and as such thinking of examples of constants (things that don't vary) is far more difficult. A rare example of a constant is the speed of light. Variables can be either quantitative or categorical. A quantitative variable is a quantity, such as height, that is typically measured by assigning a number to each individual. Other examples of quantitative variables include people's level of talkativeness, how depressed they are, and the number of siblings they have. A categorical variable is a variable defined by its quality; this is why categorical variables are often called qualitative variables. Categorical variables, such as chosen major, are typically measured by assigning a category label to each individual (e.g., Psychology, English, Nursing, etc.). Other examples include people's nationality, their occupation, and whether they are receiving psychotherapy.

After the researcher generates their hypothesis and selects the variables they want to manipulate and measure, the researcher needs to find ways to actually measure the variables of interest. This requires an operational definition—a description of the variable in terms of precisely how it is to be measured. Many variables that social science researchers are interested in studying cannot be directly observed or measured and this poses a problem because empiricism (observation) is at the heart of the scientific method. Operationally defining a variable involves taking a abstract construct like depression that cannot be directly observed and transforming it into something that can be directly observed and measured. Most variables can be operationally defined in many different ways. For example, depression can be operationally defined as people's scores on a paper-and-pencil depression scale such as the Beck Depression Inventory, the number of depressive symptoms they are experiencing, or whether they have been diagnosed with major depressive disorder. Researchers are wise to choose an operational definition that has been used extensively in the research literature.

Experimental vs. Non-Experimental Research

The next step a researcher must take is to decide which type of approach they will use to collect the data. As you will learn in your research methods course there are many different approaches to research that can be divided in many different ways. One of the most fundamental distinctions is between experimental and non-experimental research.

Experimental Research

Researchers who want to test hypotheses about causal relationships between variables (i.e., their goal is to explain) need to use an experimental method. This is because the experimental method is the only method that allows us to determine causal relationships. Using the experimental approach, researchers first manipulate one or more variables while attempting to control extraneous variables, and then they measure how the manipulated variables affect participants' responses.

Independent and Dependent Variables

The terms independent variable and dependent variable are used in the context of experimental research. The independent variable (IV) what the researcher believes is causing changes in the outcome. Technically, the IV is manipulated by the experimenter manipulates, but we often call the variable that we think is the cause the IV whether or not the experimenter had control over it. For

example, if we thought that social media made people sadder, then social media would be the IV. If the researcher randomly assigned half of the participants to limit their social media usage and half to use social media normally, then social media would be a true IV. However, often researchers will merely ask participants how often they use social media; this is still often called the IV, even though it should be called a predictor variable. The operational definition of the IV determines whether it is a true IV or a predictor variable.

What do we think that the IV causes? The dependent variable is that outcome variable that we're usually trying to improve. The DV is the variable that the experimenter measures, the presumed effect. In our example of social media and sadness, the DV would be sadness. The operational definition could be self-report 1-5 Likert scale (with 1 being very happy and 5 being very sad), how many times the participant cried that week, or an open-ended self-report question asking about how they feel that participants would complete each time they close a social media app.

✔ Example 2.3.3.1

Can blueberries slow down aging? A study indicates that antioxidants found in blueberries may slow down the process of aging. In this study, 19-month-old rats (equivalent to 60-year-old humans) were fed either their standard diet or a diet supplemented by either blueberry, strawberry, or spinach powder. After eight weeks, the rats were given memory and motor skills tests. Although all supplemented rats showed improvement, those supplemented with blueberry powder showed the most notable improvement.

1. What is the independent variable?
2. What are the dependent variables?

Solution

1. IV: dietary supplement: none, blueberry, strawberry, and spinach
2. DVs: memory test and motor skills test

One more together, and then one on your own:

✔ Example 2.3.3.2

Does beta-carotene protect against cancer? Beta-carotene supplements have been thought to protect against cancer. However, a study published in the Journal of the National Cancer Institute suggests this is false. The study was conducted with 39,000 women aged 45 and up. These women were randomly assigned to receive a beta-carotene supplement or a placebo, and their health was studied over their lifetime. Cancer rates for women taking the betacarotene supplement did not differ systematically from the cancer rates of those women taking the placebo.

1. What is the independent variable?
2. What is the dependent variable?

Solution

1. IV: Supplements: beta-carotene or placebo
2. DV: Cancer occurrence

Your turn!

? Exercise 2.3.3.1

How bright is right? An automobile manufacturer wants to know how bright brake lights should be in order to minimize the time required for the driver of a following car to realize that the car in front is stopping and to hit the brakes.

1. What is the independent variable?
2. What is the dependent variable?

Answer

1. IV: Brightness of brake lights DV:
2. Time to hit brakes

If an experiment compares an experimental treatment with a control treatment, then the independent variable (type of treatment) has two levels: experimental and control. If an experiment were comparing five types of diets, then the independent variable (type of diet) would have 5 levels. In general, the number of levels of an independent variable is the number of experimental conditions.

Confounds

Extraneous variables are any variable other than the independent variable or the dependent variable. Confounds are a specific type of extraneous variable that systematically varies along with the independent variable, and therefore provides an alternative explanation for the results. When researchers design an experiment they need to ensure that they control for confounds; they need to ensure that extraneous variables don't become confounding variables because in order to make a causal conclusion they need to make sure alternative explanations for the results have been ruled out.

As an example, if we manipulate the lighting in the room and examine the effects of that manipulation on workers' productivity, then the lighting conditions (bright lights vs. dim lights) would be considered the independent variable and the workers' productivity would be considered the dependent variable. If the bright lights are noisy then that noise would be a confound since the noise would be present whenever the lights are bright and the noise would be absent when the lights are dim. If noise is varying systematically with light then we wouldn't know if a difference in worker productivity across the two lighting conditions is due to noise or light. So confounds are bad, they disrupt our ability to make causal conclusions about the nature of the relationship between variables. However, if there is noise in the room both when the lights are on and when the lights are off then noise is merely an extraneous variable (it is a variable other than the independent or dependent variable) and we don't worry much about extraneous variables. This is because unless a variable varies systematically with the manipulated independent variable it cannot be a competing explanation for the results.

Non-Experimental Research

Researchers who are interested in describing the current situation in detail or describing or predicting relationships between variables can use non-experimental research. Using the non-experimental approach, the researcher measures variables as they naturally occur, but they do not manipulate them. For instance, if I just measured the number of traffic fatalities in America last year that involved the use of a cell phone but I did not actually manipulate cell phone use then this would be categorized as non-experimental research. Alternatively, if I stood at a busy intersection and recorded drivers' genders and whether or not they were using a cell phone when they passed through the intersection to see whether men or women are more likely to use a cell phone when driving, then this would be non-experimental research. Qualitative research is a type of non-experimental research. Qualitative research typically involves formulating broad research questions, collecting large amounts of data from a small number of participants, and summarizing the data using nonstatistical techniques.

It is important to point out that non-experimental does not mean nonscientific. Non-experimental research is still scientific research. It can be used to fulfill two of the three goals of science (to describe and to predict). However, unlike with experimental research, we cannot make *causal* conclusions using this method; we cannot say that one variable *causes* another variable using this method.

Laboratory vs. Field Research

The next major distinction between research methods is between laboratory and field studies. A laboratory study is research done in a controlled environment, like a laboratory, classroom, or other situation in which the experimenter has control. In contrast, a field study is a study that is conducted in the real-world, in a natural environment. In these field experiments, the experimenter usually does not have full control over the situation.

Laboratory experiments typically have high internal validity. In terms of an experiment, internal validity refers to the degree to which we can confidently infer a causal relationship between variables. When we conduct an experimental study in a laboratory environment we have very high internal validity because we manipulate one variable while controlling all other outside extraneous variables. When we manipulate an independent variable and observe an effect on a dependent variable and we control for everything else so that the only difference between our experimental groups or conditions is the one manipulated variable then we

can be quite confident that it is the independent variable that is causing the change in the dependent variable. In contrast, because field studies are conducted in the real-world, the experimenter typically has less control over the environment and potential extraneous variables, and this decreases internal validity, making it less appropriate to arrive at causal conclusions.

But there is typically a trade-off between internal and external validity. External validity refers to the degree to which we can generalize the findings to other circumstances or settings, like the real-world environment. When internal validity is high, external validity tends to be low; and when internal validity is low, external validity tends to be high. So laboratory studies are typically low in external validity, while field studies are typically high in external validity. Since field studies are conducted in the real-world environment it is far more appropriate to generalize the findings to that real-world environment than when the research is conducted in the more artificial sterile laboratory.

Finally, there are field studies which are non-experimental in nature because nothing is manipulated. But there are also field *experiments* where an independent variable is manipulated in a natural setting and extraneous variables are controlled. Depending on their overall quality and the level of control of extraneous variables, such field experiments can have high external and high internal validity.

This page titled 2.3.3: Empirical Study is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **2.6: Designing a Research Study** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

2.3.4: Data Analysis

Learning Objectives

1. Distinguish between descriptive and inferential statistics
2. Identify the different kinds of descriptive statistics researchers use to summarize their data
3. Describe the purpose of inferential statistics.
4. Distinguish between Type I and Type II errors.

Once the study is complete and the observations have been made and recorded the researchers need to analyze the data and draw their conclusions. Typically, data are analyzed using both descriptive and inferential statistics. Descriptive statistics are used to summarize the data and inferential statistics are used to generalize the results from the sample to the population. In turn, inferential statistics are used to make conclusions about whether or not a theory has been supported, refuted, or requires modification.

Descriptive Statistics

Descriptive statistics are used to organize or summarize a set of data. Examples include percentages, measures of central tendency (mean, median, mode), measures of dispersion (range, standard deviation, variance), and correlation coefficients.

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

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

Typically means and standard deviations are computed for experimental research studies in which an independent variable was manipulated to produce two or more groups and a dependent variable was measured quantitatively. The means from each experimental group or condition are calculated separately and are compared to see if they differ.

For non-experimental research, simple percentages may be computed to describe the percentage of people who engaged in some behavior or held some belief. But more commonly non-experimental research involves computing the correlation between two variables. A **correlation coefficient** describes the strength and direction of the relationship between two variables. The values of a correlation coefficient can range from −1.00 (the strongest possible negative relationship) to +1.00 (the strongest possible positive relationship). A value of 0 means there is no relationship between the two variables. Positive correlation coefficients indicate that as the values of one variable increase, so do the values of the other variable. A good example of a positive correlation is the correlation between height and weight, because as height increases weight also tends to increase. Negative correlation coefficients indicate that as the value of one variable increase, the values of the other variable decrease. An example of a negative correlation is the correlation between stressful life events and happiness; because as stress increases, happiness is likely to decrease.

Inferential Statistics

As you learned in the section of this chapter on sampling, typically researchers sample from a population but ultimately they want to be able to generalize their results from the sample to a broader population. Researchers typically want to infer what the population is like based on the sample they studied. **Inferential statistics** are used for that purpose. Inferential statistics allow researchers to draw conclusions about a population based on data from a sample. Inferential statistics are crucial because the effects (i.e., the differences in the means or the correlation coefficient) that researchers find in a study may be due simply to random chance variability or they may be due to a real effect (i.e., they may reflect a real relationship between variables or a real effect of an independent variable on a dependent variable).

Researchers use inferential statistics to determine whether their effects are statistically significant. A **statistically significant** effect is one that is unlikely due to random chance and therefore likely represents a real effect in the population. More specifically results

that have less than a 5% chance of being due to random error are typically considered statistically significant. When an effect is statistically significant it is appropriate to generalize the results from the sample to the population. In contrast, if inferential statistics reveal that there is more than a 5% chance that an effect could be due to chance error alone then the researcher must conclude that their result is not statistically significant.

It is important to keep in mind that statistics are probabilistic in nature. They allow researchers to determine whether the chances are low that their results are due to random error, but they don't provide any absolute certainty. Hopefully, when we conclude that an effect is statistically significant it is a real effect that we would find if we tested the entire population. And hopefully when we conclude that an effect is not statistically significant there really is no effect and if we tested the entire population we would find no effect. And that 5% threshold is set at 5% to ensure that there is a high probability that we make a correct decision and that our determination of statistical significance is an accurate reflection of reality.

But mistakes can always be made. Specifically, two kinds of mistakes can be made. First, researchers can make a **Type I error**, which is a false positive. It is when a researcher concludes that their results are statistically significant (so they say there is an effect in the population) when in reality there is no real effect in the population and the results are just due to chance (they are a fluke). When the threshold is set to 5%, which is the convention, then the researcher has a 5% chance or less of making a Type I error. You might wonder why researchers don't set it even lower to reduce the chances of making a Type I error. The reason is when the chances of making a Type I error are reduced, the chances of making a Type II error are increased. A **Type II error** is a missed opportunity. It is when a researcher concludes that their results are not statistically significant when in reality there is a real effect in the population and they just missed detecting it. Once again, these Type II errors are more likely to occur when the threshold is set too low (e.g., set at 1% instead of 5%) and/or when the sample was too small.

This page titled 2.3.4: Data Analysis is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **2.7: Analyzing the Data** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

2.3.5: Conclusions

Learning Objectives

1. Identify the conclusions researchers can make based on the outcome of their studies.
2. Describe why scientists avoid the term “scientific proof.”

Now that you've read the research literature, modified your research question into a research hypothesis, designed a study, and analyzed your data, you're at the final step of the scientific method! You draw conclusions based on your research study and the available research literature, then share your research with the scientific community to enhance the research literature.

Drawing Conclusions

Since statistics are probabilistic in nature and findings can reflect type I or type II errors, we cannot use the results of a single study to conclude with certainty that a theory is true. Rather theories are supported, refuted, or modified based on the results of research.

If the results are statistically significant and consistent with the hypothesis and the theory that was used to generate the hypothesis, then researchers can conclude that the theory is supported. Not only did the theory make an accurate prediction, but there is now a new phenomenon that the theory accounts for. If a hypothesis is disconfirmed in a systematic empirical study, then the theory has been weakened. It made an inaccurate prediction, and there is now a new phenomenon that it does not account for.

Although this seems straightforward, there are some complications. First, confirming a hypothesis can strengthen a theory but it can never prove a theory. In fact, scientists tend to avoid the word “prove” when talking and writing about theories. One reason for this avoidance is that the result may reflect a type I error. Another reason for this avoidance is that there may be other plausible theories that imply the same hypothesis, which means that confirming the hypothesis strengthens all those theories equally. A third reason is that it is always possible that another test of the hypothesis or a test of a new hypothesis derived from the theory will be disconfirmed. This difficulty is a version of the famous philosophical “problem of induction.” One cannot definitively prove a general principle (e.g., “All swans are white.”) just by observing confirming cases (e.g., white swans)—no matter how many. It is always possible that a disconfirming case (e.g., a black swan) will eventually come along. For these reasons, scientists tend to think of theories—even highly successful ones—as subject to revision based on new and unexpected observations.

A second complication has to do with what it means when a hypothesis is disconfirmed. According to the strictest version of the hypothetico-deductive method, disconfirming a hypothesis disproves the theory it was derived from. In formal logic, the premises “if *A* then *B* ” and “not *B* ” necessarily lead to the conclusion “not *A* .” If *A* is the theory and *B* is the hypothesis (“if *A* then *B* ”), then disconfirming the hypothesis (“not *B* ”) must mean that the theory is incorrect (“not *A* ”). In practice, however, scientists do not give up on their theories so easily. One reason is that one disconfirmed hypothesis could be a missed opportunity (the result of a type II error) or it could be the result of a faulty research design. Perhaps the researcher did not successfully manipulate the independent variable or measure the dependent variable.

A disconfirmed hypothesis could also mean that some unstated but relatively minor assumption of the theory was not met. For example, if Zajonc had failed to find social facilitation in cockroaches, he could have concluded that drive theory is still correct but it applies only to animals with sufficiently complex nervous systems. That is, the evidence from a study can be used to modify a theory. This practice does not mean that researchers are free to ignore disconfirmations of their theories. If they cannot improve their research designs or modify their theories to account for repeated disconfirmations, then they eventually must abandon their theories and replace them with ones that are more successful.

The bottom line here is that because statistics are probabilistic in nature and because all research studies have flaws there is no such thing as scientific proof, there is only scientific evidence.

Reporting the Results

The final step in the research process involves reporting the results. As described in the next section, results are typically reported in peer-reviewed journal articles and at conferences. The most prestigious way to report one’s findings is by writing a manuscript and having it published in a peer-reviewed scientific journal. Manuscripts published in psychology journals typically must adhere to the writing style of the American Psychological Association (APA style) which you will learn later in this textbook.

Another way to report findings is by writing a book chapter that is published in an edited book. Preferably the editor of the book puts the chapter through peer review but this is not always the case and some scientists are invited by editors to write book chapters.

A fun way to disseminate findings is to give a presentation at a conference. This can either be done as an oral presentation or a poster presentation. Oral presentations involve getting up in front of an audience of fellow scientists and giving a talk that might last anywhere from 10 minutes to 1 hour (depending on the conference) and then fielding questions from the audience. Alternatively, poster presentations involve summarizing the study on a large poster that provides a brief overview of the purpose, methods, results, and discussion. The presenter stands by their poster for an hour or two and discusses it with people who pass by. Presenting one's work at a conference is a great way to get feedback from one's peers before attempting to undergo the more rigorous peer-review process involved in publishing a journal article.

This page titled 2.3.5: Conclusions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **2.8: Drawing Conclusions and Reporting the Results** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

2.3.6: Sharing Your Research

Learning Objectives

1. List several ways that researchers can present their research and the situations in which they might use them.
2. Describe how final manuscripts differ from copy manuscripts.
3. Describe the purpose of talks and posters at professional conferences.
4. Recognize the typical format of a research poster.

The previous section noted that the conclusions that are made from analyzing the results of empirical research studies are reported in peer reviewed scientific journals, books, and conference presentations. This section briefly shares a little more detail into some of these topics to bring us back to where we started with the scientific method: research literature.

Types of Manuscripts

Copy Manuscripts

What we will spend a considerable amount of time reviewing will by copy manuscripts. These are manuscripts that will be submitted to a professional journal for publication. Many formatting features of a copy manuscript, such as consistent double-spacing, the running head, and the placement of tables and figures at the end, were intended to make it easier to edit and typeset on its way to publication when journal publication and review were completed on paper.

Review and Theoretical Articles

Recall that review articles summarize research on a particular topic without presenting new empirical results. When these articles present a new theory, they are often called theoretical articles. Review and theoretical articles are structured much like empirical research reports, with a title page, an abstract, references, appendixes, tables, and figures, and they are written in the same high-level and low-level style. Because they do not report the results of new empirical research, however, there is no method or results section. Of course, the body of the manuscript should still have a logical organization and include an opening that identifies the topic and explains its importance, a literature review that organizes previous research (identifying important relationships among concepts or gaps in the literature), and a closing or conclusion that summarizes the main conclusions and suggests directions for further research or discusses theoretical and practical implications. In a theoretical article, of course, much of the body of the manuscript is devoted to presenting the new theory. Theoretical and review articles are usually divided into sections, each with a heading that is appropriate to that section. The sections and headings can vary considerably from article to article (unlike in an empirical research report). But whatever they are, they should help organize the manuscript and make the argument clear.

Final Manuscripts

Until now, we have focused on the formatting of manuscripts that will be submitted to a professional journal for publication. In contrast, other types of manuscripts are prepared by the author in their final form with no intention of submitting them for publication elsewhere. These are called final manuscripts and include dissertations, theses, and other student papers. These manuscripts may look different from strictly APA style manuscripts in ways that make them easier to read, such as putting tables and figures close to where they are discussed so that the reader does not have to flip to the back of the manuscript to see them. If you read a dissertation or thesis, for example, you might notice it does not adhere strictly to APA style formatting. For student papers, it is important to check with the course instructor about formatting specifics. In a research methods course, papers are usually required to be written as though they were manuscripts being submitted for publication.

Conference Presentations

One of the ways that researchers in psychology share their research with each other is by presenting it at professional conferences. (Although some professional conferences in psychology are devoted mainly to issues of clinical practice, we are concerned here with those that focus on research.) Professional conferences can range from small-scale events involving a dozen researchers who get together for an afternoon to large-scale events involving thousands of researchers who meet for several days. Although researchers attending a professional conference are likely to discuss their work with each other informally, there are two more

formal types of presentation: oral presentations (“talks”) and posters. Presenting a talk or poster at a conference usually requires submitting an abstract of the research to the conference organizers in advance and having it accepted for presentation—although the peer review process is typically not as rigorous as it is for manuscripts submitted to a professional journal.

Oral Presentations

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

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

Posters

Another way to present research at a conference is in the form of a research poster. You may have designed a research poster in a middle school science fair, but they have gotten a little more professional and are now almost always printed in a general format. A research poster is typically presented during a one- to two-hour poster session that takes place in a large room at the conference site. Presenters set up their posters on bulletin boards arranged around the room and stand near them. Other researchers then circulate through the room, read the posters, and talk to the presenters. In essence, poster sessions are a grown-up version of the school science fair. But there is nothing childish about them. Posters are used by professional researchers in all scientific disciplines and they are becoming increasingly common. At a recent American Psychological Association Conference, nearly 2,000 posters were presented across 16 separate poster sessions. Among the reasons posters are so popular is that they encourage meaningful interaction among researchers.

Posters are typically a large size, maybe four feet wide and three feet high. The poster’s information is organized into distinct sections, including a title, author names and affiliations, an introduction, a method section, a results section, a discussion or conclusions section, references, and acknowledgments. Although posters can include an abstract, this may not be necessary because the poster itself is already a brief summary of the research. Figure **2.3.6.1** shows two different ways that the information on a poster might be organized.

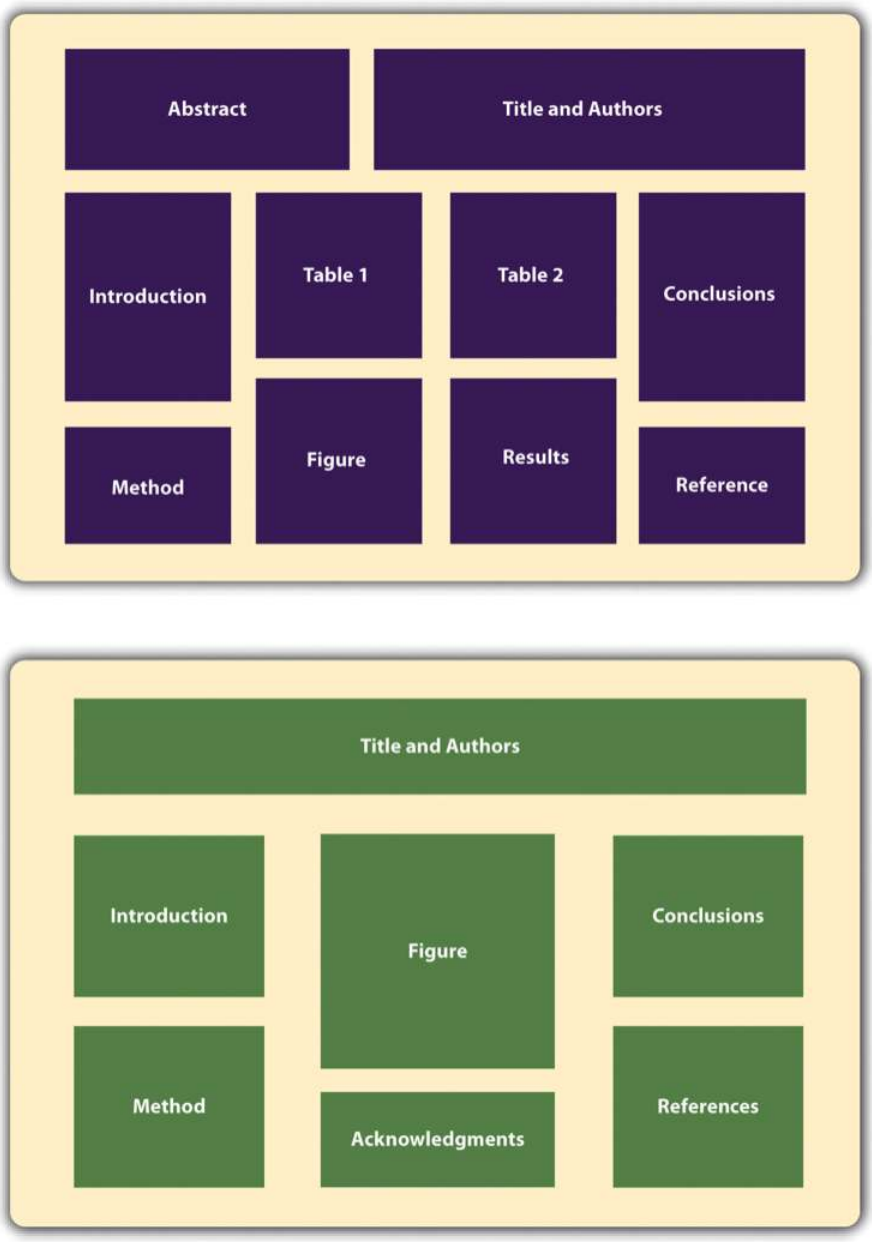


Figure **2.3.6.1**: Two Possible Ways to Organize the Information on a Poster

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

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

visitors who want more information, to offer to send them a detailed write-up, or to provide contact information so that they can request more information later.

For more information on preparing and presenting both talks and posters, see the website of the Undergraduate Advising and Research Office at Dartmouth College: http://www.dartmouth.edu/~ugar/undergrad/posterinstructions.html

Professional Conferences for Psychological Reseach

Following are links to the websites for several large national conferences in North America and also for several conferences that feature the work of undergraduate students. For a comprehensive list of psychology conferences worldwide, see the following website: http://www.conferencealerts.com/psychology.htm

Large Conferences

- Canadian Psychological Association Convention: http://www.cpa.ca/convention
- American Psychological Association Convention: http://www.apa.org/convention
- Association for Psychological Science Conference: http://www.psychologicalscience.org/index.php/convention
- Canadian Society for Brain, Behavior, and Cognitive Science Annual Meeting: https://www.csbbcs.org/meetings
- Society for Personality and Social Psychology Conference: http://meeting.spsp.org/
- Psychonomic Society Annual Meeting: http://www.psychonomic.org/annual-meeting

U.S. Regional conferences where undergraduate researchers frequently present

- Eastern Psychological Association (EPA): http://www.easternpsychological.org
- Midwestern Psychological Association (MPA): http://www.midwesternpsych.org/
- New England Psychological Association (NEPA): http://www.newenglandpsychological.org/
- Rocky Mountain Psychological Association (RMPA): http://www.rockymountainpsych.com/
- Southeastern Psychological Association (SEPA): http://www.sepaonline.com/
- Southwestern Psychological Association (SWPA): http://www.swpsych.org/
- Western Psychological Association (WPA): http://westernpsych.org/

Canadian Undergraduate Conferences

- Connecting Minds Undergraduate Research Conference: http://www.connectingminds.ca
- Science Atlantic Psychology Conference: https://scienceatlantic.ca/conferences/

To finish out this chapter overviewing the scientific method, we will learn about samples and populations.

This page titled 2.3.6: Sharing Your Research is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **11.4: Other Presentation Formats** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.
- **11.2: American Psychological Association (APA) Style** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

2.4: Populations and Samples

Because social scientists want to make people's lives better, they see what works on a small group of people, and then apply it to everyone. The small group of people is the sample, and "everyone" is the population.

Sample and Population

A sample is the small group of people that scientists test stuff on. We want at least 30 people in each group, so a study about two different ways to take notes should have at least 60 people in the sample. A population is the "everyone" that we want to apply the results to. Sometimes, "everyone" can be a pretty small group; if Dr. MO measured the GPA of one of her Research Methods classes, then the sample would be the class and the population could be students in all Research Methods classes at her he college. If you remember the vocabulary that you learned in the section about the Empirical Study step of the scientific method (in section 2.3.3), GPA would be the DV and the different types of note-taking would be the IV.

A sample is a concrete thing. You can open up a data file, and there's the data from your sample. The sample will often include the exact number of participants that the research collected data from. This is often identified as N, so a study with 60 participants would be identify the number of participants by stating N=60. A population, on the other hand, is a more abstract idea. It refers to the set of all possible people, or all possible observations, that you want to draw conclusions about, and is generally *much* bigger than the sample. In an ideal world, the researcher would begin the study with a clear idea of what the population of interest is, since the process of designing a study and testing hypotheses about the data that it produces does depend on the population about which you want to make statements. However, that doesn't always happen in practice: usually the researcher has a fairly vague idea of what the population is and designs the study as best he/she can on that basis.

Examples

In the example about note-taking from this chapter's prelude, the sample would be the class from which we got the data from, and the population would be the biggest group that they could represent. There's often more than one possible population, but I might say all college students could be a good population for this sample. Try out Exercise **2.4.1** to see if you can identify the sample and the population:

? Exercise 2.4.1

Let's say I want to know if there's a relationship between how much time students study for their course and their current GPA. Imagine that I asked 100 of my Introduction to Psychology students this semester to complete a course workload estimator (such as the Course Workload Estimator provided by Rice University) and get their current GPA from my college's institutional research office. For this scenario:

1. Who is the Sample?
2. Who could be the population? In other words, what is the biggest group that this sample could represent?

Answer

1. Who is the Sample? 100 Introduction to Psychology students
2. Who could be the population? In other words, what is the biggest group that this sample could represent? There are many possible populations, but all Introduction to Psychology students could work, or all Introduction to Psychology students at my college might make sense, too.

Sometimes it's easy to state the population of interest. In a typical a experiment, determining the population of interest is a bit more complicated. Suppose Dr. Navarro ran an experiment using 100 undergraduate students as participants. Her goal, as a cognitive scientist, is to try to learn something about how the mind works. So, which of the following would count as "the population":

- All of the undergraduate psychology students her university in Australia?
- Undergraduate psychology students in general, anywhere in the world?
- Australians currently living?
- Australians of similar ages to my sample?

- Anyone currently alive?
- Any human being, past, present or future?
- Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment?
- Any intelligent being?

Each of these defines a real group of mind-possessing entities, all of which might be of interest to me as a cognitive scientist, and it's not at all clear which one ought to be the true population of interest. Maybe surprisingly for you, there's no "right" answer! Although some of the suggestions get a little vague, they all could potentially be a population that Dr. Navarro's sample represents. Irrespective of how the population is defined, the critical point is that the sample is a subset of the population. The goal of researchers is to use our knowledge of the sample to draw inferences about the properties of the population.

? Exercise 2.4.1

Actual drug use is much higher than drug arrests suggest, so you might want to measure how many people use marijuana. If you send out a survey asking about their drug use to everyone with a driver's license in California, but only 30% fill it out:

1. Who is the Sample?
2. Who could be the population? In other words, what is the biggest group that this sample could represent?

Answer

Add texts here. Do not delete this text first.

1. Who is the Sample? 30% of Californians with driver's licenses
2. Who could be the population? In other words, what is the biggest group that this sample could represent? There are many possible populations, but Californians who have driver's licenses might make the most sense here.

This last example shows that sometimes our sample limits who we can generalize our results about, who could be our population.

In almost every situation of interest, what we have available to us as researchers is a sample of data. The data set available to us is finite, and incomplete. We can't possibly get every person in the world to do our experiment. The next section discusses different ways to identify who from the population could be in our study as the sample.

This page titled 2.4: Populations and Samples is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **10.1: Samples, Populations and Sampling** by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.
- **1.5: Populations and Samples** by Michelle Oja is licensed CC BY-SA 4.0.

2.4.1: Sampling

Learning Objectives

1. Explain the difference between major types of sampling.
2. List some techniques that can be used to increase the response rate and reduce non-response bias, and why this is important.

Now that you have a sense of a how a sample is the subset of the population of interest whom researchers actually collect data from, let's discuss different methods for recruiting those participants; this is sampl*ing*.

Sampling

Essentially all research involves sampling—selecting a sample to study from the population of interest. Sampling falls into two broad categories: Probability Sampling or Non-Probability Sampling. The first category, probability sampling, occurs when the researcher can specify the probability that each member of the population will be selected for the sample. The second category, non-probability sampling, occurs when the researcher *cannot* specify probability that each member of the population will be selected for the sample.

Probability Sampling

When conducting quantitative research are large samples, probability sampling is the preferred method because the goal of most 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 survey 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.

Probability sampling requires a very clear specification of the population; the population of interest depends on the research questions to be answered. The population might be all registered voters in California, all American consumers who have used a ride share app in the past year, people in Seattle over 40 years old who have received a mammogram in the past decade, or all the alumni of a particular college. Once the population has been specified, probability sampling requires a sampling frame. A sampling frame is essentially a list of all the members of the population from which to select the respondents. Sampling frames can come from a variety of sources, including enrollment records, lists of registered voters, or hospital records. In some cases, a map can serve as a sampling frame, allowing for the selection of cities, streets, or households.

There are a variety of different probability sampling methods. Simple random sampling, often just called random sampling, is done in such a way that each individual in the population has an equal probability of being selected for the sample. This type of sampling could involve putting the names of all individuals in the sampling frame into a hat, mixing them up, and then drawing out the number needed for the sample. Given that most sampling frames take the form of computer files, random sampling is more likely to involve computerized sorting or selection of respondents. A common approach in telephone surveys is random-digit dialing, in which a computer randomly generates phone numbers from among the possible phone numbers within a given geographic area. Random sampling is sometimes also called random selection since researchers are selecting participants out of the population (through a sampling frame) randomly.

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.” Slightly more complicated, proportionate stratified random sampling can be used to select a sample in which the proportion of respondents in each of various subgroups matches the proportion in the population. For example, because about 12.6 % of the American population is African American , stratified random sampling can be used to ensure that a survey of 1,000 American adults includes about 126 African-American respondents. Disproportionate stratified random sampling can also be used to sample extra respondents from particularly small subgroups—allowing valid conclusions to be drawn about those subgroups. For example, because Asian Americans make up a relatively small percentage of the American population (about 5.6%), a simple random sample of 1,000 American adults might include too few Asian Americans to draw any conclusions about them as distinct from any other subgroup. If representation is important to the research question, however, then disproportionate stratified random sampling could be used to ensure that enough Asian-American respondents are included in the sample to draw valid conclusions about Asian American s a whole.

Non-Probability Sampling

Despite the benefit of probability samples to create samples that better represent the population of interest, most social science research involves non-probability sampling. For example, convenience sampling (studying individuals who happen to be nearby and willing to participate) is a very common form of non-probability sampling used in social science research. Other forms of non-probability sampling include snowball sampling (existing research participants help recruit additional participants for the study) and self-selection sampling (in which individuals choose to take part in the research on their own accord, without being approached by the researcher directly). Snowball sampling is often used with populations that are small or rare, such as transgender People of Color in rural areas, as current participants can connect researchers to other group members. An example of self-selection sampling could be when a friend posts their results of an online survey, and then you decide to also take the survey.

Sample Size

When you are designing a study and choosing how to recruit your participants, it is necessary to decide how many participants you want to collect data from; how large does a sample need to be? In general, this estimate 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 a practical constraint in the form of 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.

Sample Size and Population Size

Why is a sample of about 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 American adults is generally considered a good sample of the roughly 252 million adults in the American population—even though it includes only about 0.000004% 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 reduction 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.

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. For example, if you sample from all Californians with driver's licenses, you are missing teens younger than 16, many teens in cities who choose not to get their license as soon as legally possible, and drivers who have lost their license due to traffic violations.

There is one form of sampling bias that even careful random sampling is subject to. It is almost never the case that everyone selected for the sample actually responds to the survey. Some may have died or moved away, and others may decline to participate because they are too busy, are not interested in the survey topic, or do not participate in surveys on principle. If these survey non-responders differ from survey responders in systematic ways, then this difference can produce non-response bias. For example, in a mail survey on alcohol consumption, researcher Vivienne Lahaut and colleagues found that only about half the sample responded after the initial contact and two follow-up reminders (Lahaut 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, Lahaut et al. (2002) later made unannounced visits to the homes of a subset of the non-responders; they came back up to five times if they did not find the participants at home! Lahaut et al. (2002) 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 was too high.

Although there are methods for statistically correcting for non-response bias, they are based on assumptions about the non-responders; for example, that they are more similar to late responders than to early responders—which may not be correct. For this reason, the best approach to minimizing non-response bias is to minimize the number of non-responders—that is, to maximize the response rate. There is a large research literature on the factors that affect survey response rates (Groves et al., 2004). In general, in-person interviews have the highest response rates, followed by telephone surveys, and then mail and Internet surveys. Among the other factors that increase response rates are sending potential respondents a short pre-notification message informing them that they will be asked to participate in a survey in the near future and sending simple follow-up reminders to non-responders after a few weeks. The perceived length and complexity of the study can also make a difference, which is why it is important to keep survey questionnaires as short, simple, and on topic as possible. Finally, offering an incentive—especially cash—is a reliable way to increase response rates. However, ethically, there are limits to offering incentives that may be so large as to be considered coercive.

You've now gotten an overview of the scientific method. Let's summarize, and then learn about ethics in research.

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.

This page titled 2.4.1: Sampling is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **7.4: Conducting Surveys** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

2.5: Summary of the Scientific Method

Key Takeaways

- Research can be described by a simple cyclical model. A research question based on the research literature leads to an empirical study, the results of which are published and become part of the research literature.
- The research literature is all the published research, consisting primarily of articles in professional journals and scholarly books.
- Early in the research process, it is important to conduct a review of the research literature on your topic to refine your research question, identify appropriate research methods, place your question in the context of other research, and prepare to write an effective research report.
- Research questions and reading previous research will lead to a complete research hypothesis.
- It is important to evaluate how interesting a research question is before designing a study and collecting data to answer it. Factors that affect interestingness are the extent to which the answer is in doubt, whether it fills a gap in the research literature, and whether it has important practical implications.
- A sample is a small subset of a larger population that is selected to participate in the research study. There are many different ways of sampling participants including convenience sampling and simple random sampling.
- The final step of the research process involves reporting results at scientific conferences, in journal articles, and/or in books.

What's next?

Now that you have a deeper understanding of the scientific method, you will learn when and how to ethically collect data from humans (and other animals).

Exercises

- Practice: Find a description of an empirical study in a professional journal or in one of the scientific psychology blogs. Then write a brief description of the research in terms of the cyclical model presented here. One or two sentences for each part of the cycle should suffice.
- Watch the following TED Ed video, in which David H. Schwartz provides an introduction to two types of empirical studies along with some methods that scientists use to increase the reliability of their results:



A YouTube element has been excluded from this version of the text. You can view it online here: https://kpu.pressbooks.pub/psychmethods4e/?p=271

- Practice: Generate three research ideas based on each of the following: informal observations, practical problems, and topics discussed in recent issues of professional journals.
- Practice: Generate an empirical research question about each of the following behaviors or psychological characteristics: long-distance running, getting tattooed, social anxiety, bullying, and memory for early childhood events.
- Practice: Find a recent empirical research report in a professional journal. Read the introduction and highlight in different colors descriptions of theories and hypotheses.

- Practice: Using the research article you found in a professional journal identify whether the study was experimental or non-experimental. If it was experimental identify the independent and dependent variables.
- Practice: Using the research article you found in a professional journal identify which descriptive statistics were reported.
- Practice: Describe why theories can be supported but not proved.

This page titled 2.5: Summary of the Scientific Method is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **2.9: Overview of the Scientific Method (Summary)** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

3: Ethics in Research

3.1: Prelude to Research Ethics
3.2: History of Ethics Codes
3.3: APA Ethics Code Standard 8
3.4: Principles of Ethical Research
3.5: Summary of Research Ethics

3: Ethics in Research is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

3.1: Prelude to Research Ethics

In 1998 a medical journal called *The Lancet* published an article of interest to many psychologists. The researchers claimed to have shown a statistical relationship between receiving the combined measles, mumps, and rubella (MMR) vaccine and the development of autism—suggesting furthermore that the vaccine might even cause autism. One result of this report was that many parents decided not to have their children vaccinated, which of course put them at higher risk for measles, mumps, and rubella. However, follow-up studies by other researchers consistently failed to find a statistical relationship between the MMR vaccine and autism—and it is widely accepted now in the scientific community that there is no relationship. In addition, several more serious problems with the original research were uncovered. Among them were that the lead researcher stood to gain financially from his conclusions because he had patented a competing measles vaccine. He had also used biased methods to select and test his research participants and had used unapproved and medically unnecessary procedures on them. In 2010 *The Lancet* retracted the article, and the lead researcher's right to practice medicine was revoked (Burns, 2010). [1]



[1] Burns, J. F. (2010, May 24). British medical council bars doctor who linked vaccine to autism. *The New York Times*. Retrieved fromhttp://www.nytimes.com/2010/05/25/health/policy/25autism.html?ref=andrew_wakefield

3.1: Prelude to Research Ethics is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **3.1: Prelude to Research Ethics** has no license indicated.

3.2: History of Ethics Codes

Learning Objectives

1. Describe the history of ethics codes for scientific research with human participants.
2. Summarize the ethical requirements that resulted from historical atrocities and the resulting ethics codes.
3. Explain the three ethical principles described in the Belmont Report.

To understand why researchers need ethics code, lets review some of the foundational documents and their antecedent events in this brief historical overview of ethics codes.

History of Ethics Codes

Nuremburg Codes (1947)

One of the earliest ethics codes was the Nuremberg Code. The Nuremburg Code a set of 10 principles written in 1947 in conjunction with the trials of Nazi physicians accused of shockingly cruel research on concentration camp prisoners during World War II. It provided a standard against which to compare the behavior of the men on trial—many of whom were eventually convicted and either imprisoned or sentenced to death. The Nuremberg Code was particularly clear about the importance of carefully weighing risks against benefits and the need for informed consent. This cost/benefit analysis tries to see if the cost to the research participants is worth the benefit to science. More recent thinking encourages researchers to also consider the cost to the community, not just an individual. This comparison may change over time; something that used to seem too costly might not seem costly anymore, or something that used to seem easy might seem too costly now.

If the cost to the participants or their community seems higher than the importance of the knowledge hat would be gained, it easier to justify the experiment if the costs to the participants and their community can be decreased. How could a researcher decrease the cost to the participant? One way is to provide resources. For studies that may result in psychological harm, this could include counseling. For students that may cause physical harm, a resource that could be provided might be medical care. But the best way to reduce harm to the participants and their community it to find alternative testing designs that would be less harmful.

Ethical Principles of Psychologists and Code of Conduct (1953)

The American Psychological Association's (APA) Ethical Principles of Psychologists and Code of Conduct (also known as the APA Ethics Code) was first published in 1953 and has been revised several times since then, most recently in 2016 (APA, 2021). It includes about 150 specific ethical standards that psychologists and their students are expected to follow. Much of the APA Ethics Code concerns the clinical practice of psychology—advertising one’s services, setting and collecting fees, having personal relationships with clients, and so on. For our purposes, the most relevant part is Standard 8: Research and Publication. The next section will consider some of Section 8's most important aspects in more detail.

Declaration of Helsinki (1964)

The Declaration of Helsinki is a similar ethics code that was created by the World Medical Council in 1964. Among the standards that it added to the Nuremberg Code was that research with human participants should be based on a written protocol—a detailed description of the research—that is reviewed by an independent committee. This idea of an independent committee is the start of the requirement for an institutional review board (IRB). The IRB is a committee that reviews proposed research to ensure ethical treatment of participants; this will be discussed in more detail when APA's Section 8 (APA, 2021) is reviewed. The Declaration of Helsinki has been revised several times, most recently in 2024. According to the World Medical Association, the Declaration of Taipei on Ethical Considerations regarding Health Databases and Biobanks (World Medical Association, 2025) adds additional protections needed for modern research. This is an area in which behavioral research should enhance protections.

Health and Human Services Trainings

Do you want to get ahead of other students at the end of their bachelor's degree? You should join a research lab! To prepare for that, save your certificate of completion of one or more of the following trainings from the Department of Health and Human

Services's (HHS) Human Research Protection Foundational Trainings:

- Lesson 1: When HHS Regulations Applies Apply
- Lesson 2: What is human subject research?
- Lesson 3: IRBs
- Lesson 4: IRB Review of Research
- Lesson 5: Institutional Overview of Human Research

And if your school doesn't currently have an active research lab, consider working with a professor start a student club or course so that you can get research experience.

Belmont Report (1978)

In the United States, concerns about the Tuskegee study and others led to the publication in 1978 of a set of federal guidelines called the Belmont Report. A particularly tragic example is the Tuskegee syphilis study conducted by the US Public Health Service from 1932 to 1972 (Reverby, 2009). The participants in this study were poor African American men in the vicinity of Tuskegee, Alabama, who were told that they were being treated for “bad blood.” Although they were given some free medical care, they were not treated for their syphilis. Instead, they were observed to see how the disease developed in untreated patients. Even after the use of penicillin became the standard treatment for syphilis during the course of this 40-year study (in the 1940's), these men were denied treatment. They were not informed of the full methodology of the study, and local hospitals were told not to treat the men. The study was finally discontinued after details were made known to the general public by journalists and activists.

Crash Course's Black American History #29: The Tuskegee Experiment

For more information on the Tuskegee Experiment, the Crash Course series on Black American History devoted an episode (#29) to this horrific study (Smith, 2021).

The Belmont Report explicitly recognized three important ethnical principles for researchers on humans:

- Seeking Justice
- Respect for Persons
- Beneficence

The principle of seeking justice includes the importance of conducting research in a way that distributes risks and benefits fairly across different groups at the societal level. The importance of respect for persons acknowledges individuals’ autonomy and protection for those with diminished autonomy (e.g., prisoners, children), and translates to the need for informed consent. Informed consent will be discussed in more detail in the section on APA's Section 8 (APA, 2021). Finally, the principle of beneficence underscores the importance of maximizing the benefits of research while minimizing harms to participants and society.

Before Belmont

The Smithsonian Learning Lab has created an interactive website to learn about ethics in psychological research before and through the Belmont Report. Check out Before Belmont: Ethics in Experimental Psychology from the Cummings Center for the History of Psychology (2020).

Federal Policy for the Protection of Human Subjects (1991)

The Belmont Report became the basis of a set of laws:, the Federal Policy for the Protection of Human Subjects, that apply to research conducted, supported, or regulated by the federal government. Because all public universities in the United States are supported by the federal government and many research grants are from federal agencies, these policies apply to almost all research conducted in the United States. An extremely important part of these regulations is that universities, hospitals, and other institutions that receive support from the federal government must establish an institutional review board (IRB), the committee that is responsible for reviewing research protocols for potential ethical problems. More information on IRB's will be provided in the section on APA's Section 8 (APA, 2021).

Ethics Codes

Ethics Codes

The U.S. Department of Health and Human Services's Ethical Codes and Research Standards page provides links to all of these ethics codes, and more! They are all highly recommended and, with the exception of the Federal Policy, short and easy to read.

As we move from the historical justifications and early ethics codes to understanding one profession's ethical codes related to research, hopefully you can see how detailed and enforceable ethics codes provide guidance on important issues, and that these codes must be continually updated as society and technology changes.

Crash Course's Statistics Episode #12: Ethical Data Collection

For some additional examples and discussion of unethical research, the Crash Course series on Statistics discussed ethical data collection in an episode (#12) (Hill, 2018).

References

American Psychological Association. (2021, January 1). *Ethnical principles of psychologists and code of conduct*. https://www.apa.org/ethics/code/index

Health and Human Services (2025, May 7). *Human research protection foundational training*. U.S. Department of Health and Human Resources. https://www.hhs.gov/ohrp/education-a...ing/index.html

Hill, A. (2018, April 18). *Henrietta Lacks, the Tuskegee experiment, and ethical data collection: Crash Course Statistics #12*. Crash Course. https://thecrashcourse.com/courses/h...statistics-12/

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

Smith, C. (2021, December 21). *The Tuskegee experiment: Crash Course Black American History #29*. Crash Course. https://thecrashcourse.com/courses/t...an-history-29/

Cummings Center For the History of Psychology. (2020, October 22). *Before Belmont: Ethics in experimental psychology*. Smithsonian Learning Lab. https://learninglab.si.edu/collectio...ZNGQriWxb5nXDs

United States Department of Health and Human Services (2024, April 29). *Ethnical codes and research standards*. https://www.hhs.gov/ohrp/internation...rds/index.html

World Medical Association (2025). *Declaration of Helsinki*. https://www.wma.net/what-we-do/medic...n-of-helsinki/

This page titled 3.2: History of Ethics Codes is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **3.3: From Moral Principles to Ethics Codes** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

3.3: APA Ethics Code Standard 8

Learning Objectives

1. Summarize the Section 8 of the American Psychological Association Ethics Code
2. Explain the importance of ethical requirements such as informed consent, debriefing, and scholarly integrity.
3. Describe some pro's and con's of deception, research with nonhuman animals, or other controversial research practices.

APA Ethics Code

The American Psychological Association's (APA) Ethical Principles of Psychologists and Code of Conduct (APA, 2021) can be found at https://www.apa.org/ethics/code. The code was originally created in 1953, then updated 11 times. We will focus on Section 8: Research and Publication, but your profession will have a separate ethics code for people are not Psychology majors.

Example 3.3.1

If you are not majoring in psychology, you are encouraged to search for the ethics code for your future profession. Once you find it, check to see if it has ethical guidelines for research.

Solution

For example, the American Sociological Association several ethics codes (ASA, 2023), which does has several sections that are relevant to conducting and publishing research. The National Association of Social Workers also has a code of ethics (NASW, 2021). While each Ethical Principle could be applied to research, section 5 on professional responsibilities includes a specific section on evaluation and research.

In the box below is Section 8 of the APA Ethics Code in its entirety. Following this box, further detailed is provided related to specific sub-sections.

Section 8 of the APA Ethics Code (APA, 2021)

8.01 Institutional Approval

When institutional approval is required, psychologists provide accurate information about their research proposals and obtain approval prior to conducting the research. They conduct the research in accordance with the approved research protocol.

8.02 Informed Consent to Research

(a) When obtaining informed consent as required in Standard 3.10, Informed Consent, psychologists inform participants about

1. (1) the purpose of the research, expected duration, and procedures;
2. (2) their right to decline to participate and to withdraw from the research once participation has begun;
3. (3) the foreseeable consequences of declining or withdrawing;
4. (4) reasonably foreseeable factors that may be expected to influence their willingness to participate such as potential risks, discomfort, or adverse effects;
5. (5) any prospective research benefits;
6. (6) limits of confidentiality;
7. (7) incentives for participation; and
8. (8) whom to contact for questions about the research and research participants' rights. They provide opportunity for the prospective participants to ask questions and receive answers.

(See also Standards 8.03, Informed Consent for Recording Voices and Images in Research; 8.05, Dispensing with Informed Consent for Research; and 8.07, Deception in Research .)

(b) Psychologists conducting intervention research involving the use of experimental treatments clarify to participants at the outset of the research:

1. (1) the experimental nature of the treatment;

2. (2) the services that will or will not be available to the control group(s) if appropriate;
3. (3) the means by which assignment to treatment and control groups will be made;
4. (4) available treatment alternatives if an individual does not wish to participate in the research or wishes to withdraw once a study has begun; and
5. (5) compensation for or monetary costs of participating including, if appropriate, whether reimbursement from the participant or a third-party payor will be sought. (See also Standard 8.02a, Informed Consent to Research .)

8.03 Informed Consent for Recording Voices and Images in Research

Psychologists obtain informed consent from research participants prior to recording their voices or images for data collection unless

1. (1) the research consists solely of naturalistic observations in public places, and it is not anticipated that the recording will be used in a manner that could cause personal identification or harm, or
2. (2) the research design includes deception, and consent for the use of the recording is obtained during debriefing. (See also Standard 8.07, Deception in Research .)

8.04 Client/Patient, Student, and Subordinate Research Participants

(a) When psychologists conduct research with clients/patients, students, or subordinates as participants, psychologists take steps to protect the prospective participants from adverse consequences of declining or withdrawing from participation.

(b) When research participation is a course requirement or an opportunity for extra credit, the prospective participant is given the choice of equitable alternative activities.

8.05 Dispensing with Informed Consent for Research

Psychologists may dispense with informed consent only

1. (1) where research would not reasonably be assumed to create distress or harm and involves
    1. (a) the study of normal educational practices, curricula, or classroom management methods conducted in educational settings;
    2. (b) only anonymous questionnaires, naturalistic observations, or archival research for which disclosure of responses would not place participants at risk of criminal or civil liability or damage their financial standing, employability, or reputation, and confidentiality is protected; or
    3. (c) the study of factors related to job or organization effectiveness conducted in organizational settings for which there is no risk to participants' employability, and confidentiality is protected or
2. (2) where otherwise permitted by law or federal or institutional regulations.

8.06 Offering Inducements for Research Participation

(a) Psychologists make reasonable efforts to avoid offering excessive or inappropriate financial or other inducements for research participation when such inducements are likely to coerce participation.

(b) When offering professional services as an inducement for research participation, psychologists clarify the nature of the services, as well as the risks, obligations, and limitations. (See also Standard 6.05, Barter with Clients/Patients .)

8.07 Deception in Research

(a) Psychologists do not conduct a study involving deception unless they have determined that the use of deceptive techniques is justified by the study's significant prospective scientific, educational, or applied value and that effective nondeceptive alternative procedures are not feasible.

(b) Psychologists do not deceive prospective participants about research that is reasonably expected to cause physical pain or severe emotional distress.

(c) Psychologists explain any deception that is an integral feature of the design and conduct of an experiment to participants as early as is feasible, preferably at the conclusion of their participation, but no later than at the conclusion of the data collection, and permit participants to withdraw their data. (See also Standard 8.08, Debriefing .)

8.08 Debriefing

(a) Psychologists provide a prompt opportunity for participants to obtain appropriate information about the nature, results, and conclusions of the research, and they take reasonable steps to correct any misconceptions that participants may have of which the psychologists are aware.

(b) If scientific or humane values justify delaying or withholding this information, psychologists take reasonable measures to reduce the risk of harm.

(c) When psychologists become aware that research procedures have harmed a participant, they take reasonable steps to minimize the harm.

8.09 Humane Care and Use of Animals in Research

(a) Psychologists acquire, care for, use, and dispose of animals in compliance with current federal, state, and local laws and regulations, and with professional standards.

(b) Psychologists trained in research methods and experienced in the care of laboratory animals supervise all procedures involving animals and are responsible for ensuring appropriate consideration of their comfort, health, and humane treatment.

(c) Psychologists ensure that all individuals under their supervision who are using animals have received instruction in research methods and in the care, maintenance, and handling of the species being used, to the extent appropriate to their role. (See also Standard 2.05, Delegation of Work to Others .)

(d) Psychologists make reasonable efforts to minimize the discomfort, infection, illness, and pain of animal subjects.

(e) Psychologists use a procedure subjecting animals to pain, stress, or privation only when an alternative procedure is unavailable and the goal is justified by its prospective scientific, educational, or applied value.

(f) Psychologists perform surgical procedures under appropriate anesthesia and follow techniques to avoid infection and minimize pain during and after surgery.

(g) When it is appropriate that an animal's life be terminated, psychologists proceed rapidly, with an effort to minimize pain and in accordance with accepted procedures.

8.10 Reporting Research Results

(a) Psychologists do not fabricate data. (See also Standard 5.01a, Avoidance of False or Deceptive Statements .)

(b) If psychologists discover significant errors in their published data, they take reasonable steps to correct such errors in a correction, retraction, erratum, or other appropriate publication means.

8.11 Plagiarism

Psychologists do not present portions of another's work or data as their own, even if the other work or data source is cited occasionally.

8.12 Publication Credit

(a) Psychologists take responsibility and credit, including authorship credit, only for work they have actually performed or to which they have substantially contributed. (See also Standard 8.12b, Publication Credit .)

(b) Principal authorship and other publication credits accurately reflect the relative scientific or professional contributions of the individuals involved, regardless of their relative status. Mere possession of an institutional position, such as department chair, does not justify authorship credit. Minor contributions to the research or to the writing for publications are acknowledged appropriately, such as in footnotes or in an introductory statement.

(c) Except under exceptional circumstances, a student is listed as principal author on any multiple-authored article that is substantially based on the student's doctoral dissertation. Faculty advisors discuss publication credit with students as early as feasible and throughout the research and publication process as appropriate. (See also Standard 8.12b, Publication Credit .)

8.13 Duplicate Publication of Data

Psychologists do not publish, as original data, data that have been previously published. This does not preclude republishing data when they are accompanied by proper acknowledgment.

8.14 Sharing Research Data for Verification

(a) After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release. This does not preclude psychologists from requiring that such individuals or groups be responsible for costs associated with the provision of such information.

(b) Psychologists who request data from other psychologists to verify the substantive claims through reanalysis may use shared data only for the declared purpose. Requesting psychologists obtain prior written agreement for all other uses of the data.

8.15 Reviewers

Psychologists who review material submitted for presentation, publication, grant, or research proposal review respect the confidentiality of and the proprietary rights in such information of those who submitted it.

Institutional Approval

Section 8.01 is related to Institutional Review Boards (IRB). Universities, hospitals, and other institutions that receive support from the federal government must establish an institutional review board (IRB), which is a committee that is responsible for reviewing research protocols for potential ethical problems. An IRB must consist of at least five people with varying backgrounds, including members of different professions, scientists and nonscientists, men and women, and at least one person not otherwise affiliated with the institution (a community member). The IRB helps to make sure that the risks of the proposed research are minimized, the benefits outweigh the risks, the research is carried out in a fair manner, and the informed consent procedure is adequate.

The federal regulations discussed in the previous section (3.2: History of Ethics Cods) distinguish research that poses three levels of risk:

1. Exempt
2. Expedited
3. More than Minimal Risk

Exempt research is the lowest level or risk and includes research on the effectiveness of normal educational activities, the use of standard psychological measures and surveys of a nonsensitive nature that are administered in a way that maintains confidentiality, and research using existing data from public sources. It is called exempt because once approved, it is exempt from regular, continuous review. Expedited research poses a somewhat higher risk than exempt, but still exposes participants to risks that are no greater than minimal risk (those encountered by healthy people in daily life or during routine physical or psychological examinations). Expedited review is done by by one member of the IRB or by a separate committee under the authority of the IRB that can only approve minimal risk research (many departments of psychology have such separate committees). Finally, research that does not qualify for exempt or expedited review is greater than minimal risk research must be reviewed by the full board of IRB members.

Recently, federal regulations have changed to make applications and review easier for both researchers and the IRB (DeAngelis, 2024). DeAngelis (2024) notes how the Federal Policy for Protection of Human Subjects, known as the Common Rule (Health and Human Services) was revised in 2019 that makes more types of studies exempt levels of risk. Now, most typical studies of people are exempt from further review for 20 federal agencies (Office for Human Research Protections, 2025); only studies that could cause harm would face a higher level of review (DeAngelis, 2024). Some of the new exemptions include:

1. Studies in educational studies about educational practices
2. Most self-report measures, educational tests, and observation of public behavior
3. Studies that use brief and non-harmful interventions, such as responding to materials or written scenarios
4. Using secondary data (data collected by someone else)

Researchers expect that research that could cause physical, psychological, or emotional harm would most likely be longitudinal studies that involve vulnerable populations (DeAngelis, 2024). There are also exceptions to these exemptions! First, for self-report measures that ask about sensitive issues (mental health, illegal conduct, sexual activity, finances, discrimination, or other social damaging information) will undergo a limited review; this would take less time than a full IRB review. Second, studies that focus on protected populations will not be exempt. Protected populations are those that are particularly vulnerable or at a risk for coercion, such as prisoners, children, people with mental disabilities, or people who are poor or unhoused. Pregnant people and human fetuses are also considered vulnerable.

DeAngelis (2024) discusses a few other changes to the Common Rule, but they are less related to IRB.

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. Section 8.02(a) lists eight necessary components of informed consent forms. Section 8.02(b) includes additional requirements for studies that include an intervention, rather than recording observational or self-report data only.

? Exercise 3.3.1

Based on Section 8.02(b), what information seems required othat is in addition to what's already required in informed consent forms based on 8.02(a)?

Answer

Section 8.02(b) states that "Psychologists conducting intervention research involving the use of experimental treatments clarify to participants at the outset of the research:

1. The experimental nature of the treatment;
2. The services that will or will not be available to the control group(s) if appropriate;
3. The means by which assignment to treatment and control groups will be made;
4. Available treatment alternatives if an individual does not wish to participate in the research or wishes to withdraw once a study has begun; and
5. Compensation for or monetary costs of participating including, if appropriate, whether reimbursement from the participant or a third-party payor will be sought."

It seems like when there is an intervention, participants must be told what the different groups are, and how participants were assigned into a control group or who decline to participate or choose to withdraw from the study can receive similar treatment, and how the treatment and their participation is being compensated.

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 (Mann, 1994). 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 (see Section 8.05). 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, they would not need to obtain informed consent from his students.

An additional complexity arises in research on children or adults who cannot provide consent (such as inebriated adults, adults with development disabilities, or when an elderly person no longer makes their own life decisions). According to the Common Rule (Health and Human Services, 2025), people who cannot give official consent must have a guardian provide informed consent *and*

the participant should provide assent, or agreeing verbally to participate after the study is explained to them in terms that they can understand.

Necessary Components of Informed Consent Forms

1. Basic information about the study:
    1. Purpose of the research,
    2. Expected duration of the study
    3. Procedures (including details that could influence whether or not someone would choose to participate)
2. Participants' right to decline to participate and to withdraw from the research once participation has begun, and
3. Any foreseeable consequences of declining or withdrawing [Researchers should avoid any negative consequences for declining or withdrawing in order to no coerce participation.]
4. Reasonably foreseeable factors that may be expected to influence their willingness to participate such as potential risks, discomfort, or adverse effects
5. Any prospective research benefits; for example, the form can include something like "The results of this research may contribute to scientific understanding of how social media can affect mental health in young adults. Additionally, you will learn about research methods by experiencing research of this type."
6. Limits of confidentiality, which means who will have access to the data (in full or aggregated). Additionally, some states require researchers who learn of child abuse or other crimes to report this information to authorities.
7. Incentives for participation such as money, their name in a raffle, access to the therapy or program, or extra credit. Many universities require participation in several research studies as a requirement in specific courses, so fulfilling the course requirement would be considered the incentive.
8. Finally, provide contact information for the primary researcher for questions about the research and research participants' rights. Studies must also provide an opportunity for the prospective participants to ask questions and receive answers.
9. For intervention research involving the use of experimental treatments, the informed consent should also include:
    1. The experimental nature of the treatment;
    2. The services that will or will not be available to the control group(s) if appropriate;
    3. The means by which assignment to treatment and control groups will be made;
    4. Available treatment alternatives if an individual does not wish to participate in the research or wishes to withdraw once a study has begun; and
    5. Compensation for or monetary costs of participating including, if appropriate, whether reimbursement from the participant or a third-party payor will be sought."

The following is an example informed consent. Can you find the required components for an informed consent form?

Informed Consent

Please take the time to review this consent form and ask any questions you may have before deciding whether to participate.

You are invited to participate in an experimental research study aimed at assessing the effects of social media. This study will take 6 days. The survey will take about 15-20 minutes of your time, which you will take twice, plus a few minutes every day to fill in the daily log of time spent on social media. You will be randomly assigned through our course management system's quiz function into a group that either has unlimited time on social media apps or is asked to limit their social media usage to two hours a day for a length of 5 days. On the first day and the last day, you will be asked to take a survey. Participants’ data will be collected but will be anonymous. The only person with direct access to names will be the professor involved.

Your participation in this study is voluntary. You have the right to decline to participate or to withdraw from the study at any time without fear of punishment. If you choose to withdraw, you will not earn extra credit for this activity. However, there are other extra credit opportunities that you could choose instead. There are no foreseeable negative consequences if you choose to decline participation or withdraw from the study. There are other extra credit opportunities that you could choose instead.

You may find it difficult to limit your time on social media. This may result in anxiety or depression. Also, you will need to turn on the tracking function on your phone. Some may feel uncomfortable with allowing this function.

The results of this research may contribute to scientific understanding of how social media can affect mental health in young adults. Additionally, you will learn about research methods by experiencing research of this type.

Your responses will be kept confidential. When the research is completed the research will be presented in a manner that will provide summary results and you will be anonymous.

Participants will be given extra credit. Alternative extra credit opportunities are available on our course management system if you decline to participate or withdraw from the study.

If you have any questions or concerns about the study, you may contact Primary Researcher via email (PI@YourCollege.edu) or phone (123-456-7890). For further information including final results, you may contact Primary Researcher at the end of the study.

By agreeing to participate in this study, you indicate that YOU ARE 18 YEARS OLD OR OLDER and have read and understood the information provided in this form and voluntarily consent to participate.

18 years old: Yes _____ No _____

Participant Signature: _____ Date: _____

In the example, did you look for the addition five components needed for research involving experimental treatments? Do you think that this study needs the additional five components?

What do you think?

Do you think that the study described in the sample informed consent needs the additional five components? Why or why not?

Inducements

Much of social science research relies on some sort of reward for participant in the study. As noted in the Necessary Components of Informed Consent box, these rewards an be financial (money or gift cards, participants' chance to win a raffle) or access to services (such as therapy or a program that is being tested). For students, the incentive can be extra credit for students or fulfilling a required course assignment. The concern with any of these is listed in Section 8.06, is to avoid coercion. As noted in many ethics codes, participants should always have the right to withdraw from any study and to *decline to participate* in the first place. If a worker could not afford food for their family, and participating in a study would provide for that food, does the worker really have a choice to decline to participate? Section 8.06 requires researchers to not offer excessive inducements in order to make participation an option that participants could realistically decline.

Deception

Section 8.07 covers deception in research. Deception of participants in psychological research can take a variety of forms: misinforming participants about the purpose of a study, using confederates (actors who play the part of other participants but are really part of the study), using phony equipment like Milgram’s shock generator, and presenting participants with false feedback about their performance (e.g., telling them they did poorly on a test when they actually did well). Deception also includes not informing participants of the full design or true purpose of the research even if they are not actively misinformed (Sieber et al., 1995). For example, a study on incidental learning, or learning without conscious effort, might involve having participants read through a list of words in preparation for a “memory test” later. Although participants are likely to assume that the memory test will require them to recall the words, it might instead require them to recall the contents of the room or the appearance of the research assistant.

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

Note, however, that the APA Ethics Code takes a more moderate approach. First, deception must only be used when the research question cannot be answered without the use of deception. Additionally, Section 8.07 allows deception when participants cannot reasonably be expected to be harmed and participants are informed about the deception as soon as possible. This approach acknowledges that not all forms of deception are equally bad. Compare, for example, Milgram's study in which he deceived his participants in several significant ways that resulted in their experiencing severe psychological stress with an incidental learning study in which a "memory test" turns out to be slightly different from what participants were expecting. It also acknowledges that some scientifically and socially important research questions can be difficult or impossible to answer without deceiving participants. Knowing that a study concerns the extent to which they obey authority, act aggressively toward a peer, or help a stranger is likely to change the way people behave so that the results no longer generalize to the real world.

Debriefing

Section 8.08 is about debriefing. This is the process of informing research participants as soon as possible after they have completed the study of important features of the study and their participation. This could include sharing the purpose of the study, letting them know if they were in the intervention group or placebo group (control group), and correcting any other misconceptions they might have as a result of participating. This debrief is also when any deception would be revealed, and participants would be provided correct information.

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

Finally, research that involves an intervention may benefit from a manipulation check during the debriefing. A manipulation check is when the researcher asks the participants questions to confirm that they actually experienced the IV as it was intended. For example, in a study in which participants rated the personality of a character that they had read about, a researcher could ask about information from the paragraph to confirm that the participants actually read the information and remembered enough to adequately rate that character's personality based on the information provided.

Thus, while a debriefing might not seem important, it is critical for some research design to ensure honesty from both the researcher and the participant.

Nonhuman Animal Subjects

Although most contemporary research in psychology does not involve nonhuman animal subjects, a significant minority of it does. Social science research that may be likely to use animals include the study of learning and conditioning, behavioral neuroscience, and the development of drug and surgical therapies for psychological disorders. Section 8.09 is about the humane treatment and care of these nonhuman animal subjects. While this is part of the APA's ethics code, the APA must follow the National Institute of Health's (NIH) Office of Laboratory Animal Welfare guidelines, which are created by the Institutional Animal Care and Use Committee (NIH, 2024).

The use of nonhuman animal subjects in psychological research is similar to 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 zoo-like 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 CARE, which stands for Committee on Animal Research and Ethics (APA, 2024).

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. 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. Peer review is a cornerstone of scientific knowledge. In this day and age, anyone can post a research report online. We rely on scientific peers to ensure that the study is high quality and that the results do not appear to be fabricated. To learn more about what peer reviewers are looking for, Wiley Author Services provides an online guide (Wiley, 2025).

Academic careers are built on publishing research, so properly and ethically giving credit to the people who contributed the most intellectually (not necessarily the most time) to a project is critical. The order of authors' names on published research traditionally 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.

What do you think?

A fun application of the APA ethic code (APA, 2021) could be to review controversial research studies, and see if you think that they would be able to be conducted with today's ethics code. A research to start might be this article about controversial psychological studies (Jarrett, 2014).

Moral Principles

As we wrap up this chapter on ethics, let's touch on one more topic: Moral Principles

---

References

American Psychological Association. (2021, January 1). *Ethnical principles of psychologists and code of conduct*. https://www.apa.org/ethics/code/index

American Psychological Association, (2024). *Committee on Animal Research and Ethics*. https://www.apa.org/science/leadership/care/index

American Sociological Association. (2023, September 13). *Ethics*. https://www.asanet.org/about/ethics/

Baumrind, D. (1985). Research using intentional deception: Ethical issues revisited. *American Psychologist, 40*, 165–174.

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

DeAngelis, T. (2024). *A lighter load for researchers, 55*(7), 78-87.

Health and Human Services (2025, June 11). *45 Code of Federal Regulation 4*6. U.S. Department of Health and Human Services. https://www.hhs.gov/ohrp/regulations...-46/index.html

Jarrett, C. (2014, September 19). *The 10 most controversial psychology studies ever published*. The British Psychological Society. https://www.bps.org.uk/research-dige...ever-published

Mann, T. (1994). Informed consent for psychological research: Do subjects comprehend consent forms and understand their legal rights? *Psychological Science, 5*, 140–143.

Miller, N. E. (1985). The value of behavioral research on animals. *American Psychologist, 40*, 423–440.

National Association of Social Workers (2021). *Code of ethics*. https://www.socialworkers.org/about/...code-of-ethics

National Institute of Health (2024, October 30). *Institutional Animal Care and Use Committee*. Office of Laboratory Animal Welfare. https://olaw.nih.gov/resources/tutorial/iacuc.htm

Office for Human Research Protections. (2025, June 5). *Federal policy for the protection of human subjects ('Common Rule')*. U.S. Department of Health and Human Services. https://www.hhs.gov/ohrp/regulations...ule/index.html

Sieber, J. E., Iannuzzo, R., & Rodriguez, B. (1995). Deception methods in psychology: Have they changed in 23 years? *Ethics & Behavior, 5*, 67–85.

Wiley Author Services (2025). *Step by step guide to reviewing a manuscript*. https://authorservices.wiley.com/Rev...anuscript.html

---

This page titled 3.3: APA Ethics Code Standard 8 is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **3.3: From Moral Principles to Ethics Codes** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

3.4: Principles of Ethical Research

Learning Objectives

1. Describe a simple framework for thinking about ethical issues in psychological research.
2. Give examples of several ethical issues that arise in research that may affect research participants, the scientific community, or society more generally.

Ethics is the branch of philosophy that is concerned with morality—what it means to behave morally and how people can achieve that goal. It can also refer to a set of principles and practices that provide moral guidance in a particular field. There is an ethics of business, medicine, teaching, and of course, scientific research. As the opening example illustrates, many kinds of ethical issues can arise in scientific research, especially when it involves human participants. For this reason, it is useful to begin with a general framework for thinking through these issues.

A Framework for Thinking About Research Ethics

Table **3.4.1** presents a framework for thinking through the ethical issues involved in research on humans. The rows of Table **3.4.1** represent four general moral principles that apply to scientific research adapted from the APA ethics code (APA, 2021): 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 APA Ethics Code.) The columns of Table **3.4.1** 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 **3.4.1**: A Framework for Thinking About Ethical Issues in 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			

Moral Principles

Let us look more closely at each of the moral principles and how they can be applied to each of the three groups.

Weighing Risks Against Benefits

Scientific research in psychology can be ethical only if its risks are outweighed by its benefits. Among the risks to research participants are that a treatment might fail to help or even be harmful, a procedure might result in physical or psychological harm, and their right to privacy might be violated. Among the potential benefits are receiving a helpful treatment, learning about psychology, experiencing the satisfaction of contributing to scientific knowledge, and receiving money or course credit for participating. Scientific research can have risks and benefits to the scientific community and to society too (Rosenthal, 1994). 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. The research that mistakenly linked the measles, mumps, and rubella (MMR) vaccine to autism resulted in both of these kinds of harm. Of course, the benefits of scientific research to science and society are that it advances scientific knowledge and can contribute to the welfare of society.

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

Was It Worth It?

Much of the debate over the ethics of Milgram's obedience study concerns the question of whether the resulting scientific knowledge was worth the harm caused to the research participants. To get a better sense of the harm, consider Milgram's (1963) own description of it.

> In a large number of cases, the degree of tension reached extremes that are rarely seen in sociopsychological laboratory studies. Subjects were observed to sweat, tremble, stutter, bite their lips, groan, and dig their fingernails into their flesh.…Fourteen of the 40 subjects showed definite signs of nervous laughter and smiling. The laughter seemed entirely out of place, even bizarre. Full-blown uncontrollable seizures [of laughter] were observed for three subjects. On one occasion we observed a seizure so violently convulsive that it was necessary to call a halt to the experiment (p. 375).

Milgram also noted that another observer reported that within 20 minutes one participant "was reduced to a twitching, stuttering wreck, who was rapidly approaching the point of nervous collapse" (p. 377)

To Milgram's credit, he went to great lengths to debrief his participants—including returning their mental states to normal—and to show that most of them thought the research was valuable and they were glad to have participated.

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 their professional obligations, and being truthful. Acting with integrity is important because it promotes trust, which is an essential element of all effective human relationships. Participants must be able to trust that researchers are being honest with them (e.g., about what the study involves), will keep their promises (e.g., to maintain confidentiality), and will carry out their research in ways that maximize benefits and minimize risk. An important issue here is the use of deception. Some research questions (such as Milgram's) are difficult or impossible to answer without deceiving research participants. Thus acting with integrity can conflict with doing research that advances scientific knowledge and benefits society. We will consider how psychologists generally deal with this conflict shortly.

The scientific community and society must also be able to trust that researchers have conducted their research thoroughly and competently and that they have reported on it honestly. Again, the example at the beginning of the chapter illustrates what can happen when this trust is violated. In this case, other researchers wasted resources on unnecessary follow-up research and people avoided the MMR vaccine, putting their children at increased risk of measles, mumps, and rubella. Indeed, many people, including children have died as a result of parents' misinformed decisions not to vaccinate their children.

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.

At a broader societal level, members of some groups have historically faced more than their fair share of the risks of scientific research, including people who are institutionalized, are disabled, or belong to racial or ethnic minorities. The box below shared more information about the Tusegee Syphilis Study that was first discussed in section 3.3 (History of Ethics Codes) of this chapter.

"They Were Betrayed"

In 1997—65 years after the Tuskegee Syphilis Study began and 25 years after it ended—President Bill Clinton formally apologized on behalf of the US government to those who were affected. Here is an excerpt from the apology:

So today America does remember the hundreds of men used in research without their knowledge and consent. We remember them and their family members. Men who were poor and African American, without resources and with few alternatives, they believed they had found hope when they were offered free medical care by the United States Public Health Service. They were betrayed.

Read the full text of the apology at http://www.cdc.gov/tuskegee/clintonp.htm.

Respecting People's Rights and Dignity

Researchers must respect people's rights and dignity as human beings. One element of this is respecting their autonomy—their right to make their own choices and take their own actions free from coercion. Of fundamental importance here is the concept of informed consent. This means that researchers obtain and document people's agreement to participate in a study after having informed them of everything that might reasonably be expected to affect their decision. Consider the participants in the Tuskegee study. Although they agreed to participate in the study, they were not told that they had syphilis but would be denied treatment for it. Had they been told this basic fact about the study, it seems likely that they would not have agreed to participate. Likewise, had participants in Milgram's study been told that they might be "reduced to a twitching, stuttering wreck," it seems likely that many of them would not have agreed to participate. In neither of these studies did participants give true informed consent.

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. Even more ideally participants can maintain anonymity**,** which is when their name and other personally identifiable information is not collected at all. As more information is available digitally (Winerman, 2018), a new avenue of privacy must be acknowledged.

Unavoidable Ethical Conflict

It may already be clear that ethical conflict in psychological research is unavoidable. Because there is little, if any, psychological research that is completely risk-free, there will almost always be a conflict between risks and benefits. Research that is beneficial to one group (e.g., the scientific community) can be harmful to another (e.g., the research participants), creating especially difficult tradeoffs. We have also seen that being completely truthful with research participants can make it difficult or impossible to conduct scientifically valid studies on important questions.

Of course, many ethical conflicts are fairly easy to resolve. Nearly everyone would agree that deceiving research participants and then subjecting them to physical harm would not be justified by filling a small gap in the research literature. But many ethical conflicts are not easy to resolve, and competent and well-meaning researchers can disagree about how to resolve them. Consider, for example, an actual study on "personal space" conducted in a public men's room (Middlemist et al., 1976). Middlemist et al. (1976) secretly observed their participants to see whether it took them longer to begin urinating when there was another man (a confederate of the researchers) at a nearby urinal. While some critics found this to be an unjustified assault on human dignity (Koocher, 1977), Middlemist et al. (1976) had carefully considered the ethical conflicts, resolved them as best they could, and concluded that the benefits of the research outweighed the risks. For example, Middlemist et al. (1976) had interviewed some preliminary participants and found that none of them was bothered by the fact that they had been observed.

The point here is that although it may not be possible to eliminate ethical conflict completely, it is possible to deal with it in responsible and constructive ways. In general, this means thoroughly and carefully thinking through the ethical issues that are

raised, minimizing the risks, and weighing the risks against the benefits. It also means being able to explain one's ethical decisions to others, seeking feedback on them, and ultimately taking responsibility for them.

References

American Psychological Association. (2021, January 1). *Ethnical principles of psychologists and code of conduct*. https://www.apa.org/ethics/code/index

Koocher, G. P. (1977). Bathroom behavior and human dignity. *Journal of Personality and Social Psychology, 35*, 120–121.

Middlemist, R. D., Knowles, E. S., & Matter, C. F. (1976). Personal space invasions in the lavatory: Suggestive evidence for arousal. *Journal of Personality and Social Psychology, 33*, 541–546.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371–378.

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

Rosenthal, R. M. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science, 5*, 127–133.

Winerman, L. (2018). Big data gets bigger. *Monitory on Psychology, 4*9(10). https://www.apa.org/monitor/2018/11/cover-dataKoocher, G. P. (1977). Bathroom behavior and human dignity. *Journal of Personality and Social Psychology, 35*, 120–121. ↵

This page titled 3.4: Principles of Ethical Research is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **3.2: Moral Foundations of Ethical Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

3.5: Summary of Research Ethics

Key Takeaways

- A wide variety of ethical issues arise in psychological research. Thinking them through requires considering how each of four moral principles (weighing risks against benefits, acting responsibly and with integrity, seeking justice, and respecting people's rights and dignity) applies to each of three groups of people (research participants, science, and society).
- Ethical conflict in psychological research is unavoidable. Researchers must think through the ethical issues raised by their research, minimize the risks, weigh the risks against the benefits, be able to explain their ethical decisions, seek feedback about these decisions from others, and ultimately take responsibility for them.
- There are several written ethics codes for research with human participants that provide specific guidance on the ethical issues that arise most frequently. These codes include the Nuremberg Code, the Declaration of Helsinki, the Belmont Report, and the Federal Policy for the Protection of Human Subjects.
- The APA Ethics Code is the most important ethics code for researchers in psychology. It includes many standards that are relevant mainly to clinical practice, but *Standard 8* concerns informed consent, deception, debriefing, the use of nonhuman animal subjects, and scholarly integrity in research.
- Research conducted at universities, hospitals, and other institutions that receive support from the federal government must be reviewed by an institutional review board (IRB)—a committee at the institution that reviews research protocols to make sure they conform to ethical standards.
- Informed consent is the process of obtaining and documenting people's agreement to participate in a study, having informed them of everything that might reasonably be expected to affect their decision. Although it often involves having them read and sign a consent form, it is not equivalent to reading and signing a consent form.
- Although some researchers argue that deception of research participants is never ethically justified, the APA Ethics Code allows for its use when the benefits of using it outweigh the risks, participants cannot reasonably be expected to be harmed, there is no way to conduct the study without deception, and participants are informed of the deception as soon as possible.
- It is your responsibility as a researcher to know and accept your ethical responsibilities.
- You can take several concrete steps to minimize risks and deception in your research. These include making changes to your research design, prescreening to identify and eliminate high-risk participants, and providing participants with as much information as possible during informed consent and debriefing.
- Your ethical responsibilities continue beyond IRB approval. You need to monitor participants' reactions, be alert for potential violations of confidentiality, and maintain scholarly integrity through the publication process.

What's Next?

To finish out this first unit introducing you to research in the social sciences, we will learn about how to measure social and psychological variables.

Exercises

- Discussion: It has been argued that researchers are not ethically responsible for the misinterpretation or misuse of their research by others. Do you agree? Why or why not?
- Practice: Read the Nuremberg Code, the Belmont Report, and *Standard 8* of the APA Ethics Code. List five specific similarities and five specific differences among them.
- Discussion: In a study on the effects of disgust on moral judgment, participants were asked to judge the morality of disgusting acts, including people eating a dead pet and passionate kissing between a brother and sister (Haidt, et al., 1993). If you were on the IRB that reviewed this protocol, what concerns would you have with it? Refer to the appropriate sections of the APA Ethics Code.
- Practice: Explain how you could use the six steps of ethical inquiry from Qualters (2017) to an issue in your future career field.
- Discussion: How could you conduct a study on the extent to which people obey authority in a way that minimizes risks and deception as much as possible? (Note: Such a study would not have to look at all like Milgram's.)

- Practice: Find a study in a professional journal and create a consent form for that study. Be sure to include all the information in Standard 8.02.

References

Haidt, J., Koller, S. and Dias, M. (1993) Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*, 613-628. http://dx.doi.org/10.1037/0022-3514.65.4.613

Qualters, D. M. (2017, January 30). *Navigating ethical waters in the college classroom*. Faculty Focus. https://www.facultyfocus.com/article...ers-classroom/

This page titled 3.5: Summary of Research Ethics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **3.5: Research Ethics (Summary)** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

4: Measuring Variables

4.1: Prelude to Psychological Measurement
4.2: Introduction to Measuring Variables
4.2.1: Types of Validity
4.3: Introduction to Surveys
4.3.1: Constructing Surveys
4.3.2: Reliability and Validity of Measurement
4.4: Interviews and Focus Groups
4.5: Measuring Behavior through Biology
4.6: Observations
4.7: Unobtrusive Research
4.7.1: What Is Left Behind (Physical Trace)
4.7.2: Archival Data
4.8: Summary of Measuring Variables

4: Measuring Variables is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

4.1: Prelude to 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)[1]. 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.

Do You Feel You Are a Person of Worth?

The Rosenberg Self-Esteem Scale (Rosenberg, 1989)[2] 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.

1. 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. ↵
2. Rosenberg, M. (1989). *Society and the adolescent self-image* (rev. ed.). Middletown, CT: Wesleyan University Press. ↵

4.1: Prelude to Psychological Measurement is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **4.1: Prelude to Psychological Measurement** has no license indicated.

4.2: Introduction to Measuring Variables

Learning Objectives

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

What Is Measurement?

Measurement is the assignment of scores to individuals so that the scores represent some characteristic of the individuals. This very general definition is consistent with the kinds of measurement that everyone is familiar with—for example, weighing oneself by stepping onto a bathroom scale, or checking the internal temperature of a roasting turkey using a meat thermometer. It is also consistent with measurement in the other sciences. In physics, for example, one might measure the potential energy of an object in Earth's gravitational field by finding its mass and height (which of course requires measuring *those* variables) and then multiplying them together along with the gravitational acceleration of Earth (9.8 m/s2). The result of this procedure is a score that represents the object's potential energy.

This general definition of measurement is consistent with measurement in psychology too. (Psychological measurement is often referred to as **psychometrics**.) Imagine, for example, that a cognitive psychologist wants to measure a person's working memory capacity—their ability to hold in mind and think about several pieces of information all at the same time. To do this, she might use a backward digit span task, in which she reads a list of two digits to the person and asks them to repeat them in reverse order. She then repeats this several times, increasing the length of the list by one digit each time, until the person makes an error. The length of the longest list for which the person responds correctly is the score and represents their working memory capacity. Or imagine a clinical psychologist who is interested in how depressed a person is. He administers the Beck Depression Inventory, which is a 21-item self-report questionnaire in which the person rates the extent to which they have felt sad, lost energy, and experienced other symptoms of depression over the past 2 weeks. The sum of these 21 ratings is the score and represents the person's current level of depression.

The important point here is that measurement does not require any particular instruments or procedures. What it *does* require is *some* systematic procedure for assigning scores to individuals or objects so that those scores represent the characteristic of interest.

Psychological Constructs

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

Psychological constructs cannot be observed directly. One reason is that they often represent *tendencies* to think, feel, or act in certain ways. For example, to say that a particular university student is highly extraverted does not necessarily mean that she is behaving in an extraverted 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 extraverted ways (e.g., being outgoing, enjoying social interactions) across a variety of situations. Another reason psychological constructs cannot be observed directly is that they often involve internal processes. Fear, for example, involves the activation of certain central and peripheral nervous system structures, along with certain kinds of thoughts, feelings, and behaviors—none of which is necessarily obvious to an outside observer. Notice also that neither extraversion 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.

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)[1].

Big Five Dimenson	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	Straight-forwardness	Altruism	Compliance	Modesty	Tender-Mindedness
Neuroticism	Worry	Anger	Discourage-ment	Self-Consciousness	Impulsivity	Vulnerability

Figure **4.2.1**: The Big Five Personality Dimensions

The **conceptual definition** of a psychological construct describes the behaviors and internal processes that make up that construct, along with how it relates to other variables. For example, a conceptual definition of neuroticism (another one of the Big Five) would be that it is people’s tendency to experience negative emotions such as anxiety, anger, and sadness across a variety of situations. This definition might also include that it 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 important, researchers are in the business of developing definitions that are more detailed and precise—and that more accurately describe the way the world is—than the informal definitions in the dictionary. As we will see, they do this by proposing conceptual definitions, testing them empirically, and revising them as necessary. Sometimes they throw them out altogether. This is why the research literature often includes different conceptual definitions of the same construct. In some cases, an older conceptual definition has been replaced by a newer one that fits and works better. In others, researchers are still in the process of deciding which of various conceptual definitions is the best.

Operational Definitions

An **operational definition** is a definition of a variable in terms of precisely how it is to be measured. These measures generally fall into one of three broad categories. **Self-report measures** are those in which participants report on their own thoughts, feelings, and actions, as with the Rosenberg Self-Esteem Scale (Rosenberg, 1965)[2]. **Behavioral measures** are those in which some other aspect of participants’ behavior is observed and recorded. This is an extremely broad category that includes the observation of people’s behavior both in highly structured laboratory tasks and in more natural settings. A good example of the former would be measuring working memory capacity using the backward digit span task. A good example of the latter is a famous operational definition of physical aggression from researcher Albert Bandura and his colleagues (Bandura, Ross, & Ross, 1961)[3]. They let each of several children play for 20 minutes in a room that contained a clown-shaped punching bag called a Bobo doll. They filmed each child and counted the number of acts of physical aggression the child committed. These included hitting the doll with a mallet, punching it, and kicking it. Their operational definition, then, was the number of these specifically defined acts that the child committed during the 20-minute period. Finally, **physiological measures** are those that involve recording any of a wide variety of physiological processes, including heart rate and blood pressure, galvanic skin response, hormone levels, and electrical activity and blood flow in the brain.

For any given variable or construct, there will be multiple operational definitions. Stress is a good example. A rough conceptual definition is that stress is an adaptive response to a perceived danger or threat that involves physiological, cognitive, affective, and behavioral components. But researchers have operationally defined it in several ways. The Social Readjustment Rating Scale (Holmes & Rahe, 1967)[4] is a self-report questionnaire on which people identify stressful events that they have experienced in the past year and assigns points for each one depending on its severity. For example, a man who has been divorced (73 points),

changed jobs (36 points), and had a change in sleeping habits (16 points) in the past year would have a total score of 125. The Hassles and Uplifts Scale (Delongis, Coyne, Dakof, Folkman & Lazarus, 1982) [5] is similar but focuses on everyday stressors like misplacing things and being concerned about one's weight. The Perceived Stress Scale (Cohen, Kamarck, & Mermelstein, 1983) [6] is another self-report measure that focuses on people's feelings of stress (e.g., "How often have you felt nervous and stressed?"). Researchers have also operationally defined stress in terms of several physiological variables including blood pressure and levels of the stress hormone cortisol.

When psychologists use multiple operational definitions of the same construct—either within a study or across studies—they are using **converging operations**. The idea is that the various operational definitions are "converging" or coming together on the same construct. When scores based on several different operational definitions are closely related to each other and produce similar patterns of results, this constitutes good evidence that the construct is being measured effectively and that it is useful. The various measures of stress, for example, are all correlated with each other and have all been shown to be correlated with other variables such as immune system functioning (also measured in a variety of ways) (Segerstrom & Miller, 2004)[7]. This is what allows researchers eventually to draw useful general conclusions, such as "stress is negatively correlated with immune system functioning," as opposed to more specific and less useful ones, such as "people's scores on the Perceived Stress Scale are negatively correlated with their white blood counts."

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)[8]. For example, the officials at a 100-m race could simply rank order the runners as they crossed the finish line (first, second, etc.), or they could time each runner to the nearest tenth of a second using a stopwatch (11.5 s, 12.1 s, etc.). In either case, they would be measuring the runners' times by systematically assigning scores to represent those times. But while the rank ordering procedure communicates the fact that the second-place runner took longer to finish than the first-place finisher, the stopwatch procedure also communicates *how much* longer the second-place finisher took. Stevens actually suggested four different **levels of measurement** (which he called "scales of measurement") that correspond to four types of information that can be communicated by a set of scores, and the statistical procedures that can be used with the information.

The **nominal level** of measurement is used for categorical variables and involves assigning scores that are category labels. Category labels communicate whether any two individuals are the same or different in terms of the variable being measured. For example, if you ask your participants about their marital status, you are engaged in nominal-level measurement. Or if you ask your participants to indicate which of several ethnicities they identify themselves with, you are again engaged in nominal-level measurement. The essential point about nominal scales is that they do not imply any ordering among the responses. For example, when classifying people according to their favorite color, there is no sense in which green is placed "ahead of" blue. Responses are merely categorized. Nominal scales thus embody the lowest level of measurement[9].

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

On the other hand, ordinal scales fail to capture important information that will be present in the other levels of measurement we examine. In particular, the difference between two levels of an ordinal scale cannot be assumed to be the same as the difference between two other levels (just like you cannot assume that the gap between the runners in first and second place is equal to the gap between the runners in second and third place). In our satisfaction scale, for example, the difference between the responses "very dissatisfied" and "somewhat dissatisfied" is probably not equivalent to the difference between "somewhat dissatisfied" and "somewhat satisfied." Nothing in our measurement procedure allows us to determine whether the two differences reflect the same

difference in psychological satisfaction. Statisticians express this point by saying that the differences between adjacent scale values do not necessarily represent equal intervals on the underlying scale giving rise to the measurements. (In our case, the underlying scale is the true feeling of satisfaction, which we are trying to measure.)

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

Interval scales are not perfect, however. In particular, they do not have a true zero point even if one of the scaled values happens to carry the name "zero." The Fahrenheit scale illustrates the issue. Zero degrees Fahrenheit does not represent the complete absence of temperature (the absence of any molecular kinetic energy). In reality, the label "zero" is applied to its temperature for quite accidental reasons connected to the history of temperature measurement. Since an interval scale has no true zero point, it does not make sense to compute ratios of temperatures. For example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees; no interesting physical property is preserved across the two ratios. After all, if the "zero" label were applied at the temperature that Fahrenheit happens to label as 10 degrees, the two ratios would instead be 30 to 10 and 90 to 40, no longer the same! For this reason, it does not make sense to say that 80 degrees is "twice as hot" as 40 degrees. Such a claim would depend on an arbitrary decision about where to "start" the temperature scale, namely, what temperature to call zero (whereas the claim is intended to make a more fundamental assertion about the underlying physical reality).

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

Finally, the **ratio level** of measurement involves assigning scores in such a way that there is a true zero point that represents the complete absence of the quantity. Height measured in meters and weight measured in kilograms are good examples. So are counts of discrete objects or events such as the number of siblings one has or the number of questions a student answers correctly on an exam. You can think of a ratio scale as the three earlier scales rolled up in one. Like a nominal scale, it provides a name or category for each object (the numbers serve as labels). Like an ordinal scale, the objects are ordered (in terms of the ordering of the numbers). Like an interval scale, the same difference at two places on the scale has the same meaning. However, in addition, the same ratio at two places on the scale also carries the same meaning (see Table **4.2.1**).

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

Another example of a ratio scale is the amount of money you have in your pocket right now (25 cents, 50 cents, etc.). Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this actually implies the absence of money. Since money has a true zero point, it makes sense to say that someone with 50 cents has twice as much money as someone with 25 cents.

Stevens's levels of measurement are important for at least two reasons. First, they emphasize the generality of the concept of measurement. Although people do not normally think of categorizing or ranking individuals as measurement, in fact, they are as long as they are done so that they represent some characteristic of the individuals. Second, the levels of measurement can serve as a rough guide to the statistical procedures that can be used with the data and the conclusions that can be drawn from them. With nominal-level measurement, for example, the only available measure of central tendency is the mode. With ordinal-level measurement, the median or mode can be used as indicators of central tendency. Interval and ratio-level measurement are typically considered the most desirable because they permit for any indicators of central tendency to be computed (i.e., mean, median, or mode). Also, ratio-level measurement is the only level that allows meaningful statements about ratios of scores. Once again, one cannot say that someone with an IQ of 140 is twice as intelligent as someone with an IQ of 70 because IQ is measured at the interval level, but one can say that someone with six siblings has twice as many as someone with three because number of siblings is measured at the ratio level.

Table **4.2.1** Summary of Levels of Measurements

Level of Measurement	Category labels	Rank order	Equal intervals	True zero
NOMINAL	X			
ORDINAL	X	X		
INTERVAL	X	X	X	
RATIO	X	X	X	X

References

1. Costa, P. T., Jr., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4*, 5–13. ↵
2. Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press ↵
3. 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. ↵
4. Holmes, T. H., & Rahe, R. H. (1967). The Social Readjustment Rating Scale. *Journal of Psychosomatic Research, 11*(2), 213-218. ↵
5. 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*(2), 119-136. ↵
6. Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*, 386-396. ↵
7. 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. ↵
8. Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680. ↵
9. Levels of Measurement. Retrieved from http://wikieducator.org/Introduction_to_Research_Methods_In_Psychology/Theories_and_Measurement/Levels_of_Measurement↵

This page titled 4.2: Introduction to Measuring Variables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **4.2: Understanding Psychological Measurement** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

4.2.1: Types of Validity

Learning Objectives

1. Explain what internal validity is and why experiments are considered to be high in internal validity.
2. Explain what external validity is and evaluate studies in terms of their external validity.
3. Explain the concepts of construct and statistical validity.

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)[1][2]: internal validity, external validity, construct validity, and statistical validity. We will explore each validity in depth.

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.

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, McGraw, Bartels, & Warren, 2014)[3]. 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, Roberts, Noll, Quinn, & Twenge, 1998)[4]. 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) [5]. 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.

Construct Validity

In addition to the generalizability of the results of an experiment, another element to scrutinize in a study is the quality of the experiment's manipulations or the **construct validity** . The research question that Darley and Latané started with is "does helping behavior become diffused?" They hypothesized that participants in a lab would be less likely to help when they believed there were more potential helpers besides themselves. This conversion from research question to experiment design is called **operationalization** (see Chapter 4 for more information about the operational definition). Darley and Latané operationalized the independent variable of diffusion of responsibility by increasing the number of potential helpers. In evaluating this design, we would say that the construct validity was very high because the experiment's manipulations very clearly speak to the research question; there was a crisis, a way for the participant to help, and increasing the number of other students involved in the discussion, they provided a way to test diffusion.

What if the number of conditions in Darley and Latané's study changed? Consider if there were only two conditions: one student involved in the discussion or two. Even though we may see a decrease in helping by adding another person, it may not be a clear demonstration of diffusion of responsibility, just merely the presence of others. We might think it was a form of Bandura's concept of social inhibition. The construct validity would be lower. However, had there been five conditions, perhaps we would see the decrease continue with more people in the discussion or perhaps it would plateau after a certain number of people. In that situation, we may develop a more nuanced understanding of the phenomenon. But by adding still more conditions, the construct validity may not get higher. When designing your own experiment, consider how well the research question is operationalized your study.

Statistical Validity

Statistical validity concerns the proper statistical treatment of data and the soundness of the researchers' statistical conclusions. There are many different types of inferential statistics tests (e.g., *t*-tests, ANOVA, regression, correlation) and statistical validity concerns the use of the proper type of test to analyze the data. When considering the proper type of test, researchers must consider the scale of measure their dependent variable was measured on and the design of their study. Further, many inferential statistics

tests carry certain assumptions (e.g., the data are normally distributed) and statistical validity is threatened when these assumptions are not met but the statistics are used nonetheless.

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 (subsequent chapters will discuss how small samples, even of only one person, are still very illuminating for psychological research). Therefore, small sample sizes are actually a critique of statistical validity . The statistical validity speaks to whether the statistics conducted in the study are sound and support the conclusions that are made.

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.

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, Cialdini, & Griskevicius, 2008)[6]. Morling (2014) points out that many psychology studies have high internal and construct validity but sometimes sacrifice external validity.

References

1. Judd, C.M. & Kenny, D.A. (1981). *Estimating the effects of social interventions*. Cambridge, MA: Cambridge University Press. ↵
2. Morling, B. (2014, April). Teach your students to be better consumers. *APS Observer*. Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2014/april-14/teach-your-students-to-be-better-consumers.html↵
3. 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. ↵
4. 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*, 269–284. ↵
5. Cialdini, R. (2005, April). Don't throw in the towel: Use social influence research. *APS Observer*. Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2005/april-05/dont-throw-in-the-towel-use-social-influence-research.html↵
6. 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*, 472–482. ↵

This page titled 4.2.1: Types of Validity is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **5.4: Experimentation and Validity** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

4.3: Introduction to Surveys

Learning Objectives

1. Define what survey research is, including its two important characteristics.
2. Describe several different ways that survey research can be used and give some examples.

What Is Survey Research?

Survey research is a quantitative and qualitative method with two important characteristics. First, the variables of interest are measured using self-reports (using questionnaires or interviews). In essence, survey researchers ask their participants (who are often called **respondents** in survey research) to report directly on their own thoughts, feelings, and behaviors. Second, considerable attention is paid to the issue of sampling. In particular, survey researchers have a strong preference for large random samples because they provide the most accurate estimates of what is true in the population. In fact, survey research may be the only approach in psychology in which random sampling is routinely used. Beyond these two characteristics, almost anything goes in survey research. Surveys can be long or short. They can be conducted in person, by telephone, through the mail, or over the Internet. They can be about voting intentions, consumer preferences, social attitudes, health, or anything else that it is possible to ask people about and receive meaningful answers. Although survey data are often analyzed using statistics, there are many questions that lend themselves to more qualitative analysis.

Most survey research is non-experimental. It is used to describe single variables (e.g., the percentage of voters who prefer one presidential candidate or another, the prevalence of schizophrenia in the general population, etc.) and also to assess statistical relationships between variables (e.g., the relationship between income and health). But surveys can also be used within experimental research. The study by Lerner and her colleagues is a good example. Their use of self-report measures and a large national sample identifies their work as survey research. But their manipulation of an independent variable (anger vs. fear) to assess its effect on a dependent variable (risk judgments) also identifies their work as experimental.

History and Uses of Survey Research

Survey research may have its roots in English and American “social surveys” conducted around the turn of the 20th century by researchers and reformers who wanted to document the extent of social problems such as poverty (Converse, 1987) [1]. By the 1930s, the US government was conducting surveys to document economic and social conditions in the country. The need to draw conclusions about the entire population helped spur advances in sampling procedures. At about the same time, several researchers who had already made a name for themselves in market research, studying consumer preferences for American businesses, turned their attention to election polling. A watershed event was the presidential election of 1936 between Alf Landon and Franklin Roosevelt. A magazine called *Literary Digest* conducted a survey by sending ballots (which were also subscription requests) to millions of Americans. Based on this “straw poll,” the editors predicted that Landon would win in a landslide. At the same time, the new pollsters were using scientific methods with much smaller samples to predict just the opposite—that Roosevelt would win in a landslide. In fact, one of them, George Gallup, publicly criticized the methods of *Literary Digest* before the election and all but guaranteed that his prediction would be correct. And of course, it was, demonstrating the effectiveness of careful survey methodology (We will consider the reasons that Gallup was right later in this chapter). Gallup’s demonstration of the power of careful survey methods led later researchers to to local, and in 1948, the first national election survey by the Survey Research Center at the University of Michigan. This work eventually became the American National Election Studies (https://electionstudies.org/) as a collaboration of Stanford University and the University of Michigan, and these studies continue today.

From market research and election polling, survey research made its way into several academic fields, including political science, sociology, and public health—where it continues to be one of the primary approaches to collecting new data. Beginning in the 1930s, psychologists made important advances in questionnaire design, including techniques that are still used today, such as the Likert scale. (See “What Is a Likert Scale?” in Section 7.2 “Constructing Survey Questionnaires” .) 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 **4.3.1** 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 **4.3.1**: Some Lifetime Prevalence Results From the National Comorbidity Survey

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.**

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.

References

1. Converse, J. M. (1987). *Survey research in the United States: Roots and emergence, 1890–1960*. Berkeley, CA: University of California Press. ↵

This page titled 4.3: Introduction to Surveys is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **7.2: Overview of Survey Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

4.3.1: Constructing Surveys

Learning Objectives

1. Describe the cognitive processes involved in responding to a survey item.
2. Explain what a context effect is and give some examples.
3. Create a simple survey questionnaire based on principles of effective item writing and organization.

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.

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.

A Cognitive Model

Figure **4.3.1.1** presents a model of the cognitive processes that people engage in when responding to a survey item (Sudman, Bradburn, & Schwarz, 1996) [1]. 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 **4.3.1.1**: Model of the Cognitive Processes Involved in Responding to a Survey Item

Consider, for example, the following questionnaire item:

How many alcoholic drinks do you consume in a typical day?

- _____ a lot more than average
- _____ somewhat more than average
- _____ average
- _____ somewhat fewer than average
- _____ a lot fewer than average

Although this item at first seems straightforward, it poses several difficulties for respondents. First, they must interpret the question. For example, they must decide whether "alcoholic drinks" include beer and wine (as opposed to just hard liquor) and whether a "typical day" is a typical weekday, typical weekend day, or both . Even though Chang and Krosnick (2003)[2] found that asking about "typical" behavior has been shown to be more valid than asking about "past" behavior, their study compared "typical week" to "past week" and may be different when considering typical weekdays or weekend days) . Once respondents have interpreted the question, they must retrieve relevant information from memory to answer it. But what information should they retrieve, and how should they go about retrieving it? They might think vaguely about some recent occasions on which they drank alcohol, they might carefully try to recall and count the number of alcoholic drinks they consumed last week, or they might retrieve some existing beliefs that they have about themselves (e.g., "I am not much of a drinker"). Then they must use this information to arrive at a tentative judgment about how many alcoholic drinks they consume in a typical day. For example, this mental calculation might mean dividing the number of alcoholic drinks they consumed last week by seven to come up with an average number per day. Then they must format this tentative answer in terms of the response options actually provided. In this case, the options pose additional problems of interpretation. For example, what does "average" mean, and what would count as "somewhat more" than

average? Finally, they must decide whether they want to report the response they have come up with or whether they want to edit it in some way. For example, if they believe that they drink a lot more than average, they might not want to report that for fear of looking bad in the eyes of the researcher, so instead, they may opt to select the “somewhat more than average” response option.

From this perspective, what at first appears to be a simple matter of asking people how much they drink (and receiving a straightforward answer from them) turns out to be much more complex.

Context Effects on Survey Responses

Again, this complexity can lead to unintended influences on respondents’ answers. These are often referred to as **context effects** because they are not related to the content of the item but to the context in which the item appears (Schwarz & Strack, 1990) [3]. For example, there is an **item-order effect** when the order in which the items are presented affects people’s responses. One item can change how participants interpret a later item or change the information that they retrieve to respond to later items. For example, researcher Fritz Strack and his colleagues asked college students about both their general life satisfaction and their dating frequency (Strack, Martin, & Schwarz, 1988) [4]. When the life satisfaction item came first, the correlation between the two was only −.12, suggesting that the two variables are only weakly related. But when the dating frequency item came first, the correlation between the two was +.66, suggesting that those who date more have a strong tendency to be more satisfied with their lives. Reporting the dating frequency first made that information more accessible in memory so that they were more likely to base their life satisfaction rating on it.

The response options provided can also have unintended effects on people’s responses (Schwarz, 1999) [5]. For example, when people are asked how often they are “really irritated” and given response options ranging from “less than once a year” to “more than once a month,” they tend to think of major irritations and report being irritated infrequently. But when they are given response options ranging from “less than once a day” to “several times a month,” they tend to think of minor irritations and report being irritated frequently. People also tend to assume that middle response options represent what is normal or typical. So if they think of themselves as normal or typical, they tend to choose middle response options. For example, people are likely to report watching more television when the response options are centered on a middle option of 4 hours than when centered on a middle option of 2 hours. To mitigate against order effects, rotate questions and response items when there is no natural order. Counterbalancing or randomizing the order of presentation of the questions in online surveys are good practices for survey questions and can reduce response order effects that show that among undecided voters, the first candidate listed in a ballot receives a 2.5% boost simply by virtue of being listed first[6]!

Writing Survey Items

Types of Items

Questionnaire items can be either open-ended or closed-ended. **Open-ended items** simply ask a question and allow participants to answer in whatever way they choose. The following are examples of open-ended questionnaire items.

- “What is the most important thing to teach children to prepare them for life?”
- “Please describe a time when you were discriminated against because of your age.”
- “Is there anything else you would like to tell us about?”

Open-ended items are useful when researchers do not know how participants might respond or when they want to avoid influencing their responses. Open-ended items are more qualitative in nature, so they tend to be used when researchers have more vaguely defined research questions—often in the early stages of a research project. Open-ended items are relatively easy to write because there are no response options to worry about. However, they take more time and effort on the part of participants, and they are more difficult for the researcher to analy z e because the answers must be transcribed, coded, and submitted to some form of qualitative analysis, such as content analysis. Another disadvantage is that respondents are more likely to skip open-ended items because they take longer to answer. It is best to use open-ended questions when the answer is unsure or for quantities which can easily be converted to categories later in the analysis.

Closed-ended items ask a question and provide a set of response options for participants to choose from. The alcohol item just mentioned is an example, as are the following:

How old are you?

- _____ Under 18
- _____ 18 to 34
- _____ 35 to 49
- _____ 50 to 70
- _____ Over 70

On a scale of 0 (no pain at all) to 10 (worst pain ever experienced), how much pain are you in right now?

Have you ever in your adult life been depressed for a period of 2 weeks or more? Yes No

Closed-ended items are used when researchers have a good idea of the different responses that participants might make. They are more quantitative in nature, so they are also used when researchers are interested in a well-defined variable or construct such as participants' level of agreement with some statement, perceptions of risk, or frequency of a particular behavior. Closed-ended items are more difficult to write because they must include an appropriate set of response options. However, they are relatively quick and easy for participants to complete. They are also much easier for researchers to analyze because the responses can be easily converted to numbers and entered into a spreadsheet. For these reasons, closed-ended items are much more common.

All closed-ended items include a set of response options from which a participant must choose. For categorical variables like sex, race, or political party preference, the categories are usually listed and participants choose the one (or ones) to which they belong. For quantitative variables, a rating scale is typically provided. A **rating scale** is an ordered set of responses that participants must choose from. Figure **4.3.1.2** shows several examples. The number of response options on a typical rating scale ranges from three to 11—although five and seven are probably most common. Five-point scales are best for unipolar scales where only one construct is tested, such as frequency (Never, Rarely, Sometimes, Often, Always). Seven-point scales are best for bipolar scales where there is a dichotomous spectrum, such as liking (Like very much, Like somewhat, Like slightly, Neither like nor dislike, Dislike slightly, Dislike somewhat, Dislike very much). For bipolar questions, it is useful to offer an earlier question that branches them into an area of the scale; if asking about liking ice cream, first ask "Do you generally like or dislike ice cream?" Once the respondent chooses like or dislike, refine it by offering them relevant choices from the seven-point scale. Branching improves both reliability and validity (Krosnick & Berent, 1993)[7]. Although you often see scales with numerical labels, it is best to only present verbal labels to the respondents but convert them to numerical values in the analyses. Avoid partial labels or length or overly specific labels. In some cases, the verbal labels can be supplemented with (or even replaced by) meaningful graphics. The last rating scale shown in Figure 7.3 is a visual-analog scale, on which participants make a mark somewhere along the horizontal line to indicate the magnitude of their response.

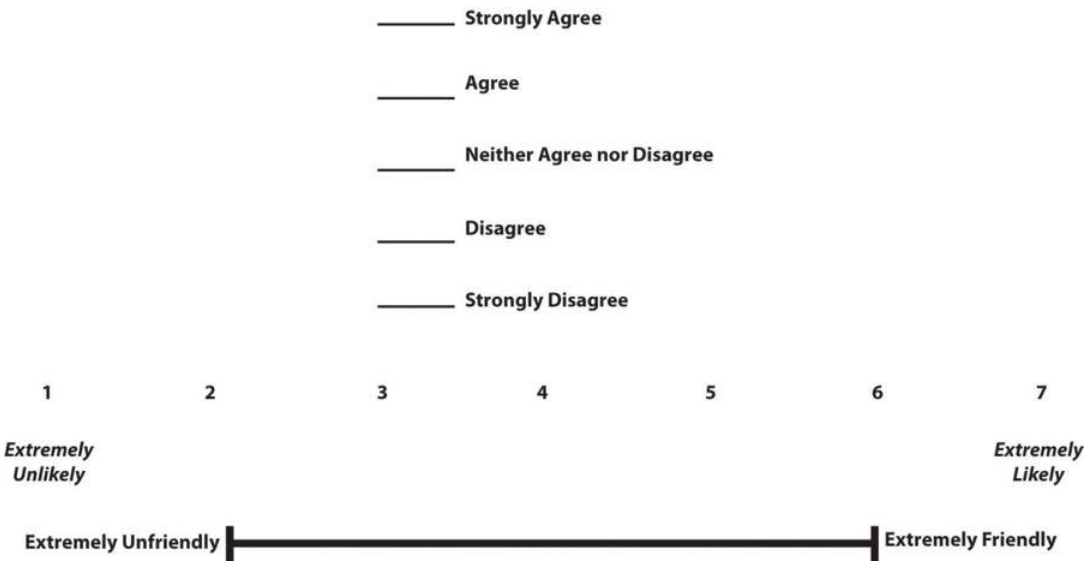


Figure **4.3.1.2**: Example Rating Scales for Closed-Ended Questionnaire Items

What Is a Likert Scale?

In reading about psychological research, you are likely to encounter the term *Likert scale* . Although this term is sometimes used to refer to almost any rating scale (e.g., a 0-to-10 life satisfaction scale), it has a much more precise meaning.

In the 1930s, researcher Rensis Likert (pronounced LICK-ert) created a new approach for measuring people's attitudes (Likert, 1932) [8]. It involves presenting people with several statements—including both favorable and unfavorable statements—about some person, group, or idea. Respondents then express their agreement or disagreement with each statement on a 5-point scale: *Strongly Agree* , *Agree* , *Neither Agree nor Disagree* , *Disagree* , *Strongly Disagree* . Numbers are assigned to each response a nd then summed across all items to produce a score representing the attitude toward the person, group, or idea. For items that are phrased in an opposite direction (e.g., negatively worded statements instead of positively worded statements), reverse coding is used so that the numerical scoring of statements also runs in the opposite direction. The entire set of items came to be called a Likert scale.

Thus unless you are measuring people's attitude toward something by assessing their level of agreement with several statements about it, it is best to avoid calling it a Likert scale. You are probably just using a "rating scale."

Writing Effective Items

We can now consider some principles of writing questionnaire items that minimize unintended context effects and maximize the reliability and validity of participants' responses. A rough guideline for writing questionnaire items is provided by the BRUSO model (Peterson, 2000)[9]. 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 barrelled." They ask about two conceptually separate issues but allow only one response. For example, "Please rate the extent to which you have been feeling anxious and depressed." This item should probably be split into two separate items—one about anxiety and one about depression. Finally, effective questionnaire items are *objective* in the sense that they do not reveal the researcher's own opinions or lead participants to answer in a particular way. Table **4.3.1.1** 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 **4.3.1.1**: BRUSO Model of Writing Effective Questionnaire Items, Plus Examples

Criterion	Poor	Effective
B—Brief	"Are you now or have you ever been the possessor of a firearm?"	"Have you ever owned a gun?"
R—Relevant	"What is your sexual orientation?"	Do not include this item unless it is clearly relevant to the research.
U—Unambiguous	"Are you a gun person?"	"Do you currently own a gun?"
S—Specific	"How much have you read about the new gun control measure and sales tax?"	"How much have you read about the new sales tax?"
O—Objective	"How much do you support the new gun control measure?"	"What is your view of the new gun control measure?"

For closed-ended items, it is also important to create an appropriate response scale. For categorical variables, the categories presented should generally be mutually exclusive and exhaustive. Mutually exclusive categories do not overlap. For a religion item,

for example, the categories of *Christian* and *Catholic* are not mutually exclusive but *Protestant* and *Catholic* are mutually exclusive. Exhaustive categories cover all possible responses. Although *Protestant* and *Catholic* are mutually exclusive, they are not exhaustive because there are many other religious categories that a respondent might select: *Jewish* , *Hindu* , *Buddhist* , and so on. In many cases, it is not feasible to include every possible category, in which case an *Other* category, with a space for the respondent to fill in a more specific response, is a good solution. If respondents could belong to more than one category (e.g., race), they should be instructed to choose all categories that apply.

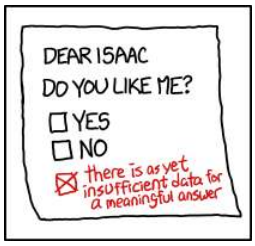
For rating scales, five or seven response options generally allow about as much precision as respondents are capable of. However, numerical scales with more options can sometimes be appropriate. For dimensions such as attractiveness, pain, and likelihood, a 0-to-10 scale will be familiar to many respondents and easy for them to use. Regardless of the number of response options, the most extreme ones should generally be "balanced" around a neutral or modal midpoint. An example of an unbalanced rating scale measuring perceived likelihood might look like this:

Unlikely | *Somewhat Likely* | *Likely* | *Very Likely* | *Extremely Likely*

A balanced version might look like this:

Extremely Unlikely | *Somewhat Unlikely* | *As Likely as Not* | *Somewhat Likely* | *Extremely Likely*

Note, however, that a middle or neutral response option does not have to be included. Researchers sometimes choose to leave it out because they want to encourage respondents to think more deeply about their response and not simply choose the middle option by default. However, including middle alternatives on bipolar dimensions can be used to allow people to choose an option that is neither.



"Question" retrieved from http://imgs.xkcd.com/comics/question.png (CC-BY-NC 2.5)

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) [10]. One is to encourage respondents to participate in the survey. In many types of research, such encouragement is not necessary either because participants do not know they are in a study (as in naturalistic observation) or because they are part of a subject pool and have already shown their willingness to participate by signing up and showing up for the study. Survey research usually catches respondents by surprise when they answer their phone, go to their mailbox, or check their e-mail—and the researcher must make a good case for why they should agree to participate. Thus the introduction should briefly explain the purpose of the survey and its importance, provide information about the sponsor of the survey (university-based surveys tend to generate higher response rates), acknowledge the importance of the respondent's participation, and describe any incentives for participating.

The second function of the introduction is to establish informed consent. Remember that this involves describing to respondents everything that might affect their decision to participate. This includes the topics covered by the survey, the amount of time it is likely to take, the respondent's option to withdraw at any time, confidentiality issues, and so on. Written consent forms are not always used in survey research (when the research is of minimal risk and completion of the survey instrument is often accepted by the IRB as evidence of consent to participate), so it is important that this part of the introduction be well documented and presented clearly and in its entirety to every respondent.

The introduction should be followed by the substantive questionnaire items. But first, it is important to present clear instructions for completing the questionnaire, including examples of how to use any unusual response scales. Remember that the introduction is the point at which respondents are usually most interested and least fatigued, so it is good practice to start with the most important items for purposes of the research and proceed to less important items. Items should also be grouped by topic or by type. For

example, items using the same rating scale (e.g., a 5-point agreement scale) should be grouped together if possible to make things faster and easier for respondents. Demographic items are often presented last because they are least interesting to participants but also easy to answer in the event respondents have become tired or bored. Of course, any survey should end with an expression of appreciation to the respondent.

References

1. Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass. ↵
2. Chang, L., & Krosnick, J.A. (2003). Measuring the frequency of regular behaviors: Comparing the 'typical week' to the 'past week'. *Sociological Methodology, 33*, 55-80. ↵
3. 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). Chichester, UK: Wiley. ↵
4. Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: The social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology, 18*, 429–442. ↵
5. Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93–105. ↵
6. Miller, J.M. & Krosnick, J.A. (1998). The impact of candidate name order on election outcomes. *Public Opinion Quarterly, 62*(3), 291-330. ↵
7. Krosnick, J.A. & Berent, M.K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science*, *27*(3), 941-964. ↵
8. Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology,140*, 1–55. ↵
9. Peterson, R. A. (2000). *Constructing effective questionnaires*. Thousand Oaks, CA: Sage. ↵
10. Peterson, R. A. (2000). *Constructing effective questionnaires*. Thousand Oaks, CA: Sage. ↵

This page titled 4.3.1: Constructing Surveys is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **7.3: Constructing Surveys** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

4.3.2: Reliability and Validity of Measurement

Learning Objectives

1. Define reliability, including the different types and how they are assessed.
2. Define validity, including the different types and how they are assessed.

Again, 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 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. Researchers 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: reliability and validity. This example is about a scale, but most research on reliability and validity are assessing the reliability of surveys (are the results consistent?) and the validity of surveys (are the results accurate?).

Reliability

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

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. This is typically done by graphing the data in a scatterplot and computing the correlation coefficient. Figure **4.3.2.1** shows the correlation between two sets of scores of several university students on the Rosenberg Self-Esteem Scale, administered two times, a week apart. The correlation coefficient for these data is +.95. In general, a test-retest correlation of +.80 or greater is considered to indicate good reliability.

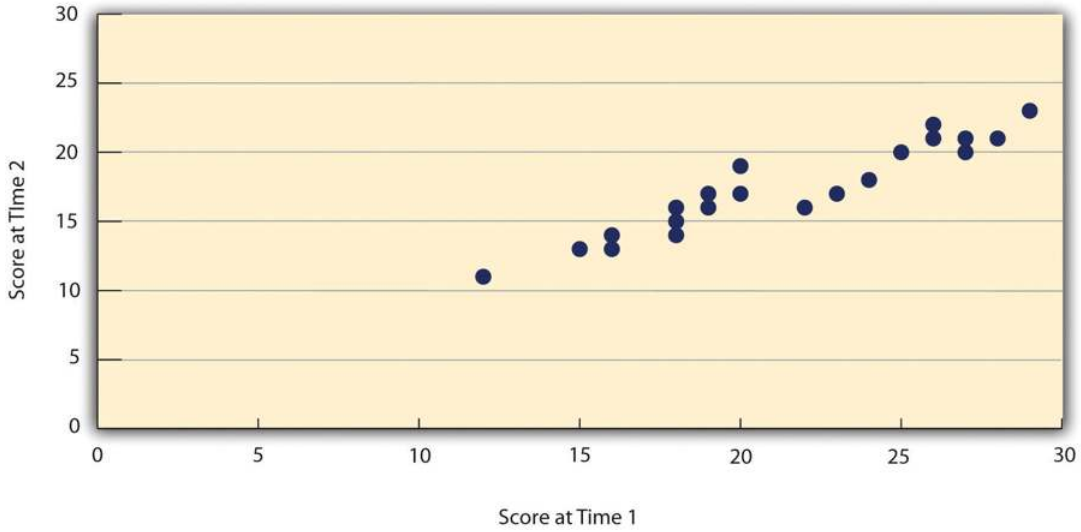


Figure **4.3.2.1**: Test-Retest Correlation Between Two Sets of Scores of Several College Students on the Rosenberg Self-Esteem Scale, Given Two Times a Week Apart

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.

Internal Consistency

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

Like test-retest reliability, internal consistency can only be assessed by collecting and analyzing data. One approach is to look at a split-half correlation. This involves splitting the items into two sets, such as the first and second halves of the items or the even- and odd-numbered items. Then a score is computed for each set of items, and the relationship between the two sets of scores is examined. For example, Figure **4.3.2.2** shows the split-half correlation between several university students' scores on the even-numbered items and their scores on the odd-numbered items of the Rosenberg Self-Esteem Scale. The correlation coefficient for these data is +.88. A split-half correlation of +.80 or greater is generally considered good internal consistency.

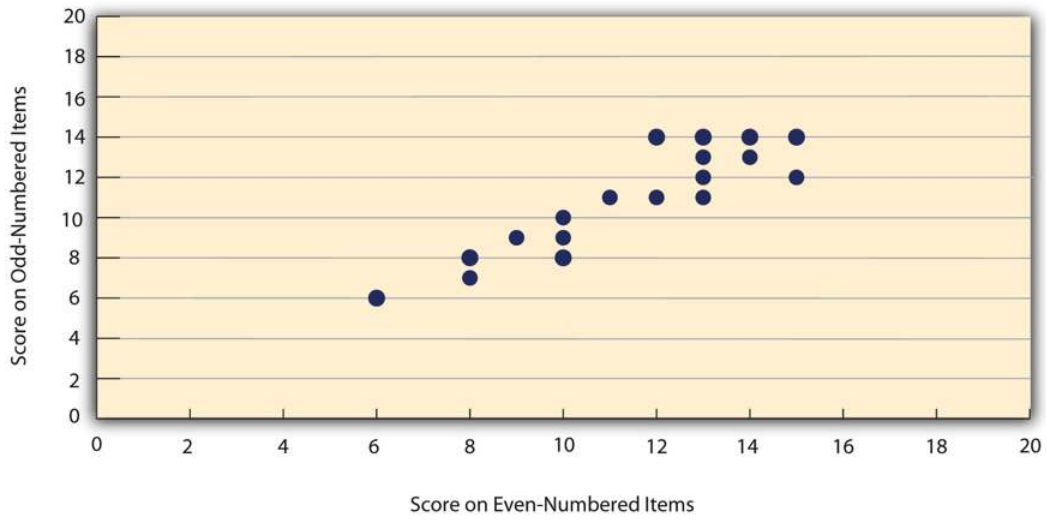


Figure **4.3.2.2**: Split-Half Correlation Between Several College Students' Scores on the Even-Numbered Items and Their Scores on the Odd-Numbered Items of the Rosenberg Self-Esteem Scale

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

Interrater 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 university students' social skills, you could make video recordings of them as they interacted with another student whom they are meeting for the first time. Then you could have two or more observers watch the videos and rate each student's level of social skills. To the extent that each participant does, in fact, have some level of social skills that can be detected by an attentive observer, different observers' ratings should be highly correlated with each other. Inter-rater reliability would also have been measured in Bandura's Bobo doll study. In this case, the observers' ratings of how many acts of aggression a particular child committed while playing with the Bobo doll should have been highly positively correlated. Interrater reliability is often assessed using Cronbach's α when the judgments are quantitative or an analogous statistic called Cohen's κ (the Greek letter kappa) when they are categorical.

Construct Validity

You've learned about four different types of validity in section 4.2.1. For surveys, we focus on construct validity. Construct 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.

Discussions of validity usually divide it into several distinct "types." But a good way to interpret these types is that they are other kinds of evidence—in addition to reliability—that should be taken into account when judging the validity of a measure. Here we consider three basic kinds: face validity, content validity, and criterion validity.

Face Validity

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

Face validity is at best a very weak kind of evidence that a measurement method is measuring what it is supposed to. One reason is that it is based on people’s intuitions about human behavior, which are frequently wrong. It is also the case that many established measures in psychology work quite well despite lacking face validity. The Minnesota Multiphasic Personality Inventory-2 (MMPI-2) measures many personality characteristics and disorders by having people decide whether each of over 567 different statements applies to them—where many of the statements do not have any obvious relationship to the construct that they measure. For example, the items “I enjoy detective or mystery stories” and “The sight of blood doesn’t frighten me or make me sick” both measure the suppression of aggression. In this case, it is not the participants’ literal answers to these questions that are of interest, but rather whether the pattern of the participants’ responses to a series of questions matches those of individuals who tend to suppress their aggression.

Content Validity

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

Criterion Validity

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 clarity and simplicity, think about criteria as behaviors that can be measured, rather than information from a different survey (which would be convergent validity). 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 behavioral variable that one has reason to think should be related with the construct being measured, and there will usually be many of them. For example, one would expect test anxiety scores to be negatively related with exam performance and course grades and positively related 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 related 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).

Convergent Validity

When the criteria is a different survey of a similar topic, we call this convergent validity. For example, one would expect new measures of test anxiety or physical risk taking to be positively correlated with existing established measures of the same constructs. Assessing convergent 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 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).

Discriminant Validity

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

When they created the Need for Cognition Scale, Cacioppo and Petty (1982) 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.

References

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

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.

This page titled 4.3.2: Reliability and Validity of Measurement is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **4.3: Reliability and Validity of Measurement** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

4.4: Interviews and Focus Groups

Learning Objectives

1. Describe the strengths and weaknesses of using interviews to gather self-report information.
2. Describe the strengths and weaknesses of using focus groups to gather self-report information.

Surveys are a common way to gather self-report information from participants, but it is not the only way. In this section, we will discuss interviews and focus groups.

Interviews

For research purposes, interviews are usually a set of questions that are asked in real time (either in-person or virtually) by interviewer to an interviewee (respondent). Interviews can be structured or unstructured. In structured interviews, there is a strict script that the interviewer does not deviate from; each respondent is asked the same questions in the same order. In contrast, unstructured interviews consist of a small number of general questions or prompts that allow participants to talk about what is of interest to them. 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. In their research on the families of suicide victims, Linquist et al. (2008) were aware that how much was disclosed about such a sensitive topic should be led by the families, not by the researchers, so they chose unstructured interviews.

Exercise 4.4.1

In comparison to surveys, what are some benefits to interviews?

Answer

Some benefits to interviews, in comparison to surveys, are that:

- The interviewer can more easily answer participant questions than when completing a survey.
- The interviewer can see if the respondent looks confused or uncomfortable.
- The interviewer can build rapport, meaning a harmonious relationship, which may help the respondent answer sensitive questions.

Did you come up with any other benefits that weren't listed in the answer to Exercise 4.4.1? In Exercise 4.4.2, you'll be asked to consider some downsides to gathering self-report information with interviews.

Exercise 4.4.2

In comparison to surveys, what are some disadvantages to interviews?

Answer

Some disadvantages to using interviews rather than surveys could include:

- Interviews usually take longer than surveys (hours rather than minutes).
- It may be more difficult to get participants to volunteer for an interview compared to completing a survey.
- Interviews are more costly. This is partly because getting participants to volunteer for an interview can take more time and money than for surveys, but also because good interviewers are trained and cost more than hosting or printing a survey.

Did you come up with any other disadvantages that weren't listed in the answer to Exercise 4.4.2?

Focus Groups

When small groups of people who participate together in interviews focused on a particular topic or issue, they are called focus groups. The interaction among participants in a focus group can sometimes bring out more information than can be learned in a one-on-one interview. The use of focus groups has become a standard technique in business and industry among those who want to understand consumer tastes and preferences. The content of all focus group interviews is usually recorded and transcribed to facilitate later analyses.

? Exercise 4.4.3

In comparison to surveys and interviews, what are some benefits to focus groups?

Answer

Some benefits to focus groups are that:

- The other focus group members may spark ideas or lead to important topics that the researcher and no one individual participant may have come up with.
- In comparison to interviews, information from more participants is gathered in a shorter amount of time.

Did you come up with any other benefits that weren't listed in the answer to Exercise **4.4.3**? 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. In Exercise **4.4.4**, you'll be asked to consider some downsides to gathering self-report information with interviews.

? Exercise 4.4.4

In comparison to surveys and interviews, what are some disadvantages to focus groups?

Answer

Some disadvantages to using focus groups could include:

- 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.
- Focus groups may work better for extraverts, and may not work well for those who are shy, socially anxious, or paranoid.
- Like interviews but not surveys, focus groups must have a quality facilitator to ensure that no one participant dominates, and that everyone feels comfortable sharing their opinions and experiences.

Did you come up with any other disadvantages that weren't listed in the answer to Exercise **4.4.4**? If you are considering running your own focus groups, there are some tips from Esposito and Evans-Winters's (2021) book Introduction to Intersectional Qualitative Research in the box:

Tips for designing Focus groups (Esposito & Evans-Winters, 2021)

- Ideal size: 6-8 participants
- Establishing personalized contact with participants is an essential step for the facilitator
  - Personalize communication/invitations
  - Explain why you want them to come
  - Incentives/food, etc.
  - Helps to begin the process of establishing trust
- Stay alert and facilitate the space (ensure everyone has a chance to talk/no one dominates, shorter open-ended questions, time limits)
- At the end, have the group clarify or summarize discussions

Life video-chatting is a newer technology, so there is not a lot of research yet comparing in-person interviews or focus groups to synchronous video chats. Maybe you can research these topics!

References

Esposito, J. & Evans-Winters, V. (2021). *Introduction to intersectional qualitative research*. SAGE Publications.

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*, 26. Retrieved from http://www.biomedcentral.com/1471-244X/8/2

This page titled 4.4: Interviews and Focus Groups is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **6.5: Qualitative Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

4.5: Measuring Behavior through Biology

This page is a draft and under active development. Please forward any questions, comments, and/or feedback to the ASCCC OERI (oeri@asccc.org).

Learning Objectives

1. Identify some biological measures that reflect thoughts or feelings.

This section briefly describes a few ways that biological measures, such as of measures of the brain, breathing and heart rate, or hormone levels, can be used in research on behavior.

Measuring the Brain

There are many ways to categorize the techniques that are used when studying the brain, but neuroscientists tend to differentiate between non-invasive techniques versus invasive techniques. As social science researchers, very few will use invasive techniques so they will not be discussed here. Instead, we cover some non-invasive techniques in which we are able to study the brain without getting direct physical access to the brain (think of fixing a broken pipe in a wall without having to open the wall up). Techniques have been developed that allow us to see what the brain looks like, as a still image (structurally) or in action (functionally). While not reviewed here, following ethical guidelines for both psychological and medical research should be followed when using physiological measures of thoughts, feelings, or behavior; these are sometime called psychobiological measures.

Structural Imaging

Structural imaging include a variety of tools that take pictures (images) of the physical features of the brain. These techniques are useful in many situations such as locating tumors, sites of physical brain damage, or finding size differences between the structures of the brain between various groups. Magnetic resonance imaging (MRI), for example, is one such technique that is commonly used to study the brain and to diagnosis knee and shoulder injuries. Structural imaging techniques allow us to look inside the brain (or body) without having to go inside.

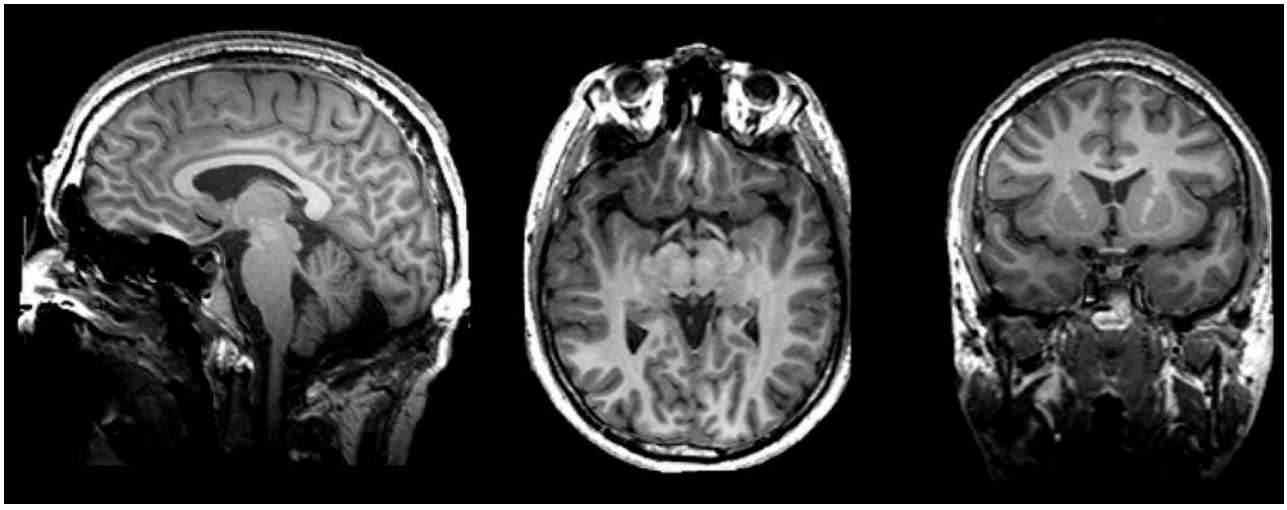


Figure 4.5.1: An MRI (A structural image) of the brain. Note the ability to see specific gyri (bumps) and sulci (grooves).

A series of MRI images can be used to create a picture of the brain.

Functional Imaging

Many researchers are also interested in how the brain works. Some studies begin with the scientific question of “what does this part do?” Or more commonly, “Where in the brain does this happen?” Functional imaging techniques allow researchers to learn about the brain activity during various tasks by creating images based on the electrical activity or the absorption of various substances

that occurs while a subject is engaging in a task. Such techniques can be used, for example, to visualize the parts of the brain that respond when we're exposed to stimuli that upset us or make us happy.

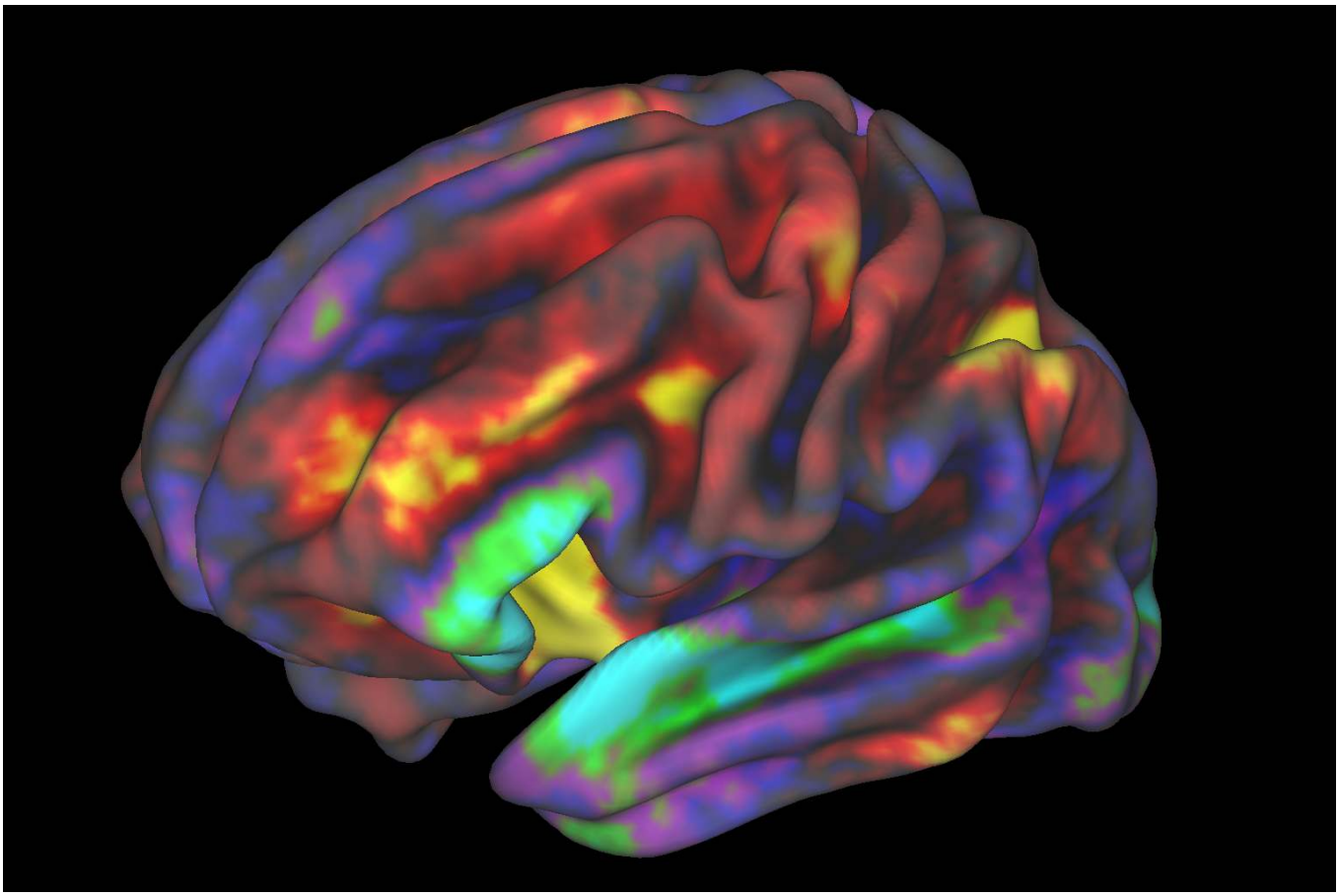


Figure **4.5.2**: Functional MRI of a teenage brain conducting a working memory test. The regions in red and yellow, primarily in the frontal lobe and occipital lobe, are the most active.

Biological Chemicals

Another way to measure the brain is through neurotransmitters. Neurotransmitters one way that the brain communicates with itself and the rest of the nervous system. Table **4.5.1** (CC-BY OpenStax) shows a sampling of how the neurotransmitters may potentially affect behavior.

Table **4.5.1**: Major Neurotransmitters and How They Affect Behavior (CC-BY OpenStax)

Neurotransmitter	Involved in	Potential Effect on Behavior
Acetylcholine	Muscle action, memory	Increased arousal, enhanced cognition
Beta-endorphin	Pain, pleasure	Decreased anxiety, decreased tension
Dopamine	Mood, sleep, learning	Increased pleasure, suppressed appetite
Gamma-aminobutyric acid (GABA)	Brain function, sleep	Decreased anxiety, decreased tension
Glutamate	Memory, learning	Increased learning, enhanced memory
Norepinephrine	Heart, intestines, alertness	Increased arousal, suppressed appetite
Serotonin	Mood, sleep	Modulated mood, suppressed appetite

In addition to brain chemistry, a more long-lasting way that chemicals affect behavior is through hormones. According to Williams Jones (2025), hormones can be measured as a way to chemically measure potential behaviors related to aggression, stress, bonding, or sexual behavior. For example, you could use a survey or interview to ask respondents how stressed they feel, or you could measure their levels of cortisol.

Additional Biological Measures of Behavior

Other typical psychobiological measures include pupil dilation, heart rate, blood pressure, galvanic skin response to measure mood. If galvanic skin response (GSR) is new to you, it's a measurement of how the sweat on your skin conducts electricity. All of these biological measures are affected by physiological arousal, such as when someone is upset or excited. If changes are shown in these measures to different stimuli, then we can assume that the stimulus affected mood somehow.

Advantages and Disadvantages of Psychobiological Measures

What do you think?

What do you think are some advantages of using psychobiological measurements to assess thoughts, feelings, or behaviors? What could be some disadvantages?

What benefits of using biological measurement to assess psychological phenomena did you come up with? One advantage of these types of measures is that they are objective, rather than self-report. Rather than have someone complete cognitive tests, we could see their brain's structure or ask them. Rather than ask which picture is preferred, the pictures can be shown while participants are in an fMRI to see what parts of their brain are most active.Rather than have someone fill out a self-report depression inventory, their dopamine and serotonin neurotransmitters could be measured.

What disadvantages did you think of? After reading this section, you may have noticed there is some inference involved in using biological measures to measure psychological concepts. We know that cortisol is related to stress, but there could be physiological or emotional reasons for the increase in cortisol and we can't know which. Is the person fighting an infection, or are they worrying about a future exam? Similarly, while there be a correlation between scores on certain cognitive test and structures in the brain, we don't know which came first. This is the classic "correlation versus causation" or "What about Z?" issue; maybe something caused both structural abnormalities and scores on the cognitive test (such as childhood malnutrition).

With just about everything related to designing a research study, it is up to the researcher to decide what is most appropriate for answering their research question, including how to measure the variables. As you read research articles, you should also be critically thinking about the choices that the researchers made; were their choices the best to answer their research question, or the maybe just the most convenient? Maybe psychobiological measures might be best, but MRIs are expensive and complex, so maybe a cognitive test was just simpler to use to complete the study. Or maybe we want to measure actual behavior, not correlates or proxies of behavior, so psychobiological measures aren't the best option. As you continue learning about designing studies, you'll see that there are many ways to collect data; you are encouraged to consider many options and decide on your own which option may be best for each research question.

References

Williams Jones, E. (2025). *Psychobiology: What is it and what does this science study?* Psychology for Mental Health. https://psychologyfor.com/psychobiol...science-study/

Attributions

- Psychophysiological Methods in Neuroscience by Zachary Infantolino and Gregory A. Miller is licensed under a CC BY-NC-SA 4.0 International License.
- Figure 4.5.1: "Deadstar's brain" by Daniele Oberti is licensed under CC BY-NC-ND 2.0
- Figure 4.5.2: "fMRI Image of Preteen Brain" by NIH Image Gallery is licensed under CC BY-NC 2.0
- Table 4.5.1: Major Neurotransmitters and How They Affect Behavior by OpenStax licensed under CC-BY 4.0

This page titled 4.5: Measuring Behavior through Biology is shared under a mixed license and was authored, remixed, and/or curated by Open Education Resource Initative (ASCCC Open Educational Resources Initiative (OERI)) .

4.6: Observations

Learning Objectives

1. List the various types of observational research methods and distinguish between each.
2. Describe the strengths and weakness of each observational research method.

Observational Research

Generally, the goal of observational research is to obtain a snapshot of specific characteristics of an individual, group, or setting. There are several decisions that a researcher must make before starting to actually observe participants. These related to how structured the situation will be, how aware the participants will be of the observation, and how involved the observer will be in the situation. Each of these will be discussed below.

Observation Structure

The setting of observations can range from naturalistic to structured. Naturalistic observation is an observational method that involves observing people's behavior in the environment in which it typically occurs, which makes this data collection method is a type of field research (as opposed to a type of laboratory research). Jane Goodall's famous research on chimpanzees is a classic example of naturalistic observation. Dr. Goodall spent three decades observing chimpanzees in their natural environment in East Africa. She examined such things as chimpanzee's social structure, mating patterns, gender roles, family structure, and care of offspring by observing them in the wild. However, naturalistic observation could more simply involve observing shoppers in a grocery store, children on a school playground, or psychiatric inpatients in their wards.

In structured observations**,** the investigator still makes careful observations of one or more specific behaviors in a particular setting. However, instead of a typical environment with natural behaviors, the setting is structured in some way. The researcher may observe people in the laboratory environment, or in a natural setting (like a classroom setting) that the researcher has structured some way. For example, a researcher could have a confederate drop their backpack in front of the library as the confederate passes a potential participant; this would be a natural setting. Alternatively, the researcher could observe participants as they sit in classroom laboratory while the overhead lights are slowly dimmed to see when (or if!) the participant would get up to check things out. One of the primary benefits of structured observation is that it is far more efficient than naturalistic observation. The researchers are focused on specific behaviors, and often times the environment is structured to encourage the behaviors of interest which again means that researchers do not have to invest as much time in waiting for the behaviors of interest to naturally occur. Finally, researchers using this approach can clearly exert greater control over the environment. However, when researchers exert more control over the environment it may make the environment less natural which decreases external validity; in other words, it is less clear for instance whether structured observations made in a laboratory environment will generalize to a real world environment.

Participant Awareness

Researchers engaged in 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 observation. Ethically, this method of data collection without informed consent is considered to be acceptable if the participants remain anonymous and the behavior occurs in a public setting where people would not normally have an expectation of privacy. Grocery shoppers putting items into their shopping carts, for example, are engaged in public behavior that is easily observable by store employees and other shoppers. For this reason, most researchers would consider it ethically acceptable to observe them for a study. On the other hand, one of the arguments against the ethicality of the naturalistic observation of bathroom behavior, like hand-washing or talking on the phone in a stall, is that people have a reasonable expectation of privacy even in a public restroom and that this expectation was violated.

Alternatively, researchers can conduct undisguised observation where the participants are made aware of the researcher presence and monitoring of their behavior. However, one concern with undisguised naturalistic observation is reactivity. Reactivity, sometimes referred to as the Hawthorne Effect, refers to when a measure changes participants' behavior. In the case of undisguised observation, participants may act differently than they normally would. For instance, participants being observed at the gym may

act much differently, maybe lifting more weight or running faster or longer than they normally would, if the participants knew that someone was observing you and recording your behaviors. This would make the observations less valid since we can't be sure how different the participants' behavior was. 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.

Naturalistic observation can easily accommodate either disguised or undisguised observations. When observing adults play basketball on an outdoor court, you could sit on a nearby bench and watch unobtrusively (disguised), or you could sit on a share near the court and let them know that you're keeping statistics on their playing. Structured observations could include undisguised observations, but it makes more sense to be disguised so that participants aren't reactive to everything that could be an intervention. For example, people probably wouldn't act normally if you said that you were observing people to how helpful they were, and then dropped your backpack.

Observer Involvement

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. 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.

Just how naturalistic and structured observations could have participant aware of being observed (undisguised observation) or unaware (disguised observation), participant observation can also be disguised or not. Rosenhan's study (1973) of the experience of people in a psychiatric ward would be considered disguised participant observation because Rosenhan and his pseudopatients were admitted into psychiatric hospitals on the pretense of being patients so that they could observe the way that psychiatric patients are treated by staff. The staff and other patients were unaware of their true identities as researchers. In another famous example of disguised participant observation, Leon Festinger and his colleagues infiltrated a doomsday cult known as the Seekers, whose members believed that the apocalypse would occur on December 21, 1954. Interested in studying how members of the group would cope psychologically when the prophecy inevitably failed, they carefully recorded the events and reactions of the cult members in the days before and after the supposed end of the world. Unsurprisingly, the cult members did not give up their belief but instead convinced themselves that it was their faith and efforts that saved the world from destruction. Festinger and his colleagues later published a book about this experience, which they used to illustrate the theory of cognitive dissonance (Festinger et al., 1956). Once again there are important ethical issues to consider with disguised participant observation. First no informed consent can be obtained and second deception is being used. The researcher is deceiving the participants by intentionally withholding information about their motivations for being a part of the social group they are studying. But sometimes disguised participation is the only way to access a protective group (like a cult). Further, disguised participant observation is less prone to reactivity than undisguised participant observation.

In undisguised participant observation, the researchers become a part of the group they are studying but they disclose their true identity as researchers to the group under investigation. An example of undisguised participant observation comes from a study by sociologist Amy Wilkins on a university-based religious organization that emphasized how happy its members were (Wilkins, 2008). Wilkins spent 12 months attending and participating in the group's meetings and social events, and she interviewed several group members. In her study, Wilkins identified several ways in which the group "enforced" happiness—for example, by continually talking about happiness, discouraging the expression of negative emotions, and using happiness as a way to distinguish themselves from other groups.

One of the primary benefits of participant observation is that the researchers are in a much better position to understand the viewpoint and experiences of the people they are studying when they are a part of the social group. The primary limitation with this approach is that the mere presence of the observer could affect the behavior of the people being observed. While this is also a concern with all types of observation, additional concerns arise when researchers become active members of the social group they are studying because that they may change the social dynamics and/or influence the behavior of the people they are studying. Similarly, if the researcher acts as a participant observer there can be concerns with biases resulting from developing relationships with the participants. Concretely, the researcher may become less objective resulting in more experimenter bias.

Collecting Data

In addition to reactivity, one of the difficulties of doing observational research is accurately and thoroughly collecting the data. To do this, researchers clearly define a set of target behaviors, 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.

There are few things that observers can do to help with this.

First, specific operational definitions are important. If you would like to observe how many people smile at a baby, you have to precisely define what a smile is. Does a smile include showing teeth? Is it any upturning of the mouth? Must the eyes get a little crinkly? How crinkly? You can see how difficult this might end up being! Practicing (with videos or a pilot test) can help with these issues as the target behaviors must be defined in such a way that different observers will code them in the same way. This difficulty with coding illustrates the issue of interrater reliability. Researchers are expected to demonstrate the interrater reliability of their coding procedure by having multiple raters code the same behaviors independently and then showing that the different observers are in close agreement.

As an example of operational definitions of observed behavior, Kraut and Johnston (1979) wanted to study bowlers' reactions to their shots, both when they were facing the pins and then when they turned toward their companions. 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). Kraut and Johnston video recorded a subset of their participants' reactions and had two observers independently code them. The two observers showed that they agreed on the reactions that were exhibited 97% of the time, indicating good interrater reliability. 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.

Second, limiting your observations might be necessary. You can limit who you observe; rather than observing everyone in a restaurant, you could observe one person or one table. You could limit what behaviors you are observing; rather than observing every behavior that is happening in the restaurant (even at that one table), you could focus on only smiling or only interactions with the baby. Finally, you can limit your observations by time; rather than observing for an hour, you could observe for two minutes then take a five minute break; you would repeat that cycle so that you end up observing for an hour, but you don't have to keep focus non-stop. As you might have guessed, these options can be combined. If you tried to observe a whole restaurant, every interaction with babies, for an hour, you'd probably miss a lot of what's going on. Instead, you could record every interaction with one baby for 1-minute intervals, with a break of 3 minutes, for the whole time that the baby is at the restaurant.

References

Festinger, L., Riecken, H., & Schachter, S. (1956). When prophecy fails: A social and psychological study of a modern group that predicted the destruction of the world. University of Minnesota Press.

Kraut, R. E., & Johnston, R. E. (1979). Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology*, *37*, 1539–1553.

Rosenhan, D. L. (1973). On being sane in insane places. *Science*, *179*, 250–258.

Wilkins, A. (2008). "Happier than Non-Christians": Collective emotions and symbolic boundaries among evangelical Christians. *Social Psychology Quarterly*, *71*, 281–301.

This page titled 4.6: Observations is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **6.6: Observational Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

4.7: Unobtrusive Research

Learning Objectives

- Identify some general strengths and weakness of self-report measures.
- List types of unobtrusive measures of behavior.
- Identify some general strengths and weakness of unobtrusive research.
- Apply how the Hawthorne Effect could affect self-report measures.

Alternatives to Self-Report Measures

We've spent a lot of time talking about gathering self-report information. Participants often try to be honest and accurate in their self-reports, but sometimes that's not possible: when people do not have the *ability* to report on their behavior and when people do not have the *willingness* to report on their own behavior.

When might participants not have the ability to report on themselves? The common reasons are:

- Fatigue
- Vocabulary
- Knowledge (of self, motivations)

Let's unpack those ideas. First, fatigue. While it is true that sometimes participants may be too sleep deprived to answers questions about themselves accurately, fatigue tends to affect self-report measures when the measures are really long. Many careers require applicants to take a personality test. The Minnesota Multiphasic Personality Inventory-2 (MMPI-2, Butcher et al., 1989) is a common assessment. This personality survey has 567 questions, and takes respondents about an hour to an hour and a half to complete. Imagine how mentally and emotionally tired you would be after having answers 560 questions about yourself over the course of an hour or more. I bet at that point, your self-report may not be as accurate as it was it was when you first started the survey.

Second, vocabulary. Many times, researchers ask questions of respondents using the technical language of the field of study. Respondents will try to answer honestly and accurately, but they might not know the answers because they do not know what some of the words in the questions mean. Third, self-knowledge. Imagine that a researcher asked how often you had worked out in the last week. You may think about your goal of working out three days a week, so you assume that you worked out three times in the past week. However, if the researcher checked your activity on a wearable device like a smart watch, they might notice that your heart rate and movement indicated working out only two times that week. It might not be that you intentionally provided false information, but that you didn't take the time to provide accurate information. Additionally, people explain why they behave certain ways even when they do not actually know why that do something. This TED Talk (Johansson, 2016) shows how people explain why they made a choice, even when that was not actually the choice that they made.

Finally, some people do not have the willingness to report accurately. This usually occurs when you are motivated to look good (social desirability). To avoid some of this pitfalls in self-report, we will consider ways unobtrusively measure behavior. Unobtrusive research refers to methods of collecting data that don’t interfere with the participants under study.

Unobtrusive Research

Researchers who seek evidence of what people actually *do*, as opposed to what they say they do (as in survey and interview research), might wish to consider using unobtrusive methods. The unobtrusive methods that will be discussed in the follow sub-sections are:

- Physical Trace
- Archival Data

Some researchers treat observation as unobtrusive measures, and sometimes it is. But observational research can manipulate variables (think structured observations), and sometimes are undisguised. Because of the variety of observational methodologies, this type of measurement was described separately from these other unobtrusive methods.

Benefits of Unobtrusive Research

Research participants, either intentionally or unintentionally, altering their behaviors because they know they are being studied, known as the Hawthorne effect, is not a concern for unobtrusive researchers because the participants do not know that they are being studied. While unobtrusive research projects, like all research projects, face the risk of introducing researcher bias into the work, researchers employing unobtrusive methods do not need to be concerned about the effect of the research *on their participants*. In fact, this is one of the major strengths of unobtrusive research.

Another benefit of unobtrusive research is that it *can* be relatively low-cost compared to some of the other methods we've discussed. Researchers may be able to access data without having to worry about paying participants for their time (though certainly travel to or access to some documents and archives can be costly, as well as the time it takes to make observations).

A benefit of archival data is that it is well suited to studies that focus on processes that occur over time. Archival data can provide accurate information about examine processes that occurred decades before data collection began, and are a cost-effective way to examine long-ranging processes. Unobtrusive methods do not rely on retrospective accounts, which may be subject to errors in memory, as some longitudinal surveys do.

Weaknesses of Unobtrusive Research

While there are many benefits to unobtrusive research, this method also comes with a unique set of drawbacks. Because unobtrusive researchers analyze data that may have been created or gathered for purposes entirely different from the researcher's aim, problems of validity sometimes arise in such projects. It may also be the case that data sources measuring whatever a researcher wishes to examine simply do not exist. This means that unobtrusive researchers may be forced to tweak their original research interests or questions to better suit the data that are available to them.

This page titled 4.7: Unobtrusive Research is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **11.2: Pros and Cons of Unobtrusive Research** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/sociology-comprehensive-edition.

4.7.1: What Is Left Behind (Physical Trace)

Learning Objectives

- Differentiate between types of physical traces.
- Identify potential physical traces that might be relevant for your research questions.

Another way to gather data unobtrusively is to collect what is left behind, and use that to make judgments about what behaviors occurred. In social science research, the remnants of behavior are called physical traces. According to Wikipedia (2024), these remnants could be any number of items, and are usually divided into two main categories.

Use traces indicate the use or non-use of an item. Fingerprints, for example, fall into the category of use traces, along with candy wrappers, cigarette cartons, and countless other objects (Wikipedia, 2024). Other examples of physical traces of behavior include worn paths across campus, the materials in a landfill or in someone's trash can (Rathje, 1992; Rathje & Murthy, 1992), indentations in furniture, or empty shelves in the grocery store.

In contrast, *products*, sometimes called material artifacts (especially by sociologists), are the creations or artifacts of behavior. An example of a product might be a painting, a song, a dance or television (Wikipedia, 2024). Other examples include video games and video game equipment, sculptures, mementos left on gravestones, housing structures, or even kitchen utensils. Whereas use physical traces tell us more about the behavior of an individual, products speak more to contemporary cultural themes (Wikipedia, 2024) by indicating tell us something about their beliefs, values, or norms.

In the National Museum of American History in Washington, DC. there was an exhibit displaying chef Julia Child's home kitchen, where she filmed many of her famous cooking shows. This might prompt a researcher to wonder how cooking has changed over the past few decades since Child's shows were on air. For example, how the layout of our kitchens and the utensils and appliances they contain might influence how we entertain guests, how much time we spend preparing meals, and how much time we spend cleaning up afterward. Our use of particular kitchen gadgets and utensils might even indicate something about our social class identities. Though the glass surrounding the Julia Child's exhibit prevents ideal picture taking, the photos in Figure **4.7.1.1** give you an idea of what was there. Might the organizational scheme used in this kitchen, or the appliances that are either present or missing from it, shape the answers to the questions posed above about human behaviors and interactions?



Figure **4.7.1.1**: A visit to Chef Julia Child's kitchen at the National Museum of American History inspired the author to think about how kitchens, kitchen utensils, and kitchen appliances have changed over time

One challenge with analyzing physical traces and material artifacts (products) is that you generally don't have access to the people who left the traces or created the artifacts that you are analyzing. (If you did find a way to contact them, then your research would no longer qualify as unobtrusive!) It can be especially tricky to analyze meanings of these materials if they come from some historical or cultural context other than your own. Situating the traces or artifacts you wish to analyze both in their original contexts and in your own is not always easy and can lead to problems related to validity and reliability. How do you know that you are viewing an object or physical trace in the way that it was intended to be viewed? Do you have the necessary understanding or knowledge about the background of its original creators or users to understand where they were coming from when they created it?

Imagine an alien trying to understand some aspect of Western human culture simply by examining our artifacts. Cartoonist Mark Parisi demonstrates the misunderstanding that could ensue in his drawing featuring three very small aliens standing atop a toilet.

One alien says, “Since water is the life-blood on this planet, this must be a temple of some sort.…Let’s stick around and see how they show their respect” (Parsi, 2006). Without a contextual understanding of Western human culture, the aliens have misidentified the purpose of the toilet, and they will be in for quite a surprise when someone shows up to use it!

The point is that while physical traces and material artifacts make excellent sources of data, analyzing their meaning takes more than simply trying to understand them from your own contextual position. You must also be aware of who caused the physical trace or created the artifact, when they created it, why they created, and for whom they created it. Answering these questions will require accessing materials in addition to the traces or artifacts themselves. It may require accessing historical documents or, if a contemporary trace or artifact, perhaps another method of data collection such as interviews with its creators.

References

Parisi, M. (2006, May 30). *Alien cartoon 6*. Off the Mark. https://www.offthemark.com/cartoon/l...den/2006-05-30

Rathje, W. (1992). How much alcohol do we drink? It’s a question…so to speak. *Garbage*, *4*, 18–19.

Rathje, W., & Murthy, C. (1992). Garbage demographics. *American Demographics*, *14*, 50–55.

Wikipedia. (2024, Decemer 29). Observational methods in psychology. https://en.wikipedia.org/wiki/Observ..._in_psychology

This page titled 4.7.1: What Is Left Behind (Physical Trace) is shared under a Public Domain license and was authored, remixed, and/or curated by Anonymous via source content that was edited to the style and standards of the LibreTexts platform.

- **11.3: Unobtrusive Data Collected by You** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/sociology-comprehensive-edition.

4.7.2: Archival Data

Learning Objectives

1. Define archival data, and list some examples.
2. Describe the strength and weakness of using archival data to answer research questions.

Another unobtrusive approach to data collection involves analyzing archival data, which is information that has already been collected for some other purpose. They can include school and hospital records, newspaper and magazine articles, Internet content, television shows, and many other things. In the digital age, there is a vast amount of archival data.

An example is a study by Pelham et al. (2005) on “implicit egotism,” or the the tendency for people to prefer people, places, and things that are similar to themselves. In one study, Pelham et al. (2005) 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.

The operational definitions of these records can be straightforward, or not. For example, counting the number of people named Virginia who live in various states based on Social Security records is relatively straightforward. But consider Peterson et al.'s (1988) study on the relationship between optimism and health using data that had been collected many years before for a study on adult development. In the 1940s, healthy male college students had completed an open-ended questionnaire about difficult wartime experiences. In the late 1980s, Peterson et al. (1988) reviewed the men’s questionnaire responses to obtain a measure of explanatory style (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, Peterson et al. (1988) used a procedure in which all negative events mentioned in the questionnaire responses, and any causal explanations for them, were identified. These events and explanations 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. Peterson et al. (1988) 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.

This method is an example of content analysis, a family of systematic approaches to identifying patterns in complex data, such as written or verbal responses. Just as 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.

References

Pelham, B. W., Carvallo, M., & Jones, J. T. (2005). Implicit egotism. *Current Directions in Psychological Science*, *14*, 106–110.

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*, 23–27.

This page titled 4.7.2: Archival Data is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **6.6: Observational Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

4.8: Summary of Measuring Variables

Key Takeaways

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

What's Next?

As this textbook is designed, this is the end of Unit 1 (Introduction to Research Methods). Now that you have a sense of some of the technical definitions, you have a basis to start reading academic research and then to start designing y our own research. The second unit will focus on research articles. It will include information to help you find, read, and understand empirical research articles, as well as instructions on formatting your own research reports.

Exercises

- Practice: Complete the Rosenberg Self-Esteem Scale and compute your overall score.
- Practice: Think of how you could measure a construct like *jealousy* using each of the following types of operational definitions:
  - Self-report (survey, interview, focus group)
  - Psychobiological measure
  - Unobtrusive measure (observation, physical trace, archival record)
- Practice: Think of how you could measure a construct like *depression* using each of the following types of operational definitions:
  - Self-report (survey, interview, focus group)
  - Psychobiological measure
  - Unobtrusive measure (observation, physical trace, archival record)
- Practice: Think of how you could measure a construct like wealth using each of the following types of operational definitions:
  - Self-report (survey, interview, focus group)

  - Psychobiological measure
  - Unobtrusive measure (observation, physical trace, archival record)
- 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.
- 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?
- Practice: Think of how you could unobtrusively measure

This page titled 4.8: Summary of Measuring Variables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **4.5: Psychological Measurement (Summary)** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

5: Literature Review

5.1: Prelude to Reviewing Literature

5.2: How to do a Literature Review

5.3: How to Read a Research Article

5.4: Interpreting Results (p-values and NHST)

5.5: Summary of Reviewing the Literature

5: Literature Review is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

5.1: Prelude to Reviewing Literature

Do you remember the overview of literature reviews in chapter 2 on the scientific method (2.3.1: Research Literature)? This chapter will go into much more detail how to find research articles, and then how to read them. All academic researchers find and read research on the topic that they intend to research.

For example, in the development of their Digital Stress Scale, Hall et al. (2021) reviewed research on how social media affects well-being, including what could amplify those effects. This led them to identify digital stress as a possible explanation for why social media use is harmful for some people but less harmful for others. Hall et al. (2021) then reviewed research on digital stress. They used what they learned from the previous research and theorizing, especially previous assessments of digital stress, to develop a Digital Stress Scale (Hall et al., 2021). In particular, Hall et al. (2021) identified single-item questions about digital stress that had been used in previous research, they read qualitative studies that included the concept of digital stress to develop relevant questions, and they wrote questions based on Steele et. al's (2020) conceptualization of digital stress. Through a series of studies, Hall et al. (2021) a set of questions that seemed to accurately measure five factors of digital stress (connection overload, approval anxiety, availability stress, Fear of Missing Out, and online vigilance) that seemed to capture the construct of digital stress differently than other measures of either stress in general, depression, and each sub-scale factor.

This example shows how important it is to review previous research to see what has already been found, and what information is still needed.

References

Hall, J. A., Steele, R. G., Christofferson, J. L., & Mihailova, T. (2021). Development of initial evaluation of a multidimensional digital stress scale. *Psychological Assessment, 33*(3), 230-242. https://doi.org/10.1037/pas0000979

Steele, R. G., Hall, J. A., & Christofferson, J. L. (2020). Conceptualizing digital stress in adolescents and young adults: Toward teh development of an empirically based model. *Clinical Child and Family Psychology Review, 23*, 15-26. https://doi.org/10.1007/s10567-019-00300-5

This page titled 5.1: Prelude to Reviewing Literature is shared under a CC BY-SA license and was authored, remixed, and/or curated by Michelle Oja.

5.2: How to do a Literature Review

Learning Objectives

1. Describe and use several methods for finding previous research on a particular research idea or question.
2. Ability to find several peer-reviewed, empirical research articles topics related to your research questions.

Refresh

Before we begin learning tips and tricks for finding academic research reports, let's review what we learned in Chapter 2 (Overview of the Scientific Method) in the first section literature reviews (2.3.1: Research Literature). You are encouraged to answer the questions in Exercise **5.2.1** on your own, the click on the link to see how close your answers were. If you choose not to engage with Exercise **5.2.1**, you should at least read the answers so that you are reminded of what you learned in previously.

Exercise 5.2.1

1. Without looking up the definition, describe what a literature review is.
2. Why are literature reviews important for any research study?
3. Without looking up the definition, describe what an empirical research report or a primary source is.
4. How are secondary sources different from primary sources?
5. What is peer review, and why is it importnat for science?

Answer

1. The definition is "Finding and reading previous research and summaries about a specific topic;" was your description close?
2. Literature reviews are important so that you can avoid replicating what others have already. Additionally, literature reviews provide examples of measures and research designs that have already been used successfully. New research should extend knowledge on a topic, but every new study does not have to reinvent every aspect of the design.
3. The definition of an empirical research report is "A type of research literature that reports on studies in which the authors analyzed data", and the definition of a primary source is "An original document; in research, an article in which the authors collected and analyzed their own data;" was your description close?
4. A secondary source is defined as "A variety of documents that summarize others' research or thoughts," while a primary source is the original researchers' results.
5. Peer review is defined as "In research publishing, a system of evaluation in which experts in the field volunteer to review research manuscript drafts to determine if the manuscript should be published, revised, or rejected." Peer review is important because it is an attempt to ensure that only high-quality scientific studies are published and shared with the research community.

Using Your Library's Academic Databases

The primary method used to search the research literature involves using one or more electronic databases. There are several databases that allow you search in multiple other databases at the same time, and these include Academic Search Premier, EBSCOhost, JSTOR, and ProQuest; Dr. MO's librarian suggests EBSCOhost for social science research. Here's a "how-to" video for EBSCOhost (EBSCO, 2025, found at Taft College's Databases Library Guide on EBSCOhost, which has more resources for EBSCOhost), but videos for other databases are just a quick online search away!

Then, you search within these databases for general as well as discipline-specific databases. For disciplines, you might want to look in APA PsychArticles, CINAHL (for nursing and allied health), Health Source: Nursing/Academic Edition (not the consumer addition), MEDLINE, Military and Government Collection (for criminal justice topics), and Psychology and Behavioral Sciences Collection. To ensure that you learn about research on all people, you might also want to review something like the Ethnic Diversity Source. Your library may have access to even more databases, so you are encouraged to talk with your librarian to learn more! These databases cost you college money, so you can only access them through your school's library.

While each database is organized slightly differently, they all consist of individual records for each article, video, link, or book in the database. Each record includes basic publication information, an abstract (summary of the work), key words (called "subjects," which are an index of terms that describe the contents of the article). The index terms are especially helpful because they are standardized; for example, research on note-taking is always indexed under the term “Learning Strategies.”

You use the databases by entering one or more search terms, then the database returns records that contain those search terms. (These interfaces are provided by different vendors and therefore can look somewhat different depending on the library you use.) Your search words are in bold in the record. The key word section (Subjects) can help you identify if you are using the wrong work for your topic. Given that there are millions, even billions, of records in these databases, 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 how using social media is related to stress and mental health. If you were to enter the search term “stress,” it would return far too many records to look through individually. Your first step would be to look at the key words, to see if we are using the correct terminology. When searching "stress," it looks like "psychological stress" is the standardized term. Sometimes, finding these terms takes a lot of time and guesswork! This is why your secondary sources may be beneficial (to provide lots of different words for the same construct). Your librarian is also an expert on databases and terminology, and the love to help!

Library Guides

Your librarian has probably already helped you! Go to your library's homepage and see if you can find Library Guides. These are little pages to provide quick info. At Dr. MO's college, there are at least two relevant Library Guides (sometimes called LibGuides):

- What is a literature review?
- What is a database?

What to Search

When you do a literature review, you need to be selective. Not every article relates to your research idea will be worth obtaining, reading, and integrating into your review. Instead, you want to focus on sources that help you do four basic things: (a) refine your research question, (b) identify appropriate research methods, (c) place your research in the context of previous research, and (d) write an effective research report. Several basic principles can help you find the most useful sources.

First, it is best to focus on recent research, keeping in mind that what counts as recent depends on the topic. A good general rule, however, is to start with sources published in the past five years. The main exception to this rule would be classic articles that turn up in the reference list of nearly every other source. If other researchers think that this work is important, even though it is old, then, by all means, you should include it in your review.

Second, you should look for review articles on your topic because they will provide a useful overview of it—often discussing important definitions, results, theories, trends, and controversies—giving you a good sense of where your own research fits into the literature. HOWEVER, some professors won't allow review articles as part of your own literature review. While they are great secondary sources and can help you with terminology and finding other sources (see their referenes), they are not primary source, empirical articles. These secondary sources can also include reports that provide information that can help you argue for the interestingness of your research question. For a study on the effects of cell phone use on driving ability, for example, you might look for information about how widespread cell phone use is, how frequent and costly motor vehicle crashes are, and so on. You can then go find the original articles provided in the review.

Third, read what others have read. You should look for empirical research reports addressing your question or similar questions, which can give you ideas about how to measure your variables and collect your data. As a general rule, it is good to use methods that others have already used successfully unless you have good reasons not to.

One key tip for doing these searches is to write down every set of terms that you search, with how many records were returned.

> *Record every set of terms that you search! You'll thank me later!*

Literature reviews take hours, sometimes over multiple days, which makes it unlikely that you'll remember every search term that you used. Instead, keep a running list of what terms and combinations of terms that you searched (and how many records were returned). As you search, you will add terminology to your list as you review abstracts and key words. You may also run across a sub-topic that you're particularly interested in (or several sub-topics!), which will results in more terms to add to your list. As you add terms to your list, you'll need to search for combinations with the original and new terms. See how it can get complicated quickly!

Researchers and librarians have found that you need to find about three articles for every one article that you will actually work for your final literature review. If your professor is asking for at least five empirical articles, then you should keep searching the databases until you have terms that result in about 15 empirical articles. You may think that five empirical articles is a lot, the average number of sources cited per article is about 50 (Adair & Vohra, 2003). As you start reading through your 15 empirical articles, you'll realize that some aren't on the exact topic that you ended up with, some may be too difficult for you to understand at this point, and some may not actually be empirical articles (and are instead secondary sources). Dr. MO tries asks her students to find 10-30 articles for literature reviews that required three to five empirical articles. In her experience, students who find fewer than 10 articles have to repeat their literature search once they start reading the articles. And more than 30 articles just seems like too many to read even the titles and abstracts; who has time for that?

Accessing Articles

Once you've found 10-30 research articles, you should start saving them for later use. While your list of key terms will help, you don't really want to recreate your literature review to re-find the articles. Depending on the vendor that provides the interface to the databases, you may be able to save, print, or e-mail the relevant records or even the whole article. To only search for records that contain full-text copies of the works themselves, there should be a filter option for "Full Text". If your library does not have access to an electronic version of the article, you still have options!

The first place that Dr. MO goes is to here librarian for an inter-libray loan. If the articles are available at other libraries, the librarian can usually get a PDF of the article within hours to a few days. If you are looking for a book, that can take up to a few weeks. Another option is to ask someone who goes to (or works at) a bigger or more research-oriented school; the libraries at those types of institutions have access to more databases so your article may be available. Dr. MO did this when one of her professor friends was back at school to get a doctorate; Dr. MO sent the publication information (from her literature review) to her friend, who searched in her university's libary database and found that her university had a PDF of the article! Other ways to access articles for free are listed below; these are great options for getting access to an article that you want but they are *not* a good way to do a literature review.

Other Ways to Access Articles

- Free journal access from APA: https://www.apa.org/pubs/highlights/sample/
- General online search or use Google scholar
- Openly licensed research article repositories:
  - Scientific Electronic Library Online for openly licensed research articles: https://scielo.org/en/
  - https://www.refseek.com/
  - https://link.springer.com/
  - https://www.science.gov/

CAUTION: These options may provide sources that aren't peer-reviewed and aren't primary source (empirical) articles.

Why aren't these good options for doing your literature review (rather than the complicated databases)? Because the databases are designed to provide peer-reviewed articles, and the abstracts can help you determine if the article is actually a primary source (empirical). These other options don't usually provide these options.



This tutorial explains how to use advanced features of Google Scholar. https://youtu.be/t1ZwgDeX2eQ

From Finding to Reading

The next section gives suggestions and advice for actually reading one of the empirical articles that you found during your literature review.

References

Adair, J. G., & Vohra, N. (2003). The explosion of knowledge, references, and citations: Psychology's unique response to a crisis. *American Psychologist*, *58*, 15–23.

EBSCO (2025). *Advanced searching on EBSCOhost: Tutorial*. Vimeo. https://vimeo.com/355538515

Taft College (2025, February 25). *EBSCOhost databases*. Databases Library Guide. https://lib.taftcollege.edu/databases/ebsco

---

This page titled 5.2: How to do a Literature Review is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **2.3: Finding a Research Topic** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

5.3: How to Read a Research Article

Learning Objectives

1. Identify the major sections of an APA Style research report and the basic contents of each section.
2. Decide your strategy to read APA Style research reports.

In this section, we look at how to read an APA-style empirical research report, but we will start with a video by UBC iSchool (2013) in Figure **5.3.1** with suggestions on how to read any scientific article.

How to Read an Academic Paper



Figure **5.3.1**: How to Read an Academic Paper by UBC iSchool, CC-BY via YouTube

But before you follow the video's suggestions, be aware that there are *tons* of different approaches to reading academic research. At the end of this sub-section, several more different approaches will be discussed. But before that, you'll need to know a little bit about how scientific articles are structured. The standard sections of an empirical research report provide a kind of outline. This information is based on the 7th edition of the Publication Manual of the American Psychological Association (APA, 2020), often referred to as APA Style manual. You can find more information at apastyle.apa.org or https://apastyle.apa.org/products/pu...al-7th-edition .

Sections of a Research Report

Here we consider each of these sections in detail, including what information it contains, and tips for reading and understanding each section. The following are the major sections in most quantitative research reports.

- Title page
- Abstract
- Main body, which includes:
  - Introduction
  - Methods, which includes sub-sections:
    - Participants
    - Materials
    - Design
    - Procedure
  - Results
  - Discussion/Conclusion
- Reference page
- Appendices (optional); figures and tables can be in an appendix or in the paper around where they are discussed.

Qualitative research articles may have different sections. Check out the APA's (2025) Journal Article Reporting Standards (JARS) webpage to compare what's included in quantitative reports compared to qualitative reports (Table 1 in both articles is particularly useful). The APA (2025) JARS website also includes writing standards for reports that include studies that are both qualitative and quantitative (mixed methods), and a new JARS article related to writing about race and ethnicity.

Where is it?

You are encouraged to open one of the empirical articles that you found in your literature review, then find the section and sub-sections that you'll be reading about here.

Note that the requirements and formatting for APA Style manuscripts is different from what and how each journal publishes the manuscript.

Title Page

An APA Style research report begins with a title page, sometimes called the cover page. The title page contains the title, author names, and affiliations. Once published in a journal, this information will formatted to flow into the next sections of an academic research article. The title should clearly and concisely (in about 12 words or fewer) communicate the primary variables and research questions. The authors are listed in an order that reflects their contribution to the research. When multiple authors have made equal contributions to the research, they often list their names alphabetically or in a randomly determined order. Next, the authors' institutional affiliation (the university or other institution where the authors worked when they conducted the research) is listed.

Abstract

The abstract is a summary of the study. The abstract presents the research question, a summary of the method, the basic results, and the most important conclusions. Because the abstract is usually limited to about 200 words, it can be a challenge to write a good one! The abstract is probably the first thing that you read after the title to determine if the article is related to your topic and worth spending the time reading.

> *Pro Tip: Re-read the abstract as you continue to work with the article to remind yourself of the IV, DV, and results.*

Main Body

The main body of any research report includes the introduction section, a section on the study's methods, a section detailing the results, and a conclusion or discussion section. We'll go through each of these sections separately.

Introduction Section

The introduction is not labeled in APA Style manuscripts and may or not be labeled in the published journal article. But now you know that the first section is always the introduction. This section includes these three of important sub-sections:

- Opening: Introduces the research question and explains why it is interesting
- Summary of Previous Research: Discusses relevant previous research
- Closing: Re-states the purpose, perhaps now with explicit research hypotheses

These sub-sections are typically *not* identified by headings. However, the second sub-section (summarizing what you learned in your literature review) may have sub-headings to help lead the reader from the general topic to specific research findings.

The opening, which is usually about a paragraph in length, introduces the research question and explains why it is interesting. After capturing the reader’s attention, the opening should go on to introduce the research question and explain why it is interesting. Will the answer fill a gap in the literature? Will it provide a test of an important theory? Does it have practical implications? Giving readers a clear sense of what the research is about and why they should care about it will motivate them to continue reading the literature review—and will help them make sense of it. When reading these paragraphs, try to determine what the purpose of the research study is. This will help you understand some of the context for what comes next.

Immediately after the opening comes the literature review, which describes the results of relevant previous research on the topic. The literature review is not simply a list of past studies.This summary of previous research can be anywhere from several paragraphs to several pages in length. Beginning students often take a paragraph or two to summarize one research article, while experienced researchers will usually use each paragraph to summarize many research studies on one topic. You'll see that accomplished researchers use the literature review as a kind of argument for why the research question is worth addressing, and the best way to answer the research question. By the end of the literature review, readers should be convinced that the research question makes sense and that the present study is a logical next step in the ongoing research process.

Like any effective argument, the literature review must have some kind of structure. For example, it might begin by describing a phenomenon in a general way along with several studies that demonstrate it, then describing two or more competing theories of the phenomenon, and finally presenting a hypothesis to test one or more of the theories. Or it might describe one phenomenon, then describe another phenomenon that seems inconsistent with the first one, then propose a theory that resolves the inconsistency, and finally present a hypothesis to test that theory. In applied research, it might describe a phenomenon or theory, then describe how that phenomenon or theory applies to some important real-world situation, and finally suggest a way to test whether it does, in fact, apply to that situation. If you can outline this structure, then you understand the background and argued need for the upcoming research study.

The last paragraph of the introduction section is the closing of the introduction. This paragraph usually includes two important elements. The first is a clear statement of the purpose of the research study, the main research question. or the research hypotheses. This statement tends to be more formal and precise than in the opening and is often expressed in terms of operational definitions of the key variables, so you can compare what is written here to what you you thought that the purpose was after the first paragraph. The second element in the closing paragraph is a brief overview of the method and some comment on its appropriateness. Here, for example, is how Darley and Latané (1968) concluded the introduction to their classic article on the bystander effect:

> These considerations lead to the hypothesis that the more bystanders to an emergency, the less likely, or the more slowly, any one bystander will intervene to provide aid. To test this proposition it would be necessary to create a situation in which a realistic “emergency” could plausibly occur. Each subject should also be blocked from communicating with others to prevent his getting information about their behavior during the emergency. Finally, the experimental situation should allow for the assessment of the speed and frequency of the subjects’ reaction to the emergency. The experiment reported below attempted to fulfill these conditions. (p. 378)

The brief overview of the methods is sometimes implied in research hypotheses. This happens when the research hypothesis includes how the IV will be created (or measured) and how the DV will be measured. In sum, the introduction leads smoothly into the next major section of the article—the method section.

Method

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

The method section begins immediately after the introduction ends, and usually starts immediately with the sub-heading “Participants.” The participants subsection indicates how many participants there were, and should include information like the gender diversity, ethnic diversity, some indication of the participants' age, other demographics that may be relevant to the study. This sub-section also, and how they were recruited, including any incentives given for participation.

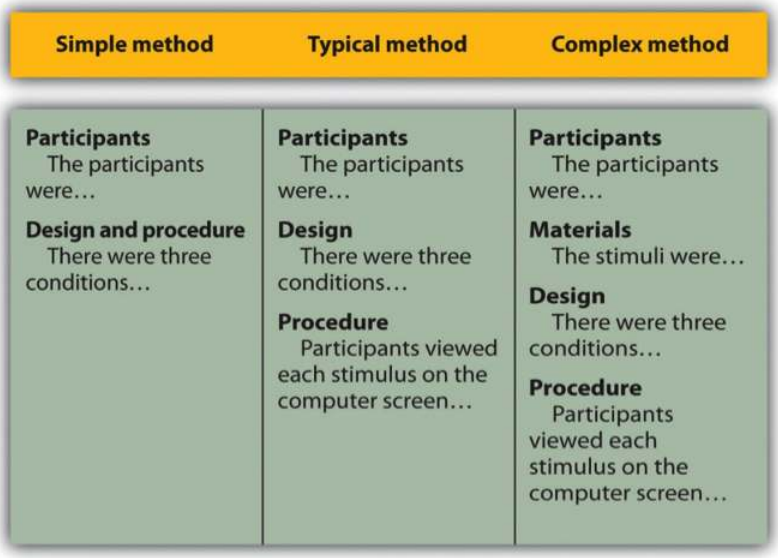


Figure **5.3.1**: Three Ways of Organizing an APA Style Method Section

After the participants section, the structure can vary a bit. Figure **5.3.1** shows three common approaches. Before we dive into the three columns in Figure **5.3.1**, let's refresh on some basic terminology in Exercise **5.3.1**

? Exercise 5.3.1

1. Without looking up the definition, can you describe what an operational definition is?
2. Without looking up the definition, can you describe what an IV, or independent variable, is?
3. Without looking up the definition, can you describe what a DV, or dependent variable, is?

Answer

1. An operational definition is "A description of the variable in terms of precisely how it is to be measured;" was your description close?
2. An IV is "The variable that the experimenter manipulates, and believes is the cause of changes in the outcome variable;" was your description close?
3. A DV is "The measured variable that the experimenter think is the effect (the thing that the IV changes). This is the outcome variable, the thing that we're trying to improve;" was your description close?

In the first column in Figure **5.3.1**), the participants section is followed by a design and procedure subsection, which describes the rest of the method. The design of a study is its overall structure. Much of this textbook will discuss different types of designs. The design could also include a description of the variables, and how they were created or measured (operational definitions). The procedure is how the study was carried out, or a description of what the participants did (not what the researchers did). For example, the participants gave their informed consent, read a set of instructions, completed a block of four practice trials, completed a block of 20 test trials, completed two questionnaires, and were debriefed and excused. This first approach (the left column of Figure **5.3.1**) This works well for methods that are relatively simple and can be described adequately in a few paragraphs, but is pretty rare.

In the second approach (middle column of Figure **5.3.1**), the participants section is followed by separate design and procedure subsections. This works well when both the design and the procedure are relatively complicated and each requires multiple paragraphs. However, is is not ideal to clearly identify the measures ("Materials") that were used to assess the DV (and possibly the IV).

In the third basic way to organize a method section (right column of Figure **5.3.1**), the participants sub-section is followed by a materials subsection before the design and procedure subsections. This works well to describe the materials used (like prompts) and the measures (like surveys). The heading of this subsection can be modified to reflect its content. Instead of “Materials,” it can be

"Questionnaires," "Stimuli," and so on. The materials subsection is also a good place to refer to the reliability and/or validity of the measures. This is where you would present test-retest correlations, Cronbach's α, or other statistics to show that the measures are consistent across time and across items and that they accurately measure what they are intended to measure.

? Exercise 5.3.2

Using the third approach to a method section (right column of Figure **5.3.1**):

1. Which sub-section would a description of an IV be?
2. Which sub-section would a description of the DV be?

Answer

1. This actually depends on what kind of study you have! If IV was created through manipulation by the researcher, then this would probably go in the Procedures sub-section. However, if the IV is more like a predictor varaible, then the researcher measures the IV rather than creates it. In this case, the measure would be described in the Materials section.
2. Measurement of the DV is usually in the Materials section.

As you are getting into the meat of the research study, this is where you might start getting confused. Before you put too much time into reading and re-reading any study if it is more complex than you are currently ready for. You'll need to understand:

- How the participants were recruited.
- What participants experienced, in what order. (This is the procedure.)
- What the IV was. (This is usually the groups that were compared, but could also be whatever was randomly assigned (manipulated) by the experimenter.)
- What the DV was. (This was what was measured.)

> *Pro Tip: As a beginning researcher, find another article if don't understand the methods.*

If you're not sure if you understand, try explaining the study to another student or a family member. You can also work with your professor, other professors in your department, or a tutor who is experienced with reading research studies.

Results

The results section is where the main results of the study are described, including the results of the statistical analyses. Students find the Result section to be very intimidating, so the next section of this textbook will go into interpreting the results of quantitative studies. But for now, you can review the structure of this section. Unlike in the Method section, there are no standard subsections. Instead, the results are (hopefully) organized logically for the reader.

Typically this section begins with certain preliminary issues. One is whether any participants or responses were excluded from the analyses and why. The rationale for excluding data should be described clearly so that other researchers can decide whether it is appropriate. A second preliminary issue is how multiple responses were combined to produce the primary variables in the analyses. While the operational definition of each measure is described in the Method section, the Result section also needs an operational definition for how variables were categorized (or not) for analysis. For example, if participants were asked to recall as many items as they could from study list of 20 words, did the analysis use a count of the number of words that were correctly recalled, the percentage of words correctly recalled, or perhaps the number correctly recalled minus the number incorrectly recalled? A final preliminary issue is whether the manipulation was successful. This is could be where you would report the results of any manipulation checks, although it also makes sense to include this information in the materials sub-section of the Method section.

The main information in the results section should then tackle the primary research questions, one at a time. Again, there should be a clear organization. One approach would be to answer the most general questions and then proceed to answer more specific ones. Another would be to answer the main question first and then to answer secondary ones. Regardless, Bem (2003) suggests the following basic structure for discussing each new result:

1. Remind the reader of the research question.
2. Give the answer to the research question in words.

3. Present the relevant statistics.
4. Qualify the answer if necessary.
5. Summarize the result.

Notice that only Step 3 necessarily involves numbers. The rest of the steps involve presenting the research question and the answer to it in words. In fact, the basic results should be clear even to a reader who skips over the numbers.

> *Pro Tip: For each research question, determine a one-sentence answer for each of Bem's (2003) five-stage structure of results.*

If the findings raise additional questions that can be answered with the current data, then this structure can be followed again. However, student researchers are often overwhelmed by the statistical analysis, so it might be a good idea to interpret the results for only the main research hypotheses for now. Some journal reviewers may ask that any additional analyses are described in the main Results section, while other journals may ask for that in an appendix or an online-only supplementary section. These additional analyses may or may not be relevant for your purposes, so don't get bogged down on reading and understanding everything.

If you're confused about information in the Results section, you can review the Discussion section to help understand what was found. However, the Discussion section is written more broadly to consider implications for the results, so the Discussion may generalize or avoid important results. The Discussion section is a good check for your understanding, but it is important to make up your own mind about what was found and what that implies based on the statistics results, not on what is written in the Discussion section.

Discussion

The discussion section is the last major section of the research report. Discussions usually consist of some combination of the following elements:

- Summary of the research results
- Limitations
- Implications of the research results
- Suggestions for future research

The discussion typically begins with a summary of the study that provides a clear answer to the research question. In a short report with a single study, this might require no more than a sentence. In a longer report with multiple studies, it might require a paragraph or even two. To qualify the results, articles often then add any problems with the study. Some of these problems may have been known before the study started, like if the researcher knew that their participant pool was limited. Some limitations could arise (or been discovered) during data collection; for example, a researcher might expect equal participation from their diverse population, but instead find that participants from only certain demographics are actually participating. But do not overdo it. All studies have limitations, and most readers will understand that a different sample or different measures might have produced different results. Unless there is good reason to think they *would* have, however, there is no reason to mention these routine issues. Instead, pick two or three limitations that seem like they could have influenced the results, explain how they could have influenced the results, and suggest ways to deal with them. Finally, researchers and reviewers may identify limitations once the study has been completed. Now is the time to discuss these issues and how they might have affected the results. If you came up with some limitations on your own while reading the Method section, it's good to check to see if the researchers thought the same thing.

The summary of what was found is often followed by a discussion of the implications of the research. These implications might more readily affect a relevant theory, or the research results could be applied to professionals in the field (practical implications). For theoretical implications, researchers might write about how the results provide support for any existing theory, or not. For practical implications, ther researcher might write about how the results be used, by whom, and to accomplish some real-world goal. After reading about the study's results and limitations, you can decide on your own how impactful the results actually are. This may lead you to question the author's perspective on the implications of their findings and overstated the result's impact, or you may decide that the authors understated how influencial their results might be.

The summary of the findings and a recognition of the study's limitations may lead naturally to the next research question and study on the topic, or this may be written after the implications. What *new* research questions has the results of study raised?

References

The Discussion section is the last section in the main body. The next section, References includes the full references, in APA Style, of everything that was cited in the entire article. The references are listed alphabetically on the page by the last name of the first author (even while the authors of each individual reference are listed in the order that they appear on the article). If you wanted to read another source that was cited in the article, this is where you'd go to find where to find that source.

Appendices

Appendices come after the references. An appendix is appropriate for material that would interrupt the flow of the research report if it were presented within any of the major sections, but is still important for understanding the study. Depending on the journal publisher, an appendix could be used to present tables, figures, lists of stimulus words, questionnaire items, detailed descriptions of special equipment or unusual statistical analyses, or picture or video screenshots. If you wanted more detail than what was provided in a specific sub-section of the main body, this is where the details would be.

Approaches to Reading Academic Research Articles

Now that you have an idea of the basic outline of most research articles, we'll discuss several approaches to reading a research article. The UCB iSchool (2013) video at the beginning of this section described taking three passes at the article. Another approach was detailed by Raff (2013) in her 11-step process. While 11 steps seems like more than three passes, you'll find that many of the suggestions are similar between UCB iSchool (2013) and Raff (2013) . Similar to Raff (2013) , some librarians suggest avoiding re-reading the abstract and instead focusing on the following four sections (in this order):

How Librarians Might Read Research Articles

1. Literature Review in the introduction section
2. Methods
3. Results
4. Discussion

As an instructor for research methods and an advisor for a research club, Dr. MO has to read a lot of research articles really quickly. Here's how she actually approaches new journal articles:

How Dr. MO Reads Research Articles

1. Abstract
2. Methods (sometimes she has to go back to the introduction section to pick up some terminology).
3. Tables and Figures in the Results; full Results if she doesn't understand the tables or figure (and the Discussion if she doesn't understand the full results)

Once she understands the basics of what happened and has read several studies on the topic and chose a few to dig in to, Dr. MO then does read the introduction section to understand the context and to see if there are other articles on the topic that she should read.

As you can see, there are many approaches to reading research articles! Now that you have some sense of the structure of research articles in the social sciences and know how others approach reading this type of technical writing, how do you think that you'll approach reading scientific journal articles? What might work best for you?

References

APA, (2020). *Publication manual of the American Psychological Association* (7th edition). American Psychological Association.

Bem, D. J. (2003). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. R. Roediger III (Eds.), *The complete academic: A practical guide for the beginning social scientist* (2nd ed.). Washington, DC: American Psychological Association.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 4*, 377–383.

Raff. J. (2013, August 25). *How to read and understand a scientific paper: A guide for non-scientists*. Violent Metaphors: Throughs from the Intersection of Science, Pseudoscience, and Conflict. https://violentmetaphors.com/2013/08...tific-paper-2/

UCB iSchool (2013, January 17). *How to read an academic paper* [Video]. YouTube. https://www.youtube.com/watch?v=SKxm2HF_-k0

This page titled 5.3: How to Read a Research Article is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **11.3: Writing a Research Report in American Psychological Association (APA) Style** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

5.4: Interpreting Results (p-values and NHST)

Learning Objectives

1. Describe inferential statistics, and how they are related to null hypothesis significance testing.
2. Interpret the results of null hypothesis significance tests in research article.

Remember?

Most students reading this textbook have already taken (or are currently taking) a statistics course. This section will remind you of what you learned that is most relevant for reading a quantitative research article. If you'd like a more in-depth refresher, LibreTexts offers a selection of openly-licensed textbooks on statistics (https://stats.libretexts.org/), including textbooks for different social sciences (https://stats.libretexts.org/Bookshe...ied_Statistics).

To understand how to interpret the results of statistical analyses, you should remember what a p-value represents. But to remember that, you might need a refresher on what a null hypothesis is.

Null Hypothesis

In general, the null hypothesis is the idea that nothing is going on: there is no effect of our treatment, no relation 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 (more on that in the next paragraph on Null Hypothesis Significance Testing). Until we have evidence against it, we must use the null hypothesis as our starting point. In sum, the null hypothesis always states that:

> *There is **no difference between the groups' means***
>
> *OR*
>
> *There is **no relationship between the variables**.*

This is enough background for our purposes, but you are welcome to learn more by reviewing the section of the openly-licensed, online textbook that this information came from (7.3: The Research Hypothesis and the Null Hypothesis in Oja, 2022).

Null Hypothesis Significance Testing

So far, so good? We develop a research hypothesis of what we expect will happen, and we have a null hypothesis that says that nothing will happen. It's at this point that things get somewhat counterintuitive; the null hypothesis seems to correspond to the opposite of what we expect, and then we focus exclusively on that. It seems like we're neglecting the thing that we're actually interested in (the research hypothesis). Based on Hall et al.'s (2021) research validating the Digital Stress Scale (DSS), we could develop a null hypothesis that says that people who experience higher levels of digital stress will have similar levels of depression than those with lower levels of digital stress (no difference between group means) or a null hypothesis could be that there is no linear relationship between scores on the DSS and scores on a depression scale (no relationship between the variables). But, we actually *do* think that digital stress would be related to depression, so the alternative to this null hypothesis is that those with higher scores on the DSS will have higher depression scores than those with lower scores on the DSS (there is a difference between the means) or we predict that as scores on the DSS increase, depression also increases (there is a positive, linear relationship between the variables). **The important thing to recognize is that the goal of a hypothesis test is *not* to show that the research 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, in my experience, is to imagine that a hypothesis test is a criminal trial… *the trial of the null hypothesis*. 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 you, the researcher, can prove beyond a reasonable doubt that it is false. You are free to design your experiment however you like, 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 doesn't get a lawyer. And given that the researcher is trying desperately to prove it to be false, *someone* has to protect it.

In sum, the purpose of null hypothesis significance testing is to be able to **reject** the *expectation* that the means of the two groups are the *same*.

- Reject the *null hypothesis*:
  - The means *are* different.
  - or
  - There *is* relationship between the variables.
- Retain the null hypothesis:
  - The means are similar (The means are *not* different).
  - or
  - The is *no* relationship between the variables.

Finally, you reject or retain the null hypothesis, and you support or or don't support the research hypothesis.

Why predict that two things are similar?

Because each sample's mean will vary around the population mean (see the first few sections of this chapter to remind yourself of this), we can't tell if our sample's mean is within a "normal" variance. But we *can* gather data to show that this sample's mean is different (enough) from the population's mean. This is *rejecting* the null hypothesis. We use statistics to determine the probability of the null hypothesis being true.

Why can't we prove that the mean of our sample is different from the mean of the population? Since different samples from the same population have different statistical qualities, we can't guarantee that the sample that we're analyzing actually has the qualities of the population. Researchers are a conservative bunch; we don't want to stake our reputation on a sample mean that could be fluke; what if our sample had a lot of people with very lower scores on digital stress? Then we couldn't be sure that what our sample showed us would also happen in the population.

But what we can show is that our sample is so extreme that it is statistically unlikely to be similar to the population. Null hypothesis significant testing is like how courts decide if defendants are Guilty or Not Guilty, not their Guilt v. Innocent. Similarly, we decide if the sample is similar to the population or not.

This is a tough concept to grasp, so we'll keep working on it. And if you never get it, that's okay, too, as long as you remember the pattern of rejecting or retaining the null hypothesis, and supporting or not supporting the research hypothesis. You are also welcome to learn more by reviewing the section of the openly-licensed, online textbook that this information came from (7.4: Null Hypothesis Significance Testing in Oja, 2022).

p-value

Okay, so you've remembered about null hypotheses and a little about Null Hypothesis Significance Testing, but how does that relate to reading the results of a quantitative research article? If you look at a research article's statistical results, you probably saw p-values. **We use p-values to show us the probability of obtaining the result from any one sample *if* the null hypothesis were true**.

? Exercise 5.4.1

What does the null hypothesis say?

Answer

The null hypothesis says that the means of the groups are similar OR that there is no relationship between the groups.

The p-value shows us the probability to determine if any effects could have happened by chance. The p-value tells us the **probability** of getting an effect this different if the sample is the same as the population. If the probability (p-value) is small

enough (p< .05), then we conclude that the sample probably is from a *different* population.

Without having to understand everything about probability distributions and the Standard Normal Distribution, what do the p-values tell us?

- A small p-value means a *small probability* that the means are similar or that there is no relationship between the variables. This suggests that that the means are *different* or that this *is* a relationship.
- A large p-value means a *large probability* that the means are similar or that there is no relationship between the variables. This suggests that there is nothing going on..

Note

In other words:

Reject null = p<.05

Retain null = p>.05

If the probability is less than 5% that you would get a sample that is *this different* from the population if the sample really is from the population, the sample is probably *not* from that population. If you took 100 samples from a population, less than 5 of the samples would be *this different* from the population if the samples were different in reality. So, the sample is probably not from the population. Probably.

You have a 5% chance that your results are wrong. You have a (small) chance that your sample is *this different* from the population, but the sample is still actually from the population.

Statisticians are okay with being wrong 5% of the time!

Let's practice!

Example 5.4.2

For each:

1. Determine whether to report “p<.05 ” or “p>.05”
2. Determine whether to retain or reject the null hypothesis.
3. Determine whether something is happening in the sample, or if the sample is from the population.

Hint:

Reject null = p<.05

Retain null = p>.05

1. p < .05
2. p = .138
3. p = .510

Solution

1. p < .05
    1. “p<.05 ”
    2. Reject the null hypothesis.
    3. Something is happening; the sample and population are different.
2. p = .138
    1. “p>.05” (because 0.138 > 0.05, 0.138 is bigger than 0.05)
    2. Retain the null hypothesis.
    3. Nothing is happening; the sample and population are similar.
3. p = .510

1. “p>.05” (We’re comparing to 0.05, or 5%, not .50, or 50%)
2. Retain the null hypothesis.
3. Nothing is happening; the sample and population are similar.

Now, try it yourself:

? Exercise 5.4.2

For each:

1. Determine whether to report “p<.05 ” or “p>.05”
2. Determine whether to retain or reject the null hypothesis.
3. Determine whether something is happening in the sample, or if the sample is from the population.

Hint:

Reject null = p<.05

Retain null = p>.05

1. p > .05
2. p = .032
3. p = .049

Answer

1. p > .05
    1. “p>.05 ”
    2. Retain the null hypothesis.
    3. Nothing is happening; the sample and population are similar.
2. p = .032
    1. “p<.05 ” or “p>.05” (because 0.32 < .05, 0.032 is less than 0.05)
    2. Reject the null hypothesis.
    3. Something is happening; the sample and population are different.
3. p = .049
    1. “p<.05 ” (even though it’s close, .049 is smaller than 0.05)
    2. Reject the null hypothesis.
    3. Something is happening; the sample and population are different.

Now that you've been reminded of what you probably already learned, let's lock in to see how this relates to reading a research article.

Reading Research Results

Even if you're still a little lost about null hypotheses and p-values, you can still interpret the results of many research articles. The main task right now is to narrow your focus. The results section can be confusing, with lots of tables and graphs and numbers. Rather than reading it all at once, go back to the introduction section to identify the research questions or research hypotheses then only look for analyses that are answering these questions. The research question or research hypothesis will imply some IVs (groups that the researcher thinks are the cause of changes) and the DV (the outcome that was measured that the researcher wants to improve). Can you identify the results of how the IV affected the DV (or didn't) in the graphs, tables, or written paragraphs? If so, then you have found the most important information in this section! Dr. MO often tries to answer the research question by only looking at the graphs or tables; if she can do that, then she understand what was found.

There may be a ton of other analyses and data provided. These may provide additional context for the results of the main research hypotheses, or they may be interested results that weren't the main point of the study at all. When first reading the results section, it is perfectly okay to ignore anything that isn't related to the original research questions. Don't get bogged down in all of the results; focus on the results related to why you originally wanted to read the article.

While narrowing your focus on the research questions or research hypotheses, you then narrow your focus to identify which analyses show p<.05.

? Exercise 5.4.3

What does p<.05 tell you about whether or not the intervention worked?

Hint: It's about the null hypothesis.

Answer

When p<.05, it means that there is a small probability that the there is no difference between those who received an intervention and those who did not. This suggests that there is something happening, that the intervention did work.

Often, tables will identify which analyses have p<.05 with an asterisk; so easy! Once you find which null hypotheses are rejected, you can find which research hypotheses (or parts of a research hypothesis) are supported. At this point, a self-check is in order. The next step after narrowing your focus to the research questions/hypotheses and p<.05 is to see if you can explain the findings to someone else. This can be someone in your class, one of your professors, or a family member. The key is that you should have enough of a grasp of the variables and the statistics that you can simplify the results for someone else.

What should you identify?

Answer the following questions to ensure that you can simplify the important information:

1. List each DV from the Methods section.
2. List one hypothesis from the introduction section.
3. Describe the results related to each of their hypothesis or purpose. Briefly describe in words what their statistical results show.

If you have trouble with any of this, the Discussion section of the article will help clarify the results. However, keep in mind that the purpose of a Discussion section is different from the purpose of a Results section. You will not get the details and specifics from the Discussion that you should be identifying in the statistical results, so only use the Discussion if you get lost in the results. A second option if you get lost is to ask for help. Your research methods professor, any professor in your department, really, has read many, many research articles. Visit them during their office hours with a printed version of your article, or email them with a PDF of your article and what you think the results are. Another resource is your college's learning center. There are probably free tutors available, so find someone who is familiar with behavioral statistics or who has experience reading social science research articles.

Advice for Reading Results Sections

Here's a summary of the tips presented in this last section:

1. Narrow your focus to:
    1. Research Questions or Research Hypotheses
    2. p<.05
2. Explain the findings to someone else.

If you get stuck on either of these, you are encouraged to:

1. Review the article's Discussion section for clarification.
2. Work with your professor, another professor in your department, or a behavioral statistics tutor.

References

Hall, J. A., Steele, R. G., Christofferson, J. L., & Mihailova, T. (2021). Development of initial evaluation of a multidimensional digital stress scale. *Psychological Assessment, 33*(3), 230-242. https://doi.org/10.1037/pas0000979

Oja, M. (2022, May 12). Behavioral statistics. LibreTexts. https://stats.libretexts.org/Courses...Sciences_(Oja)

This page titled 5.4: Interpreting Results (p-values and NHST) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

5.5: Summary of Reviewing the Literature

Key Takeaways

- Your librarian and library databases are critical tools for finding research literature on your topic.
- Secondary sources like websites and review articles may be useful at the beginning of your literature review, but you should focus your search on empirical (primary source) research articles that have been peer-reviewed.
- Research articles have a typical structure that can help you learn to read them: introduction, methods, results, conclusion, references.
- Determine your own strategy for reading scientific articles.
- To interpret the results of quantitative research articles, it's important to remember that null hypotheses state that nothing is happening, and that p-values show the probability that the null hypothesis is true. This is the basis of null hypothesis significance testing.
- Another key ot interpreting the results of quantitative research articles is to focus on the original research questions so that you don't get overwhelmed with all of the statistical analyses provided.
- There are resources if you get stuck when interpreting the results of a scientific research article, including that article's discussion section, professors in the field, or behavioral statistics tutors.

What's Next?

This chapter gave you a refresher on interpreting statistics to help you understand research articles. Next, we focus more on understanding statistical analyses and what they can tell us.

Exercises

- Practice: Try to find an empirical article (primary source) on "study skills" using your library's databases.
- Practice: With an empirical (primary source) article that you have found, identify the parts of the introduction section, method section, the results, and the parts of the discussion section.
- Practice: Imagine a study measured growth mindset and how long the high student participants spent on their math homework. What’s the null hypothesis for scoring higher on growth mindset (compared to the population of high school students) and how long students spent on their homework?
- Practice: Does a true null hypothesis say a sample mean and the population mean are similar or different?
- Practice: If we’re saying that the sample is probably different from the population, are we rejecting or retaining (failing to reject) the null hypothesis?
- Discussion: Imagine a study showing that people who eat more broccoli tend to be happier. Explain for someone who knows nothing about statistics how this could be tested through null hypothesis signficance testing?
- Discussion: Imagine that you conduct a *t*-test to compare the final grade percentage of students who go to tutoring to students (average grade of 93.3%) who did not go to tutoring (average grade of 71.7%). If the *p* value was p=.02, would you retain or reject (fail to retain) the null hypothesis? Would you recommend that the school continue to hire tutors; why or why not?

This page titled 5.5: Summary of Reviewing the Literature is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **13.6: Inferential Statistics (Summary)** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

6: Analyzing

- 6.1: Prelude to Analyzing
- 6.2: Describing Variables
- 6.3: Graphs
  - 6.3.1: Graphing Qualitative Variables- Bar Graphs
  - 6.3.2: Describing the Shape of Graphs of Quantitative Data
    - 6.3.2.1: Graphing Quantitative Data- Line Graphs
    - 6.3.2.2: Graphing Quantitative Data- Histograms
  - 6.3.3: Graphing Quantitative Data- Scatterplots
  - 6.3.4: Honesty in Graphing
- 6.4: Understanding Null Hypothesis Significance Testing
- 6.5: Comparing Means
  - 6.5.1: Effect Sizes
- 6.6: Introduction to Correlations
  - 6.6.1: Correlation versus Causation
  - 6.6.2: Factor Analysis
- 6.7: The Qualitative Approach
  - 6.7.1: Qualitative Analysis- Preparing and Managing Qualitative Data
  - 6.7.2: Qualitative Coding
  - 6.7.3: From Qualitative Data to Findings
- 6.8: Summary of Analyzing

6: Analyzing is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

6.1: Prelude to Analyzing

Recall that Matthias Mehl and his colleagues, in their study of sex differences in talkativeness, found that the women in their sample spoke a mean of 16,215 words per day and the men a mean of 15,669 words per day (Mehl, Vazire, Ramirez-Esparza, Slatcher, & Pennebaker, 2007)[1]. But despite this sex difference in their sample, they concluded that there was no evidence of a sex difference in talkativeness in the population. Recall also that Allen Kanner and his colleagues, in their study of the relationship between daily hassles and symptoms, found a correlation of +.60 in their sample (Kanner, Coyne, Schaefer, & Lazarus, 1981)[2]. But they concluded that this finding means there *is* a relationship between hassles and symptoms in the population. This assertion raises the question of how researchers can say whether their sample result reflects something that is true of the population. The answer to this question is that they use a set of techniques called inferential statistics, which is what this chapter is about.

References

1. Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, *317*, 82.
2. Kanner, A. D., Coyne, J. C., Schaefer, C., & Lazarus, R. S. (1981). Comparison of two modes of stress measurement: Daily hassles and uplifts versus major life events. *Journal of Behavioral Medicine*, *4*, 1–39.

6.1: Prelude to Analyzing is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **13.1: Prelude to Inferential Statistics** has no license indicated.

6.2: Describing Variables

Learning Objectives

1. Use frequency tables and histograms to display and interpret the distribution of a variable.
2. Compute and interpret the mean, median, and mode of a distribution and identify situations in which the mean, median, or mode is the most appropriate measure of central tendency.
3. Compute and interpret the range and standard deviation of a distribution.

Introduction to Analysis

In the chapter on reading a research article, there was a section that focused on interpreting statistics in a research article. Now, the sections in this chapter will go into more detail about organizing and describing quantitative data, null hypothesis significance testing to compare means or test (linear) relationships, and introduce you to qualitative research and analyses. So why is there another whole chapter about analyses? There are a couple reasons. First, the additional details in this chapter can help you better interpret many analyses that you may across in the articles that you're starting to read. The section in the previous chapter provided a quick refresher, but was not designed to provide much detail or context. A second reason for this chapter is a review of statistical analyses for when you are asked to analyze data for any class projects. It is expected that you previously passed a statistics class, but it is also likely that you forgot much of what you learned. This chapter will provide a refresher. This is particularly important for those of you who took a statistics class through your math department instead of a behavioral statistics course through your psychology or sociology departments; the statistical analyses are the same, but the emphasis differs between mathematicians and social scientists. The summary in this section will orient you to how social scientists approach statistical analyses.

Descriptive Statistics

The first step to analyzing quantitative data is to organize and summarize the data. Descriptive statistics refers to a set of techniques for summarizing and displaying data. Although in most cases the primary research question will be about one or more statistical relationships between variables, it is also important to describe each variable individually. For this reason, we begin by looking at some of the most common techniques for describing single variables. This should be a reminder of what you learned in many math classes throughout your educational journey.

The Distribution of a Variable

Every variable has a distribution, which is the way the scores are distributed across the levels of that variable. For example, in a sample of 100 university students, the distribution of the variable “number of siblings” might be such that 10 of them have no siblings, 30 have one sibling, 40 have two siblings, and so on. In the same sample, the distribution of the variable “sex” might be such that 44 have a score of “male” and 56 have a score of “female.”

Frequency Tables

One way to display the distribution of a variable is in a frequency table. Table **6.2.1**, for example, is a frequency table showing a hypothetical distribution of scores on the Rosenberg Self-Esteem Scale for a sample of 40 college students. The first column lists the values of the variable—the possible scores on the Rosenberg scale—and the second column lists the frequency of each score. This table shows that there were three students who had self-esteem scores of 24, five who had self-esteem scores of 23, and so on. From a frequency table like this, one can quickly see several important aspects of a distribution, including the range of scores (from 15 to 24), the most and least common scores (22 and 17, respectively), and any extreme scores that stand out from the rest.

Table **6.2.1**: Frequency Table Showing a Hypothetical Distribution of Scores on the Rosenberg Self-Esteem Scale

Self-esteem	Frequency
24	3
23	5
22	10
21	8
20	5
19	3
18	3
17	0
16	2
15	1

Frequency Charts

A frequency polygon, often referred to as a frequency line graph, is a graphical display of a distribution of quantitative variables. It presents the same information as a frequency table but in a way that is even quicker and easier to grasp. The line graph in Figure **6.2.1** presents the distribution of self-esteem scores in Table **6.2.1**. The *x*-axis of the histogram represents the variable and the *y*-axis represents frequency. Similar information can be displayed in a histogram, but there are more rules to how to create a histogram. Additionally, a histogram uses bars to represent the frequency of categories of scores. There are no gaps or spaces between the bars; this is what makes a histogram visually different from a bar chart. What makes a histogram conceptually different from a bar chart is that the frequency of quantitative variables are displayed on histogram, while the frequency of categorical variables are displayed on bar charts.

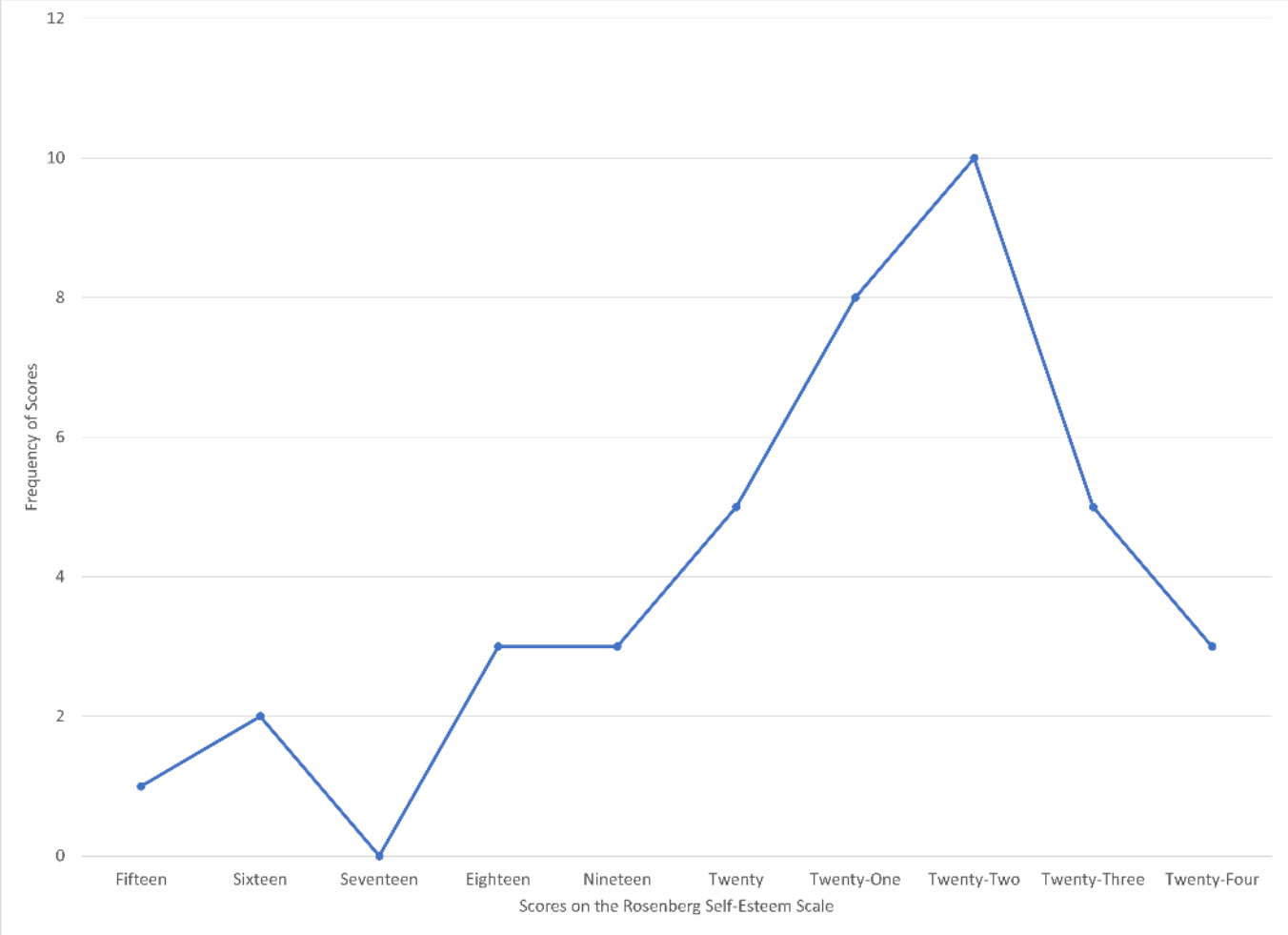


Figure **6.2.1**: Frequency Polygon Showing the Distribution of Self-Esteem Scores Presented in Table **6.2.1**

Distribution Shapes

When the distribution of a quantitative variable is displayed in a frequency polygon (frequency line graph) or a histogram, it has a shape. The shape of the distribution of self-esteem scores in Figure **6.2.1** is a little negatively skewed since there are more people who score higher, and few who score lower; in other words, there is a peak somewhere towards the right of the distribution and "tails" that taper in to the left. Figure **6.2.2** shows the two kinds of skew (negative or positive), plus what a symmetrical histogram might look like. The distribution on the left is negatively skewed, with its peak shifted toward the upper end of its range and a relatively long negative tail. The distribution on the right is positively skewed, with its peak toward the lower end of its range and a relatively long positive tail.

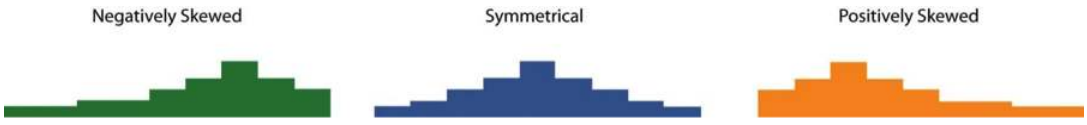


Figure **6.2.2**: Histograms Showing Negatively Skewed, Symmetrical, and Positively Skewed Distributions

An outlier is an extreme score that is much higher or lower than the rest of the scores in the distribution. Sometimes outliers represent truly extreme scores on the variable of interest. For example, on the Beck Depression Inventory, a single clinically depressed person might be an outlier in a sample of otherwise happy and high-functioning peers. However, outliers can also represent errors or misunderstandings on the part of the researcher or participant, equipment malfunctions, or similar problems.

Another thing to look at in a distribution is whether there is only one, or more, frequently occurring scores. The distribution of Figure **6.2.1** is unimodal, meaning it has one distinct peak, but distributions can also be bimodal, meaning they have two distinct

peaks. Figure **6.2.3**, for example, shows a hypothetical bimodal distribution of scores on the Beck Depression Inventory. Distributions can also have more than two distinct peaks, but these are relatively rare when your sample size is big enough.

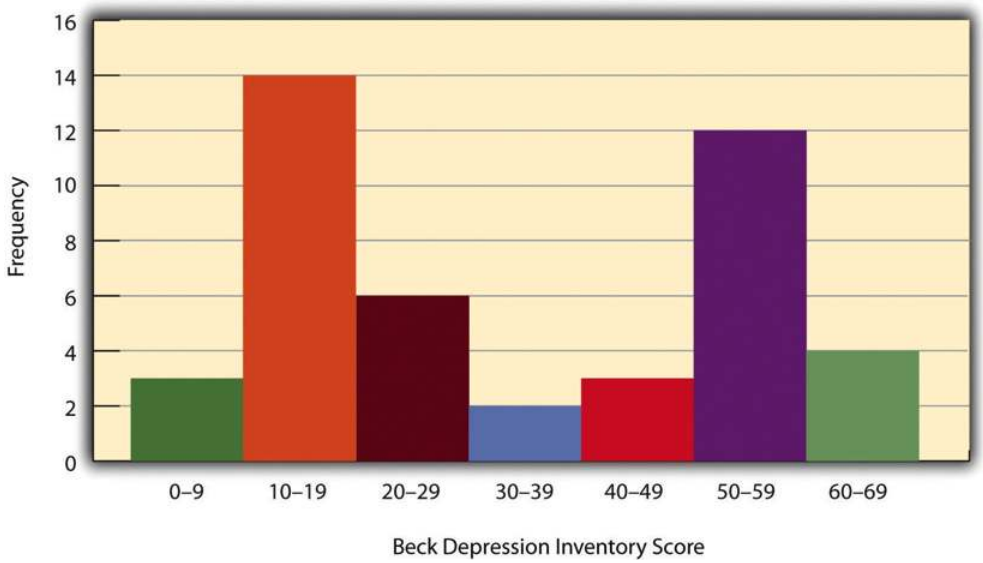


Figure **6.2.3**: Histogram Showing a Hypothetical Bimodal Distribution of Scores on the Beck Depression Inventory

Measures of Central Tendency and Variability

It is also useful to be able to describe the characteristics of a distribution more precisely. Here we look at how to do this in terms of two important characteristics: their central tendency and their variability.

Central Tendency

The central tendency of a distribution is its middle—the point around which the scores in the distribution tend to cluster. Looking back at Figure **6.2.1**, for example, we can see that the self-esteem scores tend to cluster around the values of 20 to 22. Here we will consider the three most common measures of central tendency: the mean, the median, and the mode.

The mean of a distribution (symbolized *M*) is the sum of the scores divided by the number of scores. It is an average. As a formula, it looks like this:

$$M = \frac{\sum X}{N} \tag{6.2.1}$$

In this formula, the symbol Σ (the Greek letter sigma) is the summation sign and means to sum across the values of the variable *X*. *N* represents the number of scores. The mean is by far the most common measure of central tendency, and there are some good reasons for this. It usually provides a good indication of the central tendency of a distribution, and it is easily understood by most people. In addition, the mean has statistical properties that make it especially useful in doing inferential statistics.

An alternative to the mean is the median. The median is the middle score in the sense that half the scores in the distribution are less than it and half are greater than it. The simplest way to find the median is to organize the scores from lowest to highest and locate the score in the middle. Consider, for example, the following set of seven scores:

8 4 12 14 3 2 3

To find the median, simply rearrange the scores from lowest to highest and locate the one in the middle.

2 3 3 **4** 8 12 14

In this case, the median is 4 because there are three scores lower than 4 and three scores higher than 4. When there is an even number of scores, there are two scores in the middle of the distribution, in which case the median is the value halfway between them. For example, if we were to add a score of 15 to the preceding data set, there would be two scores (both 4 and 8) in the middle of the distribution, and the median would be halfway between them (6).

One final measure of central tendency is the mode. The mode is the most frequent score in a distribution. In the self-esteem distribution presented in Table **6.2.1** and Figure **6.2.1**, for example, the mode is 22. More students had that score than any other. The mode is the only measure of central tendency that can also be used for categorical variables.

In a distribution that is both unimodal and symmetrical, the mean, median, and mode will be very close to each other at the peak of the distribution. In a bimodal or asymmetrical distribution, the mean, median, and mode can be quite different. In a bimodal distribution, the mean and median will tend to be between the peaks, while the mode will be at the tallest peak. In a skewed distribution, the mean will differ from the median in the direction of the skew (i.e., the direction of the longer tail). For highly skewed distributions, the mean can be pulled so far in the direction of the skew that it is no longer a good measure of the central tendency of that distribution. Imagine, for example, a set of four simple reaction times of 200, 250, 280, and 250 milliseconds (ms). The mean is 245 ms. But the addition of one more score of 5,000 ms—perhaps because the participant was not paying attention—would raise the mean to 1,445 ms. Not only is this measure of central tendency greater than 80% of the scores in the distribution, but it also does not seem to represent the behavior of anyone in the distribution very well. This is why researchers often prefer the median for highly skewed distributions (such as distributions of reaction times).

Keep in mind, though, that you are not required to choose a single measure of central tendency in analyzing your data. Each one provides slightly different information, and all of them can be useful.

Measures of Variability

The variability of a distribution is the extent to which the scores vary around their central tendency. Consider the two distributions in Figure **6.2.4**, both of which have the same central tendency. The mean, median, and mode of each distribution are 10. Notice, however, that the two distributions differ in terms of their variability. The top one has relatively low variability, with all the scores relatively close to the center. The bottom one has relatively high variability, with the scores are spread across a much greater range.

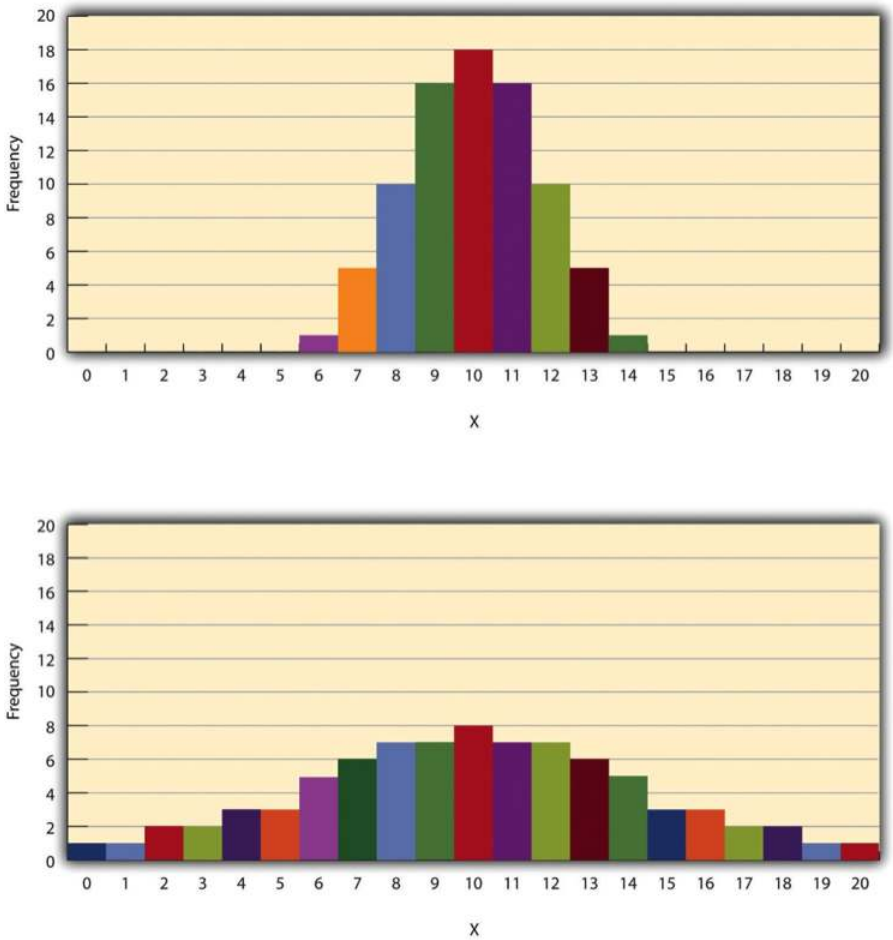


Figure **6.2.4**: Histograms Showing Hypothetical Distributions With the Same Mean, Median, and Mode (10) but With Low Variability (Top) and High Variability (Bottom)

One simple measure of variability is the range, which is simply the difference between the highest and lowest scores in the distribution. The range of the self-esteem scores in Table **6.2.1**, for example, is the difference between the highest score (24) and the lowest score (15). That is, the range is 24 − 15 = 9. Although the range is easy to compute and understand, it can be misleading when there are outliers. Imagine, for example, an exam on which all the students scored between 90 and 100. It has a range of 10. But if there was a single student who scored 20, the range would increase to 80—giving the impression that the scores were quite variable when in fact only one student differed substantially from the rest.

By far the most common measure of variability is the standard deviation. The standard deviation of a distribution is the average distance between the scores and the mean. For example, the standard deviations of the distributions in Figure **6.2.4** are 1.69 for the top distribution and 4.30 for the bottom one. That is, while the scores in the top distribution differ from the mean by about 1.69 units on average, the scores in the bottom distribution differ from the mean by about 4.30 units on average.

Computing the standard deviation involves a slight complication. Specifically, it involves finding the difference between each score and the mean, squaring each difference, finding the mean of these squared differences, and finally finding the square root of that mean. The formula for the standard deviation for a sample looks like this:

$s_{sample}=$

$$\sqrt{\frac{\sum (X-M)^2}{N-1}} \tag{6.2.2}$$

The computations for the standard deviation are illustrated for a small set of data in Table **6.2.2**. The first column is a set of eight scores that has a mean of 5. The second column is the difference between each score and the mean. The third column is the square of each of these differences. Notice that although the differences can be negative, the squared differences are always positive—meaning that the standard deviation is always positive. At the bottom of the third column is the variance (symbolized s^2), which is the sum of the squares divided by the total number of scores minus 1 (N-1). Although the variance is itself a measure of variability, it generally plays a larger role in inferential statistics than in descriptive statistics. Finally, below the variance is the square root of the variance, which is the standard deviation.

Table **6.2.3**: Computations for the Standard Deviation

X	X_M	$(X-M)^2$
3	−2	4
5	0	0
4	−1	1
2	−3	9
7	2	4
6	1	1
5	0	0
8	3	9
M = 5		s^2=28/(8-1)=4.00
		s=√4.00=2.00

If you have already taken a statistics course, you may have learned to divide the sum of the squared differences by N − 1 rather than by N when you compute the variance and standard deviation. Why is this? This is because the standard deviation of a sample tends to be a bit lower than the standard deviation of the population the sample was selected from. Dividing the sum of squares by N − 1 corrects for this tendency and results in a better estimate of the population standard deviation. Because researchers generally think of their data as representing a sample selected from a larger population—and because they are generally interested in drawing conclusions about the population—it makes sense to routinely apply this correction.

Describe to Infer

Now that you've had a refresher on descriptive the data that you collected (or understand the description in an article that you're reading), we will move to another reminder of null hypothesis significance testing, then a brief refresh about the statistical analyses to test research hypotheses. As always, you can find a selection of openly-licensed textbooks on statistics (https://stats.libretexts.org/) on LibreTexts, including textbooks for different social sciences (https://stats.libretexts.org/Bookshe...ied_Statistics).

This page titled 6.2: Describing Variables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **12.1: Describing Single Variables** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

6.3: Graphs

I have scores on a 100-point final exam from one class provided by OpenIntro.org Can you look at them in Table **6.3.1** and quickly tell how the class did?

Table **6.3.1**- Final Exam Scores

79	83	66	81
83	72	89	88
57	74	78	69
82	73	81	77
94	71	78	79

Even if you could get a general idea of how these 20 students did on the final exam, this is not the best way to understand data sets. And just imagine if the class had 40 or 100 students in it!

When you deal with data, you may have so many numbers to you that you will be overwhelmed by them. That is why we need ways to describe the data in a more manageable fashion.

We'll start that in this chapter on charts, then describe whole sets of data in only a few important numbers in the following chapter.

This page titled 6.3: Graphs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **2.1: This is what too many numbers looks like** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

6.3.1: Graphing Qualitative Variables- Bar Graphs

Learning Objectives

1. Differentiate when to use a bar chart and other types of charts.
2. Interpret bar chart charts.

Graphing Qualitative Variables

The first types of charts that will be covered are bar charts, which should be used on qualitative variables only. Qualitative variables represent categories, so these are variables in what the different options are named types. For example, if college major is the qualitative variable, then specific majors would be the different answer options. Dr. MO majored in Psychology.

Bar Charts

Bar charts are meant to display the frequencies of qualitative variables. Figure **6.3.1.1** shows the same data as the frequency table in Table 2.2.3, but with bars.

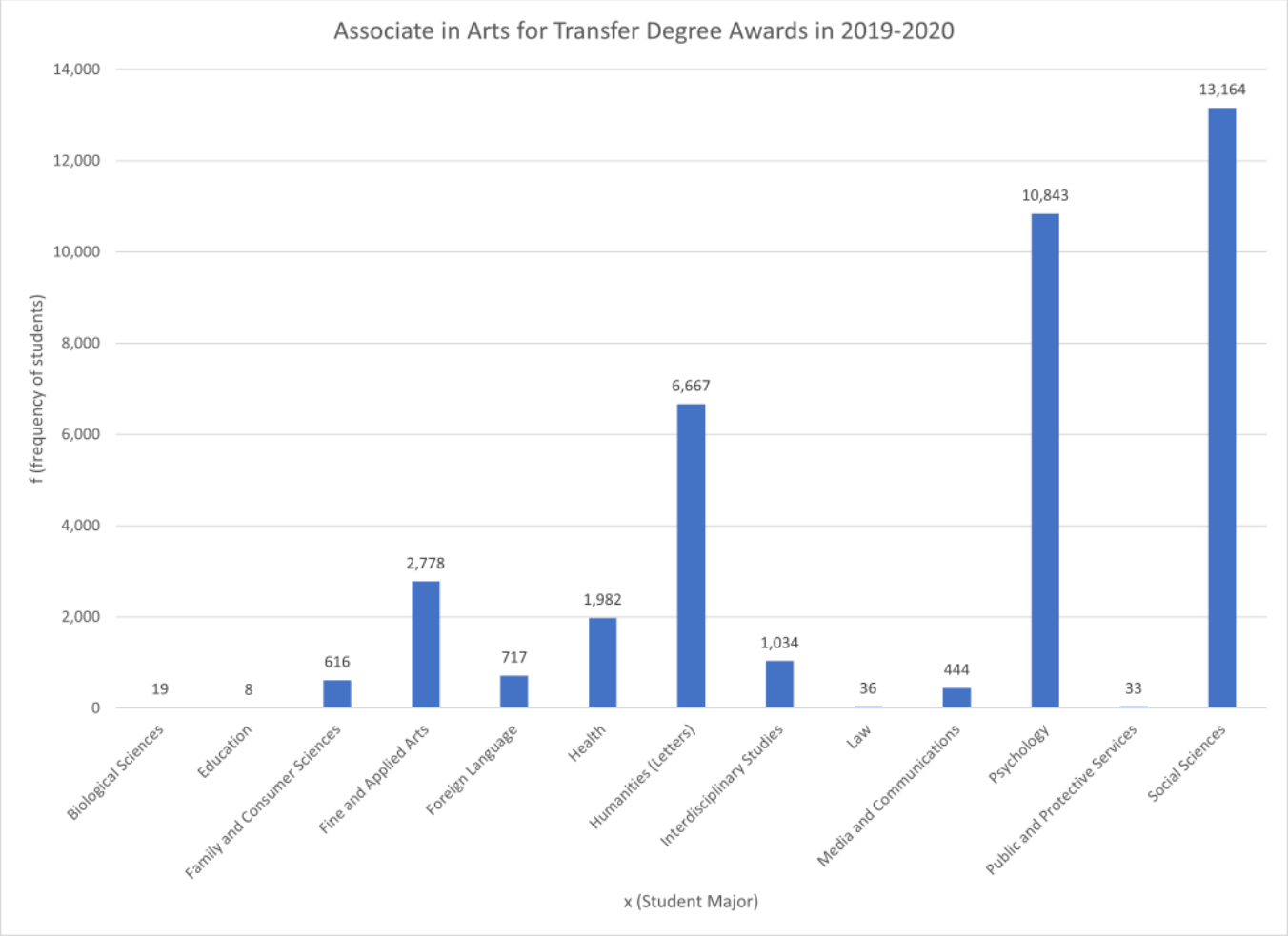


Figure **6.3.1.1**: AA-T Graduates by Major (Copyright CC0; chart created by Michelle Oja via California Community Colleges Chancellor's Office DataMart)

Let's unpack Figure **6.3.1.1** a bit.

First, the x-axis has the major that the graduating student earned the degree in. The x-axis is always the one on the bottom going from left to right. Because we are talking about a variable that has categories (nominal scale of measurement; qualitative variable),

the information on the x-axis is the category names. They can be in any order on the chart because qualitative variables don't have a natural order. As you can see, they are ordered alphabetically. This makes no sense to Dr. MO, so she usually orders the categories on the x-axis from lowest frequency to highest frequency. You can choose any order that you'd like since qualitative variables have no natural order.

Second, the y-axis in Figure **6.3.1.1** shows the number of students who earned that degree. The y-axis is always going up and down. Stand with your hands touching above your head to show that this is the axis that is going up and down, then move your hands out to a "Y" (like in the song YMCA) to remember that the y-axis is the one that goes up and down.

Third, notice that the bars do not touch; this is what makes it a bar chart and not a histogram. This chart was created with common commercial spreadsheet software (Excel), but note that this type of chart is called a "column bar chart" in the software. There are fancier types of bar charts, but often the fancier a chart looks, the more difficult it is to quickly understand what the chart is showing.

Interpreting Bar Charts

The following questions will be asked for each type of graph in this section. In most cases, the questions will also be answered so that you can get an idea of what you should be looking for in graphs. Some of these questions won't make sense for specific types of graphs, but I want you to get in the habit of answering them to yourself when reviewing graphs.

1. What kind of graph is Figure **6.3.1.1**?
    1. This is a bar graph (notice that the bars are not touching) because the variable is a qualitative category (nominal scale of measurement).
2. What does the x-axis measure in Figure **6.3.1.1**?
    1. The x-axis is the one on the bottom, and it was named "Student Major". This is the major that the student earned an Associate of Arts for Transfer degree from a California community college student during the 2019-2020 academic year.
3. What does the y-axis measure in Figure **6.3.1.1**?
    1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. In this figure, that means that the y-axis shows how many students earned the Associate of Arts for Transfer degree from each major.
4. What do you notice from Figure **6.3.1.1**? What pops out to you?
    1. What I noticed is that most students are earning the general "Social Science" degree, rather than something more specific like Sociology or Psychology. I also noticed that there were very few students earned an Associate of Arts for Transfer degree with an Education major.
5. What does Figure **6.3.1.1** make you wonder about?
    1. I wonder what kind of degree future teachers are earning, because it does not look like they are earning an Associate of Arts for Transfer! Maybe a general Associate of Arts? Or is there a different major that they are in?
6. What is a catchy headline for Figure **6.3.1.1**?
    1. Be Like the Ten Thousand Psychology Graduates- Go to a California Community College
7. How could you summarize the info in Figure **6.3.1.1** into one sentence?
    1. The most common Associate of Arts for Transfer degrees were in the Social Sciences, Psychology, Humanities, and Fine and Applied Arts.
8. Who might want to know the information in Figure **6.3.1.1**?
    1. California community college students? Potential California community college students? Faculty or administrators of California community colleges?

? Exercise 6.3.1.1

Dr. Chew, a cognitive psychologist, has a set of videos helping students use psychology to learn better. Dr. Chew's second video has a bar chart (Part 2: What Students Should Know about How People Learn). I suggest that you watch the whole video to understand the experiment, but the bar chart shows up around 3:10 minutes. Then, answer the same questions as above.

Answer

1. What kind of graph is in the video?
    1. This is a bar graph (notice that the bars are not touching).
2. What does the x-axis measure in the video?
    1. The x-axis is the one on the bottom, named “Level of Processing” and it has three levels (shallow, deep, and the control group). There are different colored bars, and they represent another variable (memory goal?) that has two levels, intentional and incidental.
3. What does the y-axis measure in the video?
    1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. However, in this video, the y-axis is showing the percentage of words that were successfully recalled (remembered).
4. What do you notice from the graph in the video? What pops out to you?
    1. What I noticed is that shallow processing (looking for the letter “e”) leads to worse recall than not having any processing goal (the control group). I also find it interesting that the intent to learn didn’t affect memory really at all. But you might have noticed other things!
5. What does the graph in the video make you wonder about?
    1. I wonder why deep processing (elaborating) isn’t more well known to help students?! But what does the graph make you wonder?
6. What is a catchy headline for the graph in the video?
    1. "Learning is Deep" seems catchy to me, but I bet you came up with something better!
7. How could you summarize the info in the graph in the video into one sentence?
    1. You could say something like, "Focusing on the meaning of words doesn’t seem to help you remember much more than just trying to remember information, but it is way better than shallow processing."
8. Who might want to know the information from the graph in the video?
    1. I imagine that both students and teachers would find this information very useful. Tutors and other learning professionals should also know about it, too.

Summary

Bar charts effectively portraying qualitative data. Bar charts are a good option when there are more than just a few categories, or for comparing two or more distributions. There's another whole mess of chart types to use on quantitative variables, but we'll cover line graphs and histograms next.

This page titled 6.3.1: Graphing Qualitative Variables- Bar Graphs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.3.2: Describing the Shape of Graphs of Quantitative Data

Statisticians and researchers look at a lot of frequency charts (usually line graphs and histograms), so they know what to look for. Khan Academy has a great video previewing what will be discussed in this section.

Skew

Since it's the more interesting of the two, let's start by talking about the skew.

The shape of a frequency chart (line graph or histogram) can tell you a lot of about that data set. The simplest things to look for are any extreme scores (called *outliers*) that seem to be much higher or much lower than most of the other scores. These outliers may affect the shape of the distribution, making it skewed. A skewed distribution is one in which many scores are bunched up to one side, and there are only a few scores on the other side. A distribution can be positively skewed (meaning that the scores are bunched to the left, and the thin tail is pointing to the right) or negatively skewed (meaning that the scores are bunched to the right, and the thin tail is pointing to the left). Figure **6.3.2.1** shows examples of a positively skewed line graph (on the right) and a negatively skewed line graph on the left.

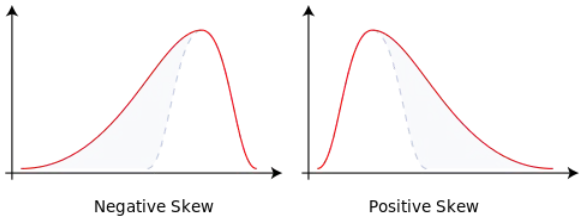


Figure **6.3.2.1**- Diagram of Skew. (CC-BY-SA Rodolfo Hermans (Godot), CC BY-SA 3.0, via Wikimedia Commons)

Skewness is basically a measure of asymmetry, and the easiest way to explain it is by drawing some pictures. As Figure **6.3.2.2** illustrates in histograms, 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), then we say that the data are *negatively skewed*. On the other hand, if there are more extremely large values than extremely small ones (right panel) we say that the data are *positively skewed*. That's the qualitative idea behind skewness.

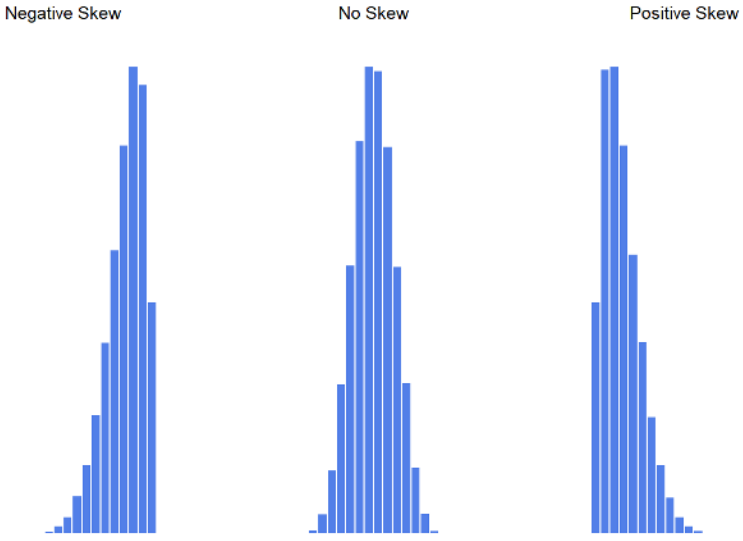


Figure **6.3.2.2**- Diagram of Skew and No Skew (CC-BY-SA Danielle Navarro from Learning Statistics with R)

Kurtosis

For distributions that are not skewed, you can look at how short and fat the distribution is, or if it is a medium bell-shaped curve, or if it is tall and narrow. This way of describing the shape of symmetrical (not-skewed) distributions, how broad the distribution is, is called kurtosis. Put simply, kurtosis is a measure of the "tailedness" of the data. As you can see in the line graph in Figure **6.3.2.3**,

there are three main types: wide, medium, or tall. Wide and flat graphs are called platykurtic. Medium, bell-shaped graphs are called mesokurtic or a normal distribution. Tall and narrow graphs are called leptokurtic.

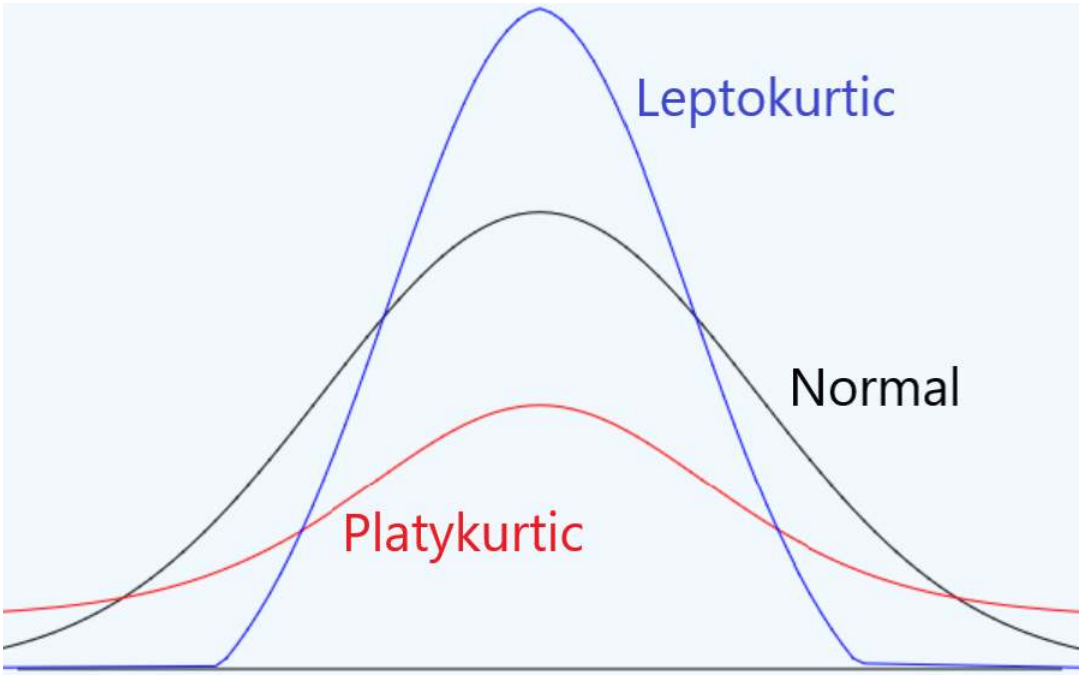


Figure **6.3.2.3**- Diagram of Three Kurtosis Types in Line Graphs (CC-BY Larry Green)

Figure **6.3.2.4** shows the three types of kurtosis in histograms. The histogram on the left shows platykurtic data, the middle histogram shows mesokurtic data, and on the right, we have a leptokurtic data set. By mathematical calculations, the "normal curve" (black lines) has zero kurtosis.

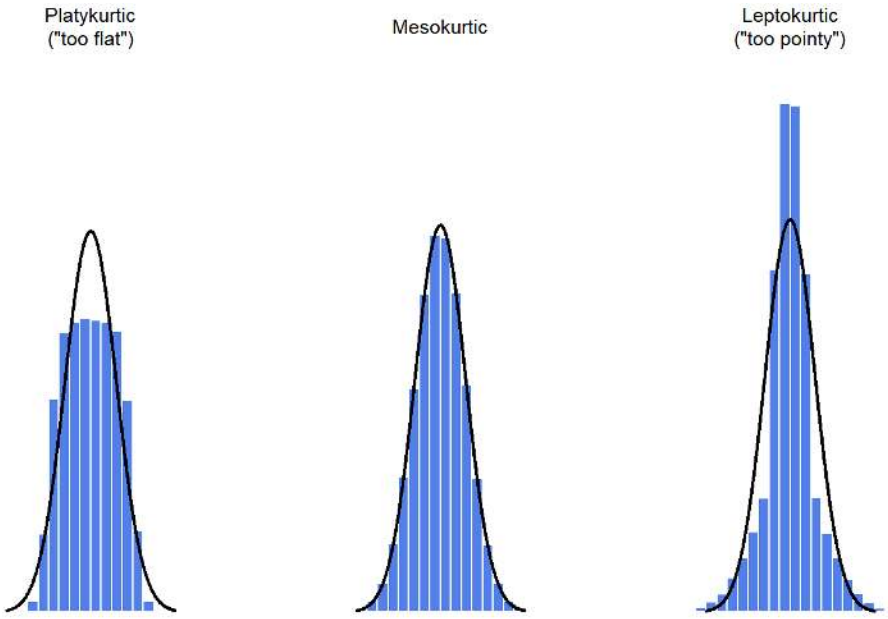


Figure **6.3.2.4**: Diagram of Three Kurtosis Types in Histograms (CC-BY-SA Danielle Navarro from Learning Statistics with R)

The types of kurtosis are summarized in Table **6.3.2.1**.

Table **6.3.2.1**- Informal Description of Types of Kurtosis

Technical Name	Informal Description
mesokurtic	just pointy enough
leptokurtic	too pointy
platykurtic	too flat

Summary

In practice, neither skew nor kurtosis is used anywhere near as frequently as the measures of central tendency and variability that will be discussed in next chapter. Skew is pretty important, though, so you do see it mentioned a fair bit; but I've actually never seen kurtosis reported in a scientific article to date

Contributors

- Danielle Navarro (University of New South Wales)
- Peter H. Westfall (Paul Whitfield Horn Professor and James and Marguerite Niver Professor, Texas Tech University)
- Dr. MO (Taft College)
- Figure **6.3.2.1** Image Attribution: Rodolfo Hermans (Godot), CC BY-SA 3.0, via Wikimedia Commons
- Figure **6.3.2.3** Image Attribution: CC-BY Larry Green from Lake Tahoe Community College

This page titled 6.3.2: Describing the Shape of Graphs of Quantitative Data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Danielle Navarro.

6.3.2.1: Graphing Quantitative Data- Line Graphs

Although line graphs are great at showing how data changes through time, we will mostly use line graphs to show frequency distributions; line graphs of frequency distributions are sometimes called *frequency polygons*, but that just reminds me of high school geometry. For example, the frequency table in Table 2.2.2 showed the frequency of final exam scores, a quantitative variable, while Figure **6.3.2.1.1** shows the same information but as a graph:

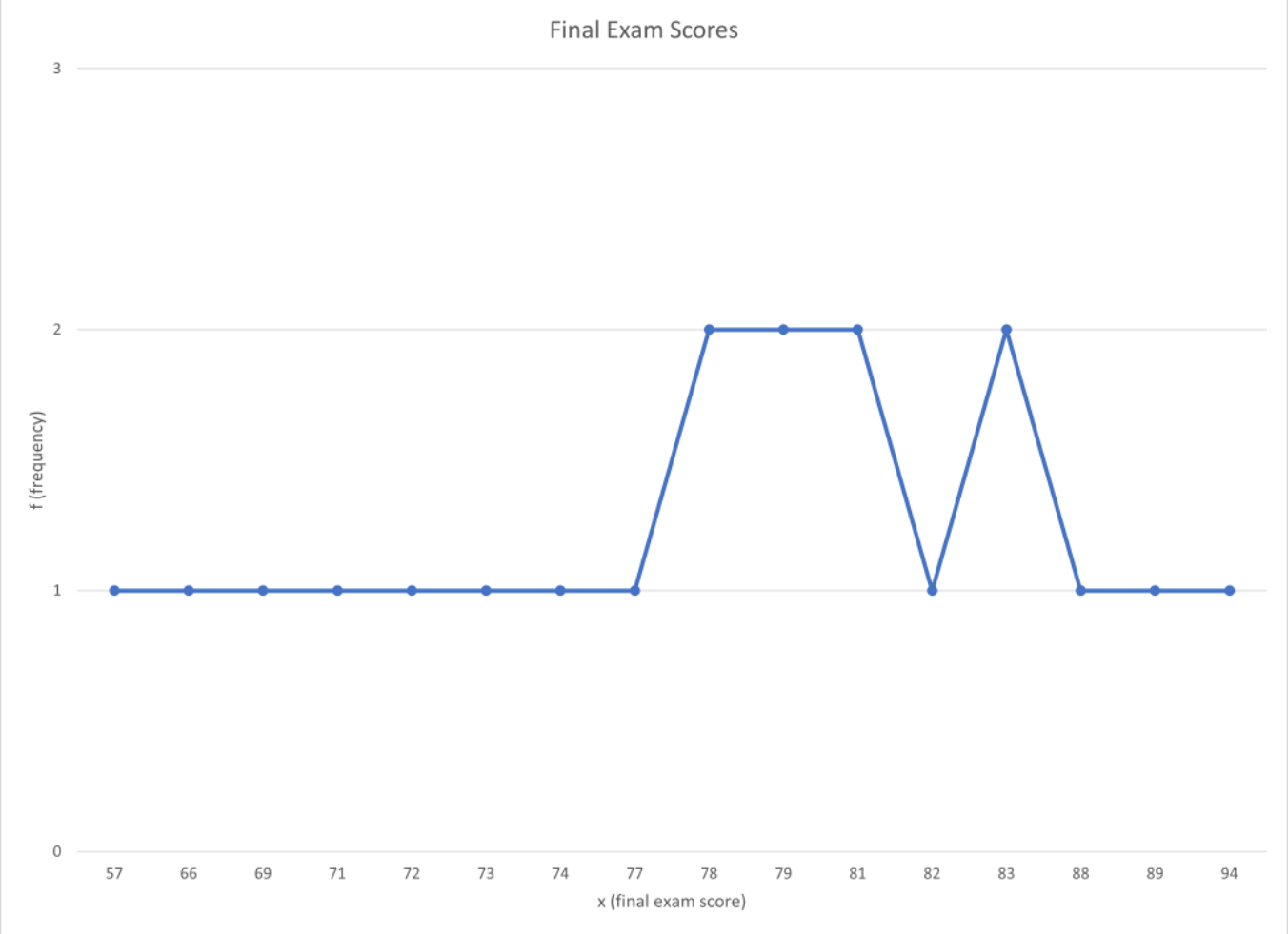


Figure **6.3.2.1.1**: Line Graph of Frequency of Final Exam Scores. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

Let's unpack Figure **6.3.2.1.1**, as well. First, the x-axis has the same information as the "x" column in Figure 2.1, each score that was earned on the final exam. The x-axis is always the one on the bottom going from left to right. Because we are talking about a numerical variable (ratio scale of measurement; quantitative variable), the information on the x-axis is *each score*. In a line graph, you did not create your own categories; for example, you would not create categories based on letter grades. The scores should be from lowest to highest. Second, the y-axis shows the same information as the "f" column in Figure 2.2.2, the number of students who earned that score. The y-axis is always going up and down. This chart was also created with a common commercial spreadsheet software (Excel), and there are again fancier types of line graph. And again, the fancier a chart looks, the more difficult it is to easily and quickly understand what the chart is showing.

Interpreting Line Graphs

I added a question to your interpretations.

Note

Can you see which question was added? Why do you think it was added for line charts?

1. What kind of graph is Figure **6.3.2.1.1**?
    1. This is a line graph. Line graphs are used to display quantitative data (or changes through time)
2. What does the x-axis measure in Figure **6.3.2.1.1**?
    1. The x-axis on line graphs show each score (or each time period). The x-axis on line graphs shows each time period. In a line graph showing changes through time, the x-axis will always be some marker of time.
3. What does the y-axis measure in Figure **6.3.2.1.1**?
    1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. However, the y-axis in Figure **6.3.2.1.2** shows the average percentages earned out of all points possible in the class each semester.
4. [New question!] Is Figure **6.3.2.1.1** skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
    1. Because the tail is so long on th left and there seems to be a pile of scores towards the right, I would say that this line graph is negatively skewed. This seems backwards to me, so I looked back at Figure 2.xxx (the picture of the skewed charts). Because it is skewed, it can't have kurtosis.
5. What do you notice from Figure **6.3.2.1.1**? What pops out to you?
    1. Two things that pop out to me. First, there is a bump of more frequent scores between 78 to 83 points. This suggests that most students in this unknown class earned a C+ or a B- on the Final Exam. The second thing I notice is that the left tail is pretty long. This means there are also quite a few students who earned low scores on the Final Exam.
6. What does Figure **6.3.2.1.1** make you wonder about?
    1. This chart makes me wonder if the professor ended up "curving" the exam since the highest score was 94 points, and that score was earned by only one student. The next highest score was 89 points, again with only one student earning that score on the Final Exam. If I haven't said this before, THIS IS NOT MY CLASS.
7. What is a catchy headline for Figure **6.3.2.1.1**?
    1. 85% of the Class Passed! [17 out of 20 students earned 70 points or higher]
8. How could you summarize the info in Figure **6.3.2.1.1** into one sentence?
    1. Almost everyone in the class passed the Final Exam, but only one student earned more than 90 points.
9. Who might want to know the information in Figure **6.3.2.1.1**?
    1. I would guess that students who are going to take this class from this professor might be interested. College administrators might also want to know so that they can see that most students are passing their classes.

Line Graphs Through Time

But for thoroughness, let's also look at a line graph that shows data through time. Figure **6.3.2.1.2** shows the percentage of points earn out of all points in Dr. MO's behavioral statistics courses in seven consecutive semesters (starting in Fall 2013, and ending in Fall 2016 when she started teaching a modified version of the course).

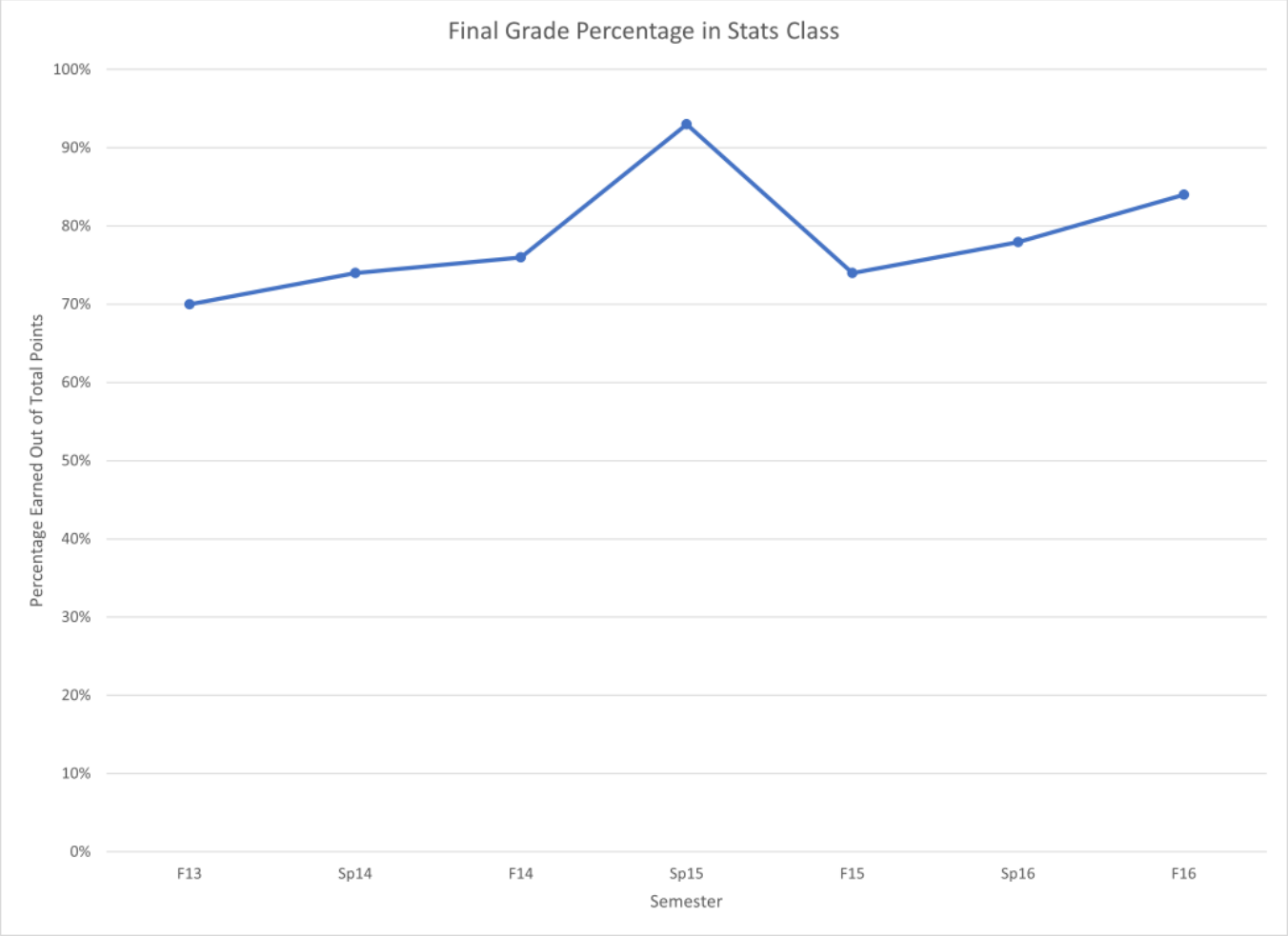


Figure **6.3.2.1.2**- Line Graph of Percentages Through Time (Copyright CC-BY-SA; Michelle Oja)

Unpacking Figure **6.3.2.1.2**, what do you see? First, let's compare Figure **6.3.2.1.2** to Figure **6.3.2.1.1**. In Figure **6.3.2.1.1**, the x-axis was each score, while in Figure **6.3.2.1.2** the x-axis is the time period. Similarly, in Figure **6.3.2.1.1**, the y-axis in the frequency, the number of students who earned that score. In Figure **6.3.2.1.2**, the y-axis is the percentage earned. Frequency distributions are really important for statistical analyses, but they aren't the only type of chart that you'll come across.

Interpreting Line Graphs

- What kind of graph is Figure **6.3.2.1.2**?
  - This is a line graph. Line graphs are used to display quantitative data (or changes through time)
- What does the x-axis measure in Figure **6.3.2.1.2**?
  - The x-axis on line graphs shows each time period. In a line graph showing changes through time, the x-axis will always be some marker of time.
- What does the y-axis measure in Figure **6.3.2.1.2**?
  - The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. However, the y-axis in Figure **6.3.2.1.2** shows the average percentages earned out of all points possible in the class each semester.
- Is Figure **6.3.2.1.2** skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
  - This question doesn't really make sense when the x-axis isn't scores (quantitative data points). In line graphs through time, we are looking for trends not kurtosis or skew.
- What do you notice from Figure **6.3.2.1.2**? What pops out to you?

- Spring 2015 was a great year for students!
- What does Figure **6.3.2.1.2** make you wonder about?
  - What happened in Spring 2015?! And why didn't it keep happening?
- What is a catchy headline for Figure **6.3.2.1.2**?
  - Average is Passing in Dr. MO's Behavioral Statistics Class
- How could you summarize the info in Figure **6.3.2.1.2** into one sentence?
  - The average percentages were all passing, with Spring 2015 being a particularly high average score.
- Who might want to know the information in Figure **6.3.2.1.2**?
  - I would guess that students who are going to take this class from this professor might be interested. College administrators might also want to know so that they can see that most students are passing this professor's class.

Another kind of graph that is used to show quantitative variables is a histogram, discussed next!

Contributors and Attributions

This page was extensively adapted by Michelle Oja (Taft College) from work by Matthew J. C. Crump (Brooklyn College of CUNY)

This page titled 6.3.2.1: Graphing Quantitative Data- Line Graphs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.3.2.2: Graphing Quantitative Data- Histograms

Another common type of graph for frequency distributions for quantitative variables (interval or ratio scales of measurement) is a histogram.

? Exercise 6.3.2.2.1

What other type of chart shows a frequency distribution for quantitative variables? Hint: We've already talked about it.

Answer

Line graph (sometimes called a frequency polygon)

? Exercise 6.3.2.2.1

Histograms look more like bar graphs, but are actually more like line graphs. Why?

Answer

Your answer probably talks about how the bars in bar charts don't touch, but the bars (or bins) in histograms do touch. Your answer should say something about the type of variable: quantitative variables (ratio or interval scale of measurements) should be graphed with histograms, and qualitative variables (nominal scale of measurement) should be graphed with bar charts.

Figure **6.3.2.2.1** shows a histogram (created by SPSS) of the frequency distribution of the final exam score data from Table 2.2.2 (and also shown in the line graph in Figure 2.8.1.

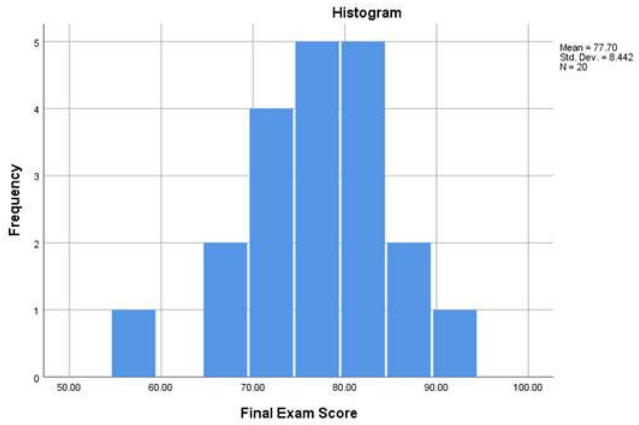


Figure **6.3.2.2.1**: Histogram of Frequency of Final Exam Scores. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

As with all frequency distributions, the y-axis of a histogram shows the frequency of each score (how many people had each score) and the x-axis has each score. Figure **6.3.2.2.1** shows the number of students who in each score category on the y-axis (Frequency), and labels the score categories on the x-axis (Final Exam Score, with each 10 points labeled). To make the histogram, we just count up the number of data points falling inside each bin, then plot those frequency counts as a function of the bins. Voila, a histogram.The 10-point categories are called bins. Bin #1 goes from about 55 to abou 59 points, bin #2 goes about 65 to 69 points, and so on until the last bin. The difference in the x-axis between a line graph and a histogram is that the scores are combined into ranges, or categories, in a histogram. This makes it easier to quickly understand what a graph is showing if there are a lot of different scores (like everyone's GPA in your class), but it does lose some accuracy.

Each bar (bin), and can be used to answer questions about the frequency of scores within these bin categories. For example, how many people scored between 90 to 100 points on the Final Exam? The seventh bar (or first on the right), the one between 90 and

100 on the x-axis, tells you how many. Look how tall that bar is. How tall is it? The height is shown on the y-axis, which provides the frequency. One person scored between 90 points and 100 points on the Final Exam.

Just like a line graph, a histogram can show the shape (kurtosis) or skew, and the range of our data set. The shape of the histogram refers to how it goes up and down. The shape tells us where the data is. For example, when the bars are low we know there isn't much data there. When the bars are high, we know there is more data there. So, where is most of the data? It looks like it's mostly in the middle two bins, between 75-85 points. The range of the data tells us the minimum (lowest score) and the maximum (highest score) of the data. In Figure **6.3.2.2.1**, most of the scores are between 65 points and 95 points, but the minimum scores is between 55-59 points.

When you make a histogram you get to choose how wide each bar will be. For example, below are four different histograms of the same data (a happiness survey). What changes is the width of the bins.

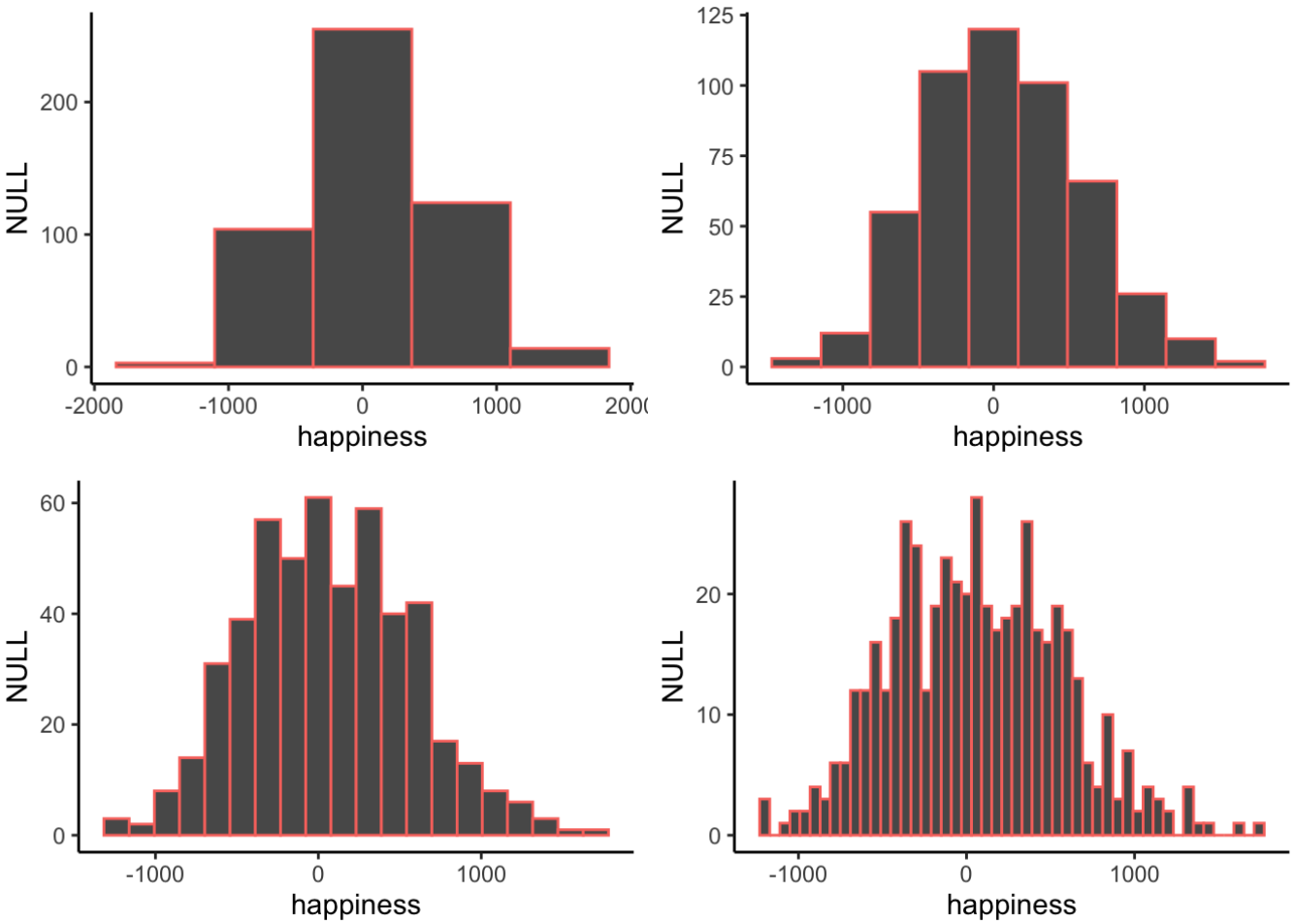


Figure **6.3.2.2.2**: Four histograms of the same data using different bin widths. (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

All of the histograms have roughly the same overall shape: From left to right, the bars start off small, then go up, then get small again. In other words, as the numbers get closer to zero, they start to occur more frequently. We see this general trend across all the histograms. But, some aspects of the trend fall apart when the bars get really narrow. For example, although the bars generally get taller when moving from -1000 to 0, there are some exceptions and the bars seem to fluctuate a little bit. When the bars are wider, there are less exceptions to the general trend. How wide or narrow should your histogram be? It's a Goldilocks question. Make it just right for your data.

Students often get confused between a bar graph and a histogram because they both have bars. There are two differences, however. First, in a histogram, the bars must touch; there should **not** be space between the categories that were created from the quantitative

data. This leads to the second difference between histograms and bar charts: Histograms show quantitative variables, while bar charts show qualitative variables (nominal scale of measurement).

Histogram Interpretation

The following are the same questions asked about the Figure 2.8.1, the line graph showing the frequency distribution of this same data. You might find it interesting how the answers change a little depending on how the data is presented.

1. What kind of graph is Figure **6.3.2.2.1**?
    1. This is a histogram. Histograms are used to display quantitative data, but it combines each score into groups. In Figure 2.xxx, the groups are 10-point ranges.
2. What does the x-axis measure in Figure **6.3.2.2.1**?
    1. The x-axis on histograms show categories based on each score. In Figure **6.3.2.2.1**, the software created 10-point categories for each Final Exam Score that a student earned.
3. What does the y-axis measure in Figure **6.3.2.2.1**?
    1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. In Figure **6.3.2.2.1**, that means that the y-axis shows how many students earned scores on the Final Exam in each 10-point range.
4. Is Figure **6.3.2.2.1** skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
    1. Because there is a gap between the lowest score category and the next one, I would say that this histrogram is negatively skewed. However, it looks less skewed than the same data in the line graph in Figure 2.8.1. In fact, just by looking at Figure **6.3.2.2.1** and not knowing the exact scores, I might have said that it is not skewed and that is has approximately normal kurtosis (mesokurtic).
5. What do you notice from Figure **6.3.2.2.1**? What pops out to you?
    1. Two things that pop out to me. First, there seemed to be the most scores around the middle of the distribution. The second thing I notice is that one column that's to the left. This shows that one student (frequency = 1) score below 60 points on the Final Exam.
6. What does Figure **6.3.2.2.1** make you wonder about?
    1. This histogram still makes me wonder if the professor ended up "curving" the exam so that the highest earning score was now like earning 100% on the Final Exam. If I haven't said this before, THIS IS NOT MY CLASS.
7. What is a catchy headline for Figure **6.3.2.2.1**?
    1. Final Exam Scores are Normal [This is catchy if you know what a Normal Distribution is; more in Ch. 4!]
8. How could you summarize the info in Figure **6.3.2.2.1** into one sentence?
    1. The class did really well on the Final Exam, although one student scored pretty low.
9. Who might want to know the information in Figure **6.3.2.2.1**?
    1. I still am guessing that students who are going to take this class from this professor might be interested. College administrators might also want to know so that they can see that most students are passing their classes.

Summary

Making a histogram is our first act of officially *summarizing* the data. We are no longer look at the individual bits of data, instead we will see how the numbers group together. Let's look at a histogram of the happiness data, and then explain it.

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 6.3.2.2: Graphing Quantitative Data- Histograms is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **2.2: Look at the data** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

6.3.3: Graphing Quantitative Data- Scatterplots

Learning Objectives

1. Recognize when to use a scatterplot.
2. Identify the general relationships between variables shown in scatterplots.
3. Interpret what a scatterplot shows about the relationship between the variables shown in the scatterplot.

So far, we've mostly covered charts that show how many people were in each category or earned each score. Scatterplots are a different kind of chart that can show you about how many people scored each variable, but the real purpose is show a person's score on *two* quantitative variables (ratio or interval scale of measurement) at once. In scatterplots, the both the x-axis and the y-axis are scores on a quantitative variable, and there is no direct measurement of frequency. Figure **6.3.3.1** shows an example with GPA and the hours that the student studied (the GPA_Study_Hours file from OpenIntro.org). Each dot represents a student, and shows both their GPA and how many hours they studied. Find the dot that's most right; that student who spent 70 hours studying, and did not quite earn a 4.0. In fact, it looks like no one who studied over 35 hours earned a 4.0; what do you think that can tell you about time spent studying? On the other hand, only a few people who spent less than 20 hours studying earned a 4.0; what can *that* tell you about how the time you spend studying leads to your GPA?

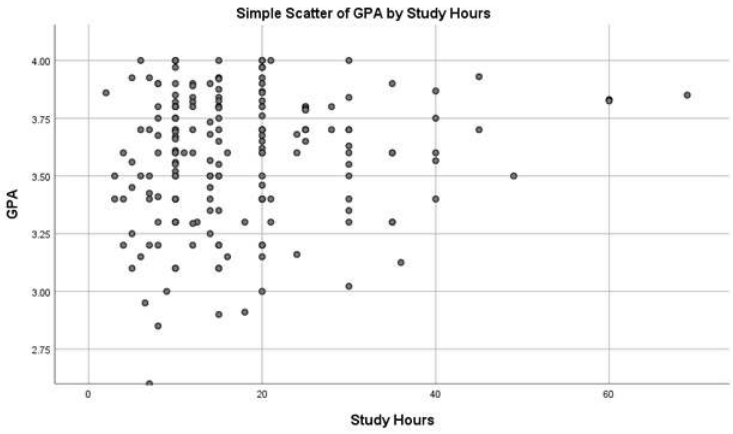


Figure **6.3.3.1**: Scatterplot of Study Hours and GPA. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

Remember IVs and DVs? In scatterplots, the cause (IV, but referred to as a predictor for correlations) is usually plotted on the x-axis, and the effect (DV, usually called the outcome in correlations) is plotted on the y-axis. What we are looking for is: what happens to the DV when the IV increases? In Figure **6.3.3.1**, as Study Hours increase, GPA tends to increase. However, this isn't a strong relationship because there are a couple students who spend a lot of time studying and don't get the GPA payoff, while there are even more students who spend 5-10 hours studying but still earn good grades. In general, however, it looks like spending more time studying is related to higher GPAs.

Note

Use this Statistics How To to practice making a scatterplot. Once you can make one, with the IV on the bottom and the DV on the side, you'll better understand what they can show.

Scatterplot Interpretation

1. What kind of graph is Figure **6.3.3.1**?
    1. This is a scatterplot. Scatterplots are used to plot two quantitative variables for each person. In Figure **6.3.3.1**, the two variables are Study Hours and GPA.
2. What does the x-axis measure in Figure **6.3.3.1**?

1. The x-axis in Figure **6.3.3.1** shows the hours that each person studied. Lower hours of studying are on the left, and higher hours of studying are on the right.

3. What does the y-axis measure in Figure **6.3.3.1**?
    1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. However, the y-axis of the scatterplot in Figure **6.3.3.1** shows another quantitative variable, GPA of each person. Lower GPAs are towards the bottom of the axis, and higher GPAs are towards the top.
4. Is Figure **6.3.3.1** skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
    1. Trick question! Because scatterplots do not show frequency distributions, they aren't described as skewed and they don't have kurtosis. Instead, we look to see if there is a general trend; as one variable goes up, does the other go up, go down, or are they unrelated? It does look like more studying is related to higher GPAs, but this isn't quite as strong of a relationship as we'd like to see.
5. What do you notice from Figure **6.3.3.1**? What pops out to you?
    1. What I first noticed is that a lot of students study 20 hours or less, and still have high GPAs. I also noticed that there are lots of scores in 5-hour increments (e.g., 5 hours, 10 hours, 15 hours, etc.). This suggests that students were estimated the number of hours that they studied, and rounding to the nearest 5-hour mark.
6. What does Figure **6.3.3.1** make you wonder about?
    1. I wonder if the results would change if the hours studied were measured more accurately, rather than having students estimate at a later time? I also wonder what was going with those two students who studied more than 60 hours (and still didn't earn a 4.0). What were they doing that they considered "studying"? Or is this a data-entry mistake? Like, did someone say that they studied 6 hours but it was entered as 60 hours?
7. What is a catchy headline for Figure **6.3.3.1**?
    1. Study Less and Still Get a 4.0!
8. How could you summarize the info in Figure **6.3.3.1** into one sentence?
    1. A lot of students studied about 5 to 20 hours and earned GPAs above 3.0.
9. Who might want to know the information in Figure **6.3.3.1**?
    1. I still am guessing that students who might be interested. Maybe tutors and folks working in a Learning Center would find this useful?

Remember those videos about the best way to learn from Dr. Chew, a cognitive psychologist? First discussed in 1.3: The Scientific Method and all 5 linked in 1.8 (What can you do to learn material for this class?). Well, Dr. Chew's first video has a scatter plot (Part 1: Beliefs that Make You Fail)! I suggest that you watch the whole video, but the experiment and scatterplot show up around 4:30 minutes.

? Exercise 6.3.3.1

Watch Dr. Chew's first video (Part 1: Beliefs that Make You Fail), then answer the same set of questions. I suggest that you watch the whole video, but the experiment and scatterplot show up around 4:30 minutes.

Answer

1. What kind of graph is in Dr. Chew's Beliefs that Make you Fail video?
    1. This is a scatterplot. Scatterplots are used to plot two quantitative variables for each person. In the scatterplot in the video, the two variables are Estimated Percentage Correct and Actual Percentage Correct on an exam.
2. What does the x-axis measure in the scatterplot in Dr. Chew's Beliefs that Make you Fail video?
    1. The x-axis in the scatterplot in Dr. Chew's Beliefs that Make you Fail video shows the percentage that students thought that they got correct on the exam. Lower scores are on the left, and higher percentages are on the right.
3. What does the y-axis measure iin the scatterplot in Dr. Chew's Beliefs that Make you Fail video?

1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. However, the y-axis of the scatterplot in the scatterplot in Dr. Chew's Beliefs that Make you Fail video shows another quantitative variable, the actual percentage correct on the exam. Lower scores are towards the bottom of the axis, and higher percentages are towards the top.

4. Is the scatterplot in Dr. Chew's Beliefs that Make you Fail video skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
    1. Trick question! Because scatterplots do not show frequency distributions, they aren't described as skewed and they don't have kurtosis. Instead, we look to see if there is a general trend; as one variable goes up, does the other go up, go down, or are they unrelated? It does look like most students estimated their percentage correct pretty close to their actual percentage correct, but were a few students who estimated that they did way better than they actually did. Dr. Chew says that these students have low meta-cognition, meaning that they are unclear on what material they actually know.
5. What do you notice from the scatterplot in Dr. Chew's Beliefs that Make you Fail video? What pops out to you?
    1. What I first noticed was that most students' estimates were pretty close to their actual percentages earned. Yay, most students had good meta-cognition!
6. What does the scatterplot in Dr. Chew's Beliefs that Make you Fail video make you wonder about?
    1. I wonder if students get better at estimating what they know throughout the semester?
7. What is a catchy headline for the scatterplot in Dr. Chew's Beliefs that Make you Fail video?
    1. Meta-Cognition: Predicting Your Future!
8. How could you summarize the info in the scatterplot in Dr. Chew's Beliefs that Make you Fail video into one sentence?
    1. Most students' estimates of their exam grades were relatively close to their actual exam grades, but there were a few students who badly overestimated their scores.
9. Who might want to know the information in the scatterplot in Dr. Chew's Beliefs that Make you Fail video?
    1. I still am guessing that students might be interested. Maybe tutors and folks working in a Learning Center would find this useful?

Positive or Negative?

Scatterplots can show when there is a linear relationship., meaning that the two variables vary together and when plotted look like a straight line. Scatterplots can show positive linear relationships, negative linear relationships, or show that there is no relationship between the two variables.

- Positive linear relationship: When one variable goes up, the other goes up.
    - Positive *doesn't* mean good!! It means that the two variables change in the **same** direction.
- Negative linear relationship: When one variable goes up, the other variable goes down.
    - Negative *doesn't* mean bad!!! It means that the two variables change in **different** directions.

Figure **6.3.3.2** shows a scatterplot for hypothetical scores on Job Satisfaction (x-axis) and Well-Being of the worker (on the y-axis). We can see from the axes that each of these variables is measured on a 10-point scale, with 10 being the highest on both variables (high satisfaction and good health and well-being) and 1 being the lowest (dissatisfaction and poor health). This scale suggests an interval scale of measurement for both variables, which means that there are two quantitative variables.

When we look at this plot, we can see that the variables do seem to be related. The higher scores on Job Satisfaction tend to also be the higher scores on Well-Being, and the same is true of the lower scores.

Figure **6.3.3.2** demonstrates a positive relation. As Job Satisfaction scores increase, Well-Being scores also tend to increase. Although this is not a perfect relation (if it were, the points would form a single straight line), it is nonetheless very clearly positive. This is one of the key benefits to scatterplots: they make it very easy to see the direction of the relation.

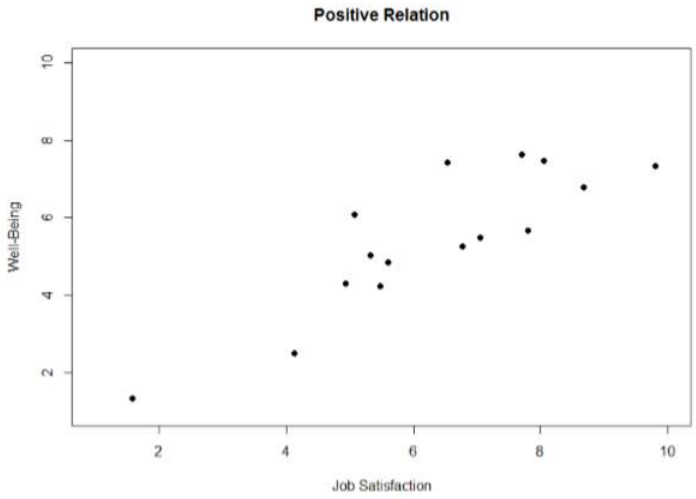


Figure **6.3.3.2**: Scatterplot of Job Satisfaction & Well-Being (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Figure **6.3.3.3**: Scatterplot of Job Satisfaction & Burnout (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

As another example, Figure **6.3.3.3** shows a negative relation between Job Satisfaction on the x-axis and Burnout on the y-axis. As we can see from Figure **6.3.3.3**, higher scores on Job Satisfaction tend to correspond to lower scores on Burnout, which is how stressed, un-energetic, and unhappy someone is at their job. As with Figure **6.3.3.2**, this is not a perfect relation, but it is still a clear one. As these figures show, points in a positive relation moves from the bottom left of the plot to the top right, and points in a negative relation move from the top left to the bottom right.

Or None?

Scatterplots can also indicate that there is no relation between the two variables. In these scatterplots (an example is shown below in Figure **6.3.3.4** plotting Job Satisfaction and Job Performance) there is no interpretable shape or line in the scatterplot. The points appear randomly throughout the plot. If we tried to draw a straight line through these points, it would basically be flat. The low scores on job satisfaction have roughly the same scores on job performance as do the high scores on job satisfaction. Scores in the middle or average range of job satisfaction have some scores on job performance that are about equal to the high and low levels and some scores on job performance that are a little higher, but the overall picture is one of inconsistency.

As we can see, scatterplots are very useful for giving us an approximate idea of whether or not there is a relation between the two variables and, if there is, if that relation is positive or negative. They are also useful for another reason: they are the only way to determine one of the characteristics of correlations that are discussed next: form.

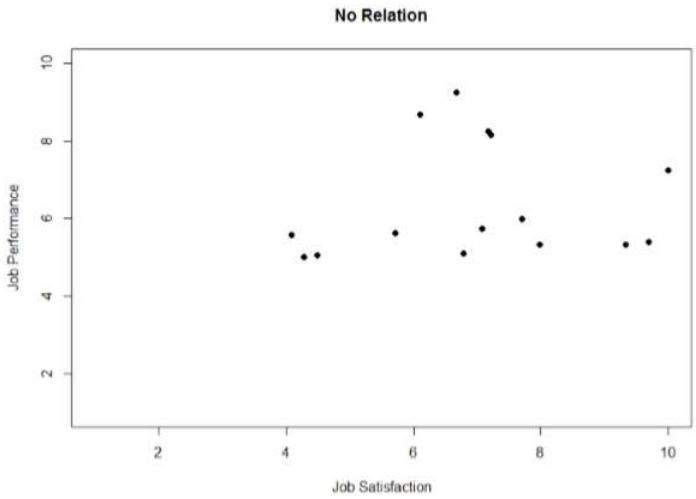


Figure **6.3.3.4**- Scatterplot of Job Satisfaction and Job Performance (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Curvilinear Relationships

A linear relation is what we saw in Figure **6.3.3.2** and Figure **6.3.3.3**. If we drew a line through the middle points in the any of the scatterplots, we would be best suited with a straight line. The term "linear" comes from the word "line". The relation between two variables can also be curvilinear. As the name suggests, a curvilinear relation is one in which a line through the middle of the points will be curved rather than straight. Two examples are presented in Figures **6.3.3.5** and **6.3.3.6**.

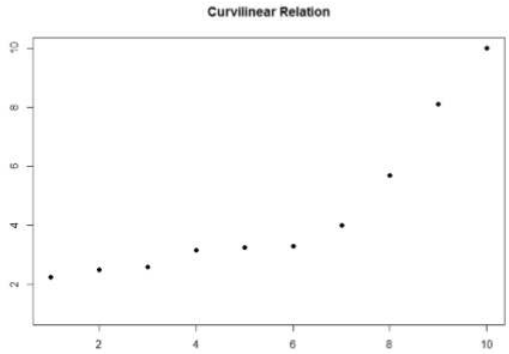


Figure **6.3.3.5**: Curvilinear Scatterplot with an Elbow (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

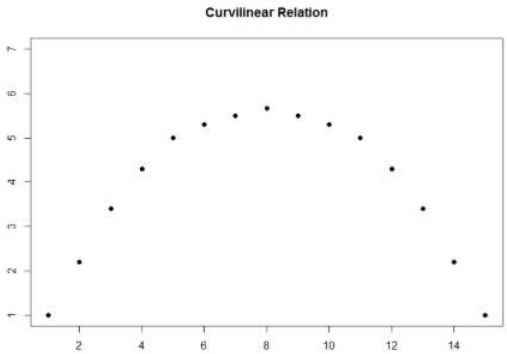


Figure **6.3.3.6**: Curvilinear Scatterplot with an Inverted-U (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics).

Curvilinear relations can take many shapes, and the two examples above are only a small sample of the possibilities. What they have in common is that they both have a very clear pattern but that pattern is not a straight line.

Sometimes when we look at scatterplots, it is tempting to get biased by a few points that fall far away from the rest of the points and seem to imply that there may be some sort of relation. These points are outliers, which we discussed before.

Contributors and Attributions

- Template:ContribIntrotoPsyStatsFoster
- Template:ContribOja

This page titled 6.3.3: Graphing Quantitative Data- Scatterplots is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.3.4: Honesty in Graphing

Intentional Lies or Accidental Misleading?

It’s easy to mislead with graphs. Some of the fancy graphics make it more difficult to see the exact numbers, but there are many ways to be misled with graphs. To avoid being fooled by a fancy chart you should also check out the axes. In a frequency distribution, the y-axis should generally be one or zero. This article “Lying with Graphs” (website address: http://chmullig.com/2011/04/lying-with-graphs/) by Chris Mulligan in 2011 shows an example of both a confusing type of chart, and mis-leading axes (plus, he didn’t think that the original article calculated the actual numbers correctly, either); always be vigilant!

Example of Misleading

We will use data from interviews of computer users to show how graphs in general, but especially bar graphs, can be misleading. Here's the research scenario: 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? Was it purchased by newcomers to the computer market> Were previous Windows users 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.

Figure **6.3.4.1** shows that three-dimensional bars are usually not as effective as two-dimensional bars because it's unclear where the top of the bar is.

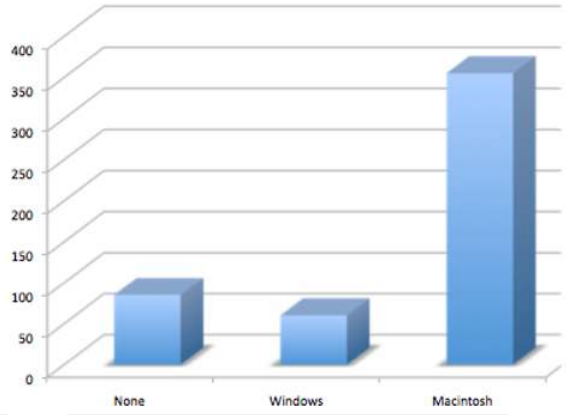


Figure **6.3.4.1**: A three-dimensional bar chart (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure **6.3.4.2** presents iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure **6.3.4.2** is misleading because the viewer's attention will be captured by the area (width) of the computers, not just the height. 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 **6.3.4.2** 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 **6.3.4.2**! Edward Tufte 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.

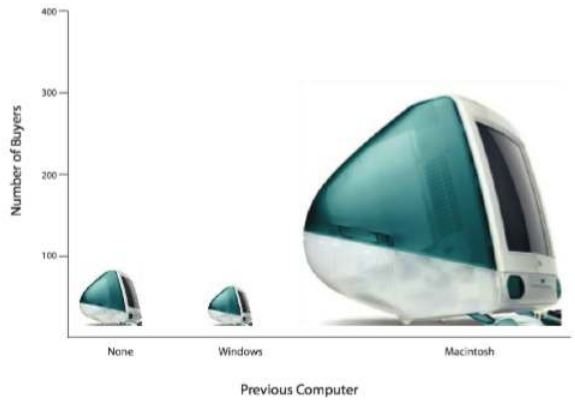


Figure **6.3.4.2**: Example with a lie factor greater than 8. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Another distortion in bar charts results from setting the bottom of the y-axis to a something other than zero. The bottom of the Y-axis should represent the least number of cases that could have occurred in a category. Normally, this number should be zero. Figure **6.3.4.3** shows the iMac data with a baseline of 50. Once again, the differences in areas suggests a different story than the true differences in percentages. The number of Windows-switchers seems minuscule compared to its true value of 12%.

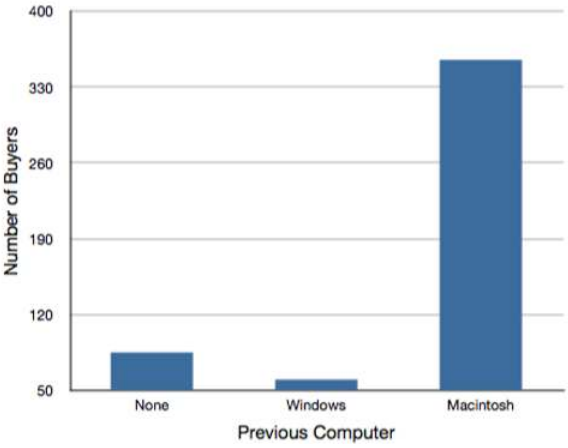


Figure **6.3.4.3**: Example with a baseline of 50. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Note

Why don't you use this iMac data to play around and create your own bar chart by hand or with computer software, without making it misleading?

Examples of Misleading, or Lies?

Example 6.3.4.1

Go through the examples from this 2014 BuzzFeed report by Katie Notopoulos (https://www.buzzfeednews.com/article/katienotopoulos/graphs-that-lied-to-us), and decide one thing that is wrong with each chart.

1. The time an upside down y-axis made "Stand Your Ground" seem much more reasonable.
2. The time 7 million was 5x more than 6 million.
3. The Govenor race where one guy's 37% was WAY more than just 37%
4. This bar graph that shows the devastating drop in this pitcher's speed after one year.

5. The time when Scotland really gave 110%.

6. This poll which gives about the same infuriating response as when you ask someone what they want for dinner.

7. The pollster who really doesn't want people to think A levels are getting harder.

8. This graph that measure..... units of "innovation"?

9. The number skipping y-axis on this totally legit graph.

10. This graph showing the astonishing growth rate to .06 hawks?

11. Whatever happened in this school paper.

12. The graph where last year, last week, and today are equally far apart.

13. This chart showing the giant gulf between 35% and 39.6%.

Solution

1. Check out the y-axis. Instead of zero on the bottom, with the higher numbers being higher, the y-axis is “upside down” from what is typical.

2. Check out the y-axis. Not only is it not labeled, but if the y-axis had started with zero, the 1 million difference would have looked like much.

3. The two 37%’s probably look difference chart doesn’t show decimal points, but if the y-axis had started with zero, the two 37%’s would be indistinguishable.

4. It’s the missing y-axis problem again, and not starting the y-axis at zero.

5. There’s nothing technically wrong with these numbers, but you should always have someone else check your math.

6. There’s nothing technically wrong with these numbers, but it is a reminder that designing survey questions is harder than it seems!

7. This is an appropriate way to use a pie chart (showing percentages that add up to close to 100%), but the pie slices need to represent the percentage, so the blue should only be 49% of the pie. It looks like 60% or about two-thirds…

8. It’s important to label your y-axis with units that make sense!

9. The y-axis should have equal intervals, meaning that the difference between the first number and the second number (2 points in this example, from 1 to 3) is the same as the difference between the highest number and the second highest number (which is 1803-516 = 1287 in this example).

10. There is nothing technically wrong with this chart, but even going from zero to 1.0 would have seemed more accurate when dealing with such small numbers on the y-axis.

11. The colors in the pie charts should take up the portion of the pie that the numbers show. Also, pie charts show 100% of the sample, but 26% + 26% + 26% +26% = 104%, so someone needed a friend to check their math here, as well.

12. Here is an example of a line graph showing changes through time, except that the x-axis doesn’t have equal intervals. It should probably have been a bar chart showing the three time categories, rather than trying to show a change through time.

13. Another example of a missing y-axis that is misleading because the missing y-axis doesn’t start at zero.

You can also read this great 2014 blog by Ravi Parikh of Heap showing some of the same examples of graphs that seem to be trying to intentionally mislead readers, and an explanation about what is inappropriate or missing: https://heap.io/blog/data-stories/how-to-lie-with-data-visualization

Be careful to avoid misleading your audience, or being misled!

Graphs Summary

You've learned about many types of graphs, and why they should be used. You also should have practiced interpreting what they can tell you.

Example 6.3.4.1

Play around with this interactive website (website address: https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm), then answer what kind of graphs they are.

Solution

Although both have lines to show a Normal Distribution, the actual graphs of the data are histograms because they are bars that touch, and the bars show the frequency of different quantitative variables.

Now that you've learned so much about charts and graphs, go through these charts from the New York Times (website address: https://www.nytimes.com/column/whats-going-on-in-this-graph) and determine if the x-axis is showing qualitative or quantitative variables and if you can describe the shape of any frequency distributions.

Contributors and Attributions

- Foster et al. (University of Missouri-St. Louis, Rice University, & University of Houston, Downtown Campus)
- Dr. MO (Taft College)

This page titled 6.3.4: Honesty in Graphing is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

6.4: Understanding Null Hypothesis Significance Testing

Learning Objectives

1. Explain the purpose of null hypothesis significance testing, including the role of sampling error.
2. Describe the basic logic of null hypothesis significance testing.
3. Describe the role of relationship strength and sample size in determining statistical significance and make reasonable judgments about statistical significance based on these two factors.

The Purpose of Null Hypothesis Testing

As we have seen, psychological research typically involves measuring one or more variables in a sample and computing descriptive summary data (e.g., means, correlation coefficients) for those variables. These descriptive data for the sample are called statistics. In general, however, the researcher's goal is not to draw conclusions about that *sample* but to draw conclusions about the *population* that the sample was selected from. Thus researchers must use sample statistics to draw conclusions about the corresponding values in the population. These corresponding values in the population are called parameters. Imagine, for example, that a researcher measures the number of depressive symptoms exhibited by each of 50 adults with clinical depression and computes the mean number of symptoms. The researcher probably wants to use this sample statistic (the mean number of symptoms for the sample) to draw conclusions about the corresponding population parameter (the mean number of symptoms for adults with clinical depression).

Unfortunately, sample statistics are not perfect estimates of their corresponding population parameters. This is because there is a certain amount of random variability in any statistic from sample to sample. The mean number of depressive symptoms might be 8.73 in one sample of adults with clinical depression, 6.45 in a second sample, and 9.44 in a third—even though these samples are selected randomly from the same population. Similarly, the correlation (Pearson's r) between two variables might be +.24 in one sample, −.04 in a second sample, and +.15 in a third—again, even though these samples are selected randomly from the same population. This random variability in a statistic from sample to sample is called sampling error. (Note that the term *error* here refers to random variability and does not imply that anyone has made a mistake. No one "commits a sampling error.")

One implication of this is that when there is a statistical relationship in a sample, it is not always clear that there is a statistical relationship in the population. A small difference between two group means in a sample might indicate that there is a small difference between the two group means in the population. But it could also be that there is no difference between the means in the population and that the difference in the sample is just a matter of sampling error. Similarly, a Pearson's r value of −.29 in a sample might mean that there is a negative relationship in the population. But it could also be that there is no relationship in the population and that the relationship in the sample is just a matter of sampling error.

In fact, any statistical relationship in a sample can be interpreted in two ways:

- There is a relationship in the population, and the relationship in the sample reflects this.
- There is no relationship in the population, and the relationship in the sample reflects only sampling error.

The purpose of null hypothesis testing is simply to help researchers decide between these two interpretations.

The Logic of Null Hypothesis Significance Testing

Null hypothesis significance testing (NHST) is a formal approach to deciding between two interpretations of a statistical relationship in a sample. One interpretation is called the null hypothesis. This is the idea that there is no relationship in the population and that the relationship in the sample reflects only sampling error. Informally, the null hypothesis is that the sample relationship "occurred by chance." The other interpretation is called the research hypothesis (often called an alternate hypothesis). This is the idea that there is a relationship in the population and that the relationship in the sample reflects this relationship in the population.

Again, every statistical relationship in a sample can be interpreted in either of these two ways: It might have occurred by chance, or it might reflect a relationship in the population. So researchers need a way to decide between them. Although there are many specific null hypothesis testing techniques, they are all based on the same general logic. The steps are as follows:

- Assume for the moment that the null hypothesis is true. There is no relationship between the variables in the population.
- Determine how likely the sample relationship would be if the null hypothesis were true.
- If the sample relationship would be extremely unlikely, then **reject the null hypothesis** in favor of the alternative hypothesis. If it would not be extremely unlikely, then **retain the null hypothesis**.

Following this logic, we can begin to understand why Mehl et al. (2007) concluded that there is no difference in talkativeness between women and men in the population. In essence, they asked the following question: "If there were no difference in the population, how likely is it that we would find a small difference of $d = 0.06$ in our sample?" Their answer to this question was that this sample relationship would be fairly likely if the null hypothesis were true. Therefore, they retained the null hypothesis—concluding that there is no evidence of a sex difference in the population. We can also see why Kanner et al. (1981) concluded that there is a correlation between hassles and symptoms in the population. They asked, "If the null hypothesis were true, how likely is it that we would find a strong correlation of +.60 in our sample?" Their answer to this question was that this sample relationship would be fairly unlikely if the null hypothesis were true. Therefore, they rejected the null hypothesis in favor of the alternative hypothesis—concluding that there is a positive correlation between these variables in the population.

A crucial step in null hypothesis significance testing is finding the probability of the sample result or a more extreme result if the null hypothesis were true. This probability is called the *p*-value. A low *p* value means that the sample or more extreme result would be unlikely if the null hypothesis were true and leads to the rejection of the null hypothesis. A *p* value that is not low means that the sample or more extreme result would be likely if the null hypothesis were true and leads to the retention of the null hypothesis. But how low must the *p* value criterion be before the sample result is considered unlikely enough to reject the null hypothesis? In null hypothesis testing, this criterion is almost always set to .05. If there is a 5% chance or less of a result at least as extreme as the sample result if the null hypothesis were true, then the null hypothesis is rejected. When this happens, the result is said to be statistically significant. If there is greater than a 5% chance of a result as extreme as the sample result when the null hypothesis is true, then the null hypothesis is retained (or failed to reject). This does not necessarily mean that the researcher *accepts* the null hypothesis as true—only that there is not *currently* enough evidence to reject it. Researchers often use the expression "fail to reject the null hypothesis" rather than "retain the null hypothesis," but they never use the expression "accept the null hypothesis."

The Misunderstood p Value

The *p* value is one of the most misunderstood quantities in psychological research (Cohen, 1994). Even professional researchers misinterpret it, and it is not unusual for such misinterpretations to appear in statistics textbooks!

The most common misinterpretation is that the *p* value is the probability that the null hypothesis is true—that the sample result occurred by chance. For example, a misguided researcher might say that because the *p* value is .02, there is only a 2% chance that the result is due to chance and a 98% chance that it reflects a real relationship in the population. But this is *incorrect*. The *p* value is really the probability of a result at least as extreme as the sample result *if* the null hypothesis *were* true. So a *p* value of .02 means that if the null hypothesis were true, a sample result this extreme would occur only 2% of the time.

You can avoid this misunderstanding by remembering that the *p* value is not the probability that any particular *hypothesis* is true or false. Instead, it is the probability of obtaining the *sample result* if the null hypothesis were true.



"Null Hypothesis" retrieved from http://imgs.xkcd.com/comics/null_hypothesis.png (CC-BY-NC 2.5)

Role of Sample Size and Relationship Strength

Recall that NHST involves answering the question, “If the null hypothesis were true, what is the probability of a sample result as extreme as this one?” In other words, “What is the *p* value?” It can be helpful to see that the answer to this question depends on just two considerations: the strength of the relationship and the size of the sample. Specifically, the stronger the sample relationship and the larger the sample, the lower the *p* value. This should make sense. Imagine a study in which a sample of 500 English majors is compared with a sample of 500 math majors in terms of some academic characteristic. If there were differences by major in the population of all college students, then we'd expect a small p-value. Even if our sample had a few wonky people who seemed more like the other major, we'd still expect a small p-value. However, what if we compared 6 English majors with 5 math majors? Any differences between the groups in the population would have to be really strong to compensate for the fact that a couple mis-matched people would greatly affect the results.

Of course, sometimes the result can be weak and the sample large, or the result can be strong and the sample small. In these cases, the two considerations trade off against each other so that a weak result can be statistically significant if the sample is large enough and a strong relationship can be statistically significant even if the sample is small. In general, researchers aim to have at lest 30 participants *in each group* to find moderately-strong effects. Weak relationships based on medium or small samples are never statistically significant, and strong relationships based on medium or larger samples are always statistically significant. This means that you'd want 30 English majors and 30 math majors if you were looking at an academic characteristics that moderately affected your outcome. However, if you had an academic characteristics that had a really weak effect, you'd want 100 to 500 participants per group.

If you keep this lesson in mind, you will often know whether a result is statistically significant based on the descriptive statistics alone. It is extremely useful to be able to develop this kind of intuitive judgment. One reason is that it allows you to develop expectations about how your null hypothesis significance tests may come out, which in turn allows you to detect problems in your analyses. For example, if your sample relationship is strong and your sample is medium, then you would expect to reject the null hypothesis. If for some reason your statistical results indicates otherwise, then you need to double-check your computations and interpretations. A second reason is that the ability to make this kind of intuitive judgment is an indication that you understand the basic logic of this approach in addition to being able to do the computations.

One important note: This is just about finding statistically significant effects! If you want external validity, or to be able generalize to a target population, you'd need to have a sample with the important characteristics for that population. For example, if we were interested in English or math majors only at community colleges, then you'd want a sample with participants that represent the characteristics of the population of community college students. Research often doesn't have this, but we pretend that the results from

Statistical Significance Versus Practical Significance

While we're talking about statistically significant effect, let's remember that a rejecting the null hypothesis does not necessarily mean that an intervention had a strong effect. Even a very weak result can be statistically significant if it is based on a large enough sample. This is closely related to Janet Shibley Hyde’s argument about sex differences (Hyde, 2007). The differences between women and men in mathematical problem solving and leadership ability are statistically significant. But the word *significant* can cause people to interpret these differences as strong and important—perhaps even important enough to influence the college courses they take or even who they vote for. As we have seen, however, these statistically significant differences are actually quite weak—perhaps even “trivial.”

This is why it is important to distinguish between the *statistical* significance of a result and the *practical* significance of that result. Practical significance refers to the importance or usefulness of the result in some real-world context. Many sex differences are statistically significant—and may even be interesting for purely scientific reasons—but they are not practically significant. In clinical practice, this same concept is often referred to as “clinical significance.” For example, a study on a new treatment for social phobia might show that it produces a statistically significant positive effect. Yet this effect still might not be strong enough to justify the time, effort, and other costs of putting it into practice—especially if easier and cheaper treatments that work almost as well already exist. Although statistically significant, this result would be said to lack practical or clinical significance.

References

Cohen, J. (1994). The world is round: $p < .05$. *American Psychologist*, *49*, 997–1003.

Hyde, J. S. (2007). New directions in the study of gender similarities and differences. *Current Directions in Psychological Science*, *16*, 259–263.

Kanner, A. D., Coyne, J. C., Schaefer, C., & Lazarus, R. S. (1981). Comparison of two modes of stress measurement: Daily hassles and uplifts versus major life events. *Journal of Behavioral Medicine*, *4*, 1–39.

Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, *317*, 82.

This page titled 6.4: Understanding Null Hypothesis Significance Testing is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **13.2: Understanding Null Hypothesis Testing** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

6.5: Comparing Means

Learning Objectives

1. Interpret one-sample, dependent-samples, and independent-samples *t*- tests.
2. Interpret the results of one-way and repeated measures ANOVAs.
3. Identify when a design would need a factorial ANOVA.

The emphasis here is on providing enough information to remember what you learned in your statistics course to be able to interpret common statistical analyses to compare means. For details on how to calculate these analyses, and when to use each one, review one of the openly-licensed textbooks on statistics on LibreTexts; textbooks for some disciplines within the social sciences might be most helpful: https://stats.libretexts.org/Bookshe...ied_Statistics You can find the the formulas for many of these in statistics textbooks, but you can also check out the Common Formulas page in the Back Matter of this book.

The *t*- Test

Many studies in the social sciences focus on the difference between two means. The most common statistical analysis to compare a quantitative DV for two different groups is the *t*- test. There are three different kinds of *t*-tests for three different kinds of groups.

One-Sample *t*- Test

The one-sample *t*-test is used to compare a sample mean (M) with a hypothetical population mean (μ, mu) that provides some interesting standard of comparison. The null hypothesis is that the mean of the sample (M) is statistically similar to the mean for the population (μ); in symbols, this is represented as: $M=\mu$. There could be three different research hypothesis:

1. The mean of the sample (M) is statistically different than the mean for the population (μ):, $M\neq\mu$
2. The mean of the sample (M) is statistically smaller than the mean for the population (μ): $M<\mu$
3. The mean of the sample (M) is statistically bigger than the mean for the population (μ): $M>\mu$

The first research hypothesis is non-directional, and the last two research hypotheses are directional.

The reason the *t*-statistic (or any test statistic) is useful is that we know how it is distributed when the null hypothesis is true. As shown in Figure **6.5.1**, the distribution of *t*, or the *t*-distribution, is unimodal and symmetrical, and it has a mean of 0. Its precise shape depends on a statistical concept called the degrees of freedom, which for a one-sample *t*-test is $N - 1$. (There are 24 degrees of freedom for the distribution shown in Figure **6.5.1**.) The important point is that knowing this distribution makes it possible to find the *p*-value for any *t*-score, or the probability of obtaining this result if the null hypothesis were true (that the means were similar). Consider, for example, a *t* score of 1.50 based on a sample of 25. The probability of a *t*-score at least this extreme is given by the proportion of *t*-scores in the distribution that are at least this extreme. For now, let us define *extreme* as being far from zero in either direction. Thus the *p*-value is the proportion of *t*-scores that are 1.50 or above *or* that are −1.50 or below—a value that turns out to be .14.

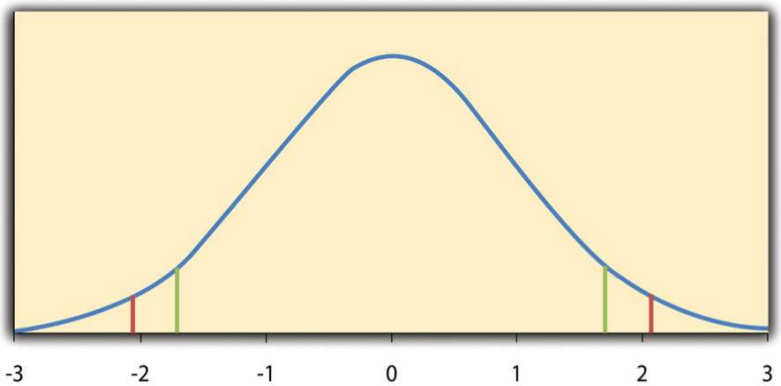


Figure **6.5.1**: Distribution of t Scores (With 24 Degrees of Freedom) When the Null Hypothesis Is True. The red vertical lines represent the two-tailed critical values, and the green vertical lines the one-tailed critical values when α = .05.

If p is equal to or less than .05 ($p<.05$), we reject the null hypothesis and conclude that the population mean differs from the hypothetical mean of interest. If p is greater than .05 ($p>.05$), we retain the null hypothesis and conclude that there is not enough evidence to say that the population mean differs from the hypothetical mean of interest. (Again, technically, we conclude only that we do not have enough evidence to conclude that it *does* differ.)

A two-tailed test is when we reject the null hypothesis if the *t*-score for the sample is extreme in either direction. This test makes sense when we believe that the sample mean might differ from the population mean but we do not have good reason to expect the difference to go in a particular direction. This follows our non-directional research hypothesis ($M \neq \mu$). But it is also possible to do a one-tailed test, where we reject the null hypothesis only if the *t*-score for the sample is extreme in one direction that we specify before collecting the data; our directional research hypotheses are examples of these. This test makes sense when we have good reason to expect the sample mean will differ from the hypothetical population mean in a particular direction.

Here is how it works. Each one-tailed critical value in Table **6.5.1** can again be interpreted as a pair of values: one positive and one negative. A *t*-score below the lower critical value is in the lowest 5% of the distribution, and a *t*-score above the upper critical value is in the highest 5% of the distribution. For 24 degrees of freedom, these values are −1.711 and 1.711. (These are represented by the green vertical lines in Figure **6.5.1**.) However, for a one-tailed test, we must decide before collecting data whether we expect the sample mean to be lower than the population mean (Research Hypothesis #2: $M<\mu$), in which case we would use only the lower critical value, or we expect the sample mean to be greater than the population mean (Research Hypothesis #3: $M>\mu$), in which case we would use only the upper critical value. Notice that we still reject the null hypothesis when the *t*-score for our sample is in the most extreme 5% of the t scores we would expect if the null hypothesis were true—so are probability remains at .05. We have simply redefined *extreme* to refer only to one tail of the distribution. The advantage of the one-tailed test is that critical values are less extreme. If the sample mean differs from the population mean in the expected direction, then we have a better chance of rejecting the null hypothesis. The disadvantage is that if the sample mean differs from the population mean in the unexpected direction, then there is no chance at all of rejecting the null hypothesis.

The Dependent-Samples *t*– Test

The dependent-samples *t*-test (sometimes called the paired-samples *t*-test) is used to compare two means for the same sample tested at two different times or under two different conditions. This comparison is appropriate for pretest-posttest designs or repeated measures experiments. The null hypothesis is that the means at the two times or under the two conditions are the same in the population. Again, there could be three different research hypothesis:

1. The mean of one sample (M_1) is statistically different than the mean for the other sample (M_2):, $M_1 \neq M_2$
2. The mean of one sample (M_1) is statistically smaller than the mean for the other sample (M_2):, $M_1 < M_2$
3. The mean of one sample (M_1) is statistically bigger than the mean for the other sample (M_2):, $M_1 > M_2$

Again, the first research hypothesis is non-directional, and the last two research hypotheses are directional, so this test can also be either two-tailed (non-directional) or one-tailed (directional); choose a directional research hypothesis if there's good reason to expect the difference goes in a particular direction.

The Independent-Samples *t*- Test

The independent-samples *t*-test is used to compare the means of two separate samples (M_A and M_B). The two samples might have been tested under different conditions in a between-subjects experiment, or they could be pre-existing groups in a cross-sectional design (e.g., women and men, extraverts and introverts). The null hypothesis is that the means of the two populations are the same: $M_A = M_B$. Again, there could be three different research hypothesis:

1. The mean of one sample (M_A) is statistically different than the mean for the other sample (M_B):, $M_A \neq M_B$
2. The mean of one sample (M_A) is statistically smaller than the mean for the other sample (M_B):, $M_A < M_B$
3. The mean of one sample (M_A) is statistically bigger than the mean for the other sample (M_B):, $M_A > M_B$

Again, the first research hypothesis is non-directional, and the last two research hypotheses are directional, so this test can also be either two-tailed (non-directional) or one-tailed (directional).

But what happens if you want to compare the means of more than two groups? Then you use an ANOVA.

ANOVA: The Analysis of Variance

Research designs comparing one quantitative DV between two different groups is very common. However, when you want to compare the means of more than two groups or conditions, the most common null hypothesis test is the analysis of variance (ANOVA). In this section, we look at the one-way ANOVA, within-subjects ANOVA, and factorial ANOVA. We will spend more time on factorial research designs later in this textbook.

One-Way ANOVA

The one-way ANOVA is used to compare the means of more than two samples (M_A, M_B…M_∞) in a between-subjects design. The null hypothesis is that all the means are equal in the population: $\mu_A = \mu_B = \ldots = M_\infty$. For research hypotheses, there is the non-directional version that not all the means in the population are equal (at least one mean differs from at least one other mean). For these more complex designs, the directional research hypotheses could be based on theory or prior research results. The easiest way to think about this is that every group's mean would be compared with every other groups mean. When there are three groups, this might look like:

1. $M_A < M_B$
2. $M_A < M_C$
3. $M_B > M_C$

This gets even more complicated if there are more than three groups since every group should have a research hypothesis that predicts a direction for every other group!

? Exercise 6.5.1

What would a directional research hypothesis look like if you compared four IV groups?

Answer

1. $M_A < M_B$
2. $M_A < M_C$
3. $M_A < M_D$
4. $M_B < M_C$
5. $M_B < M_D$
6. $M_C < M_D$

In this example, the group on the left was always hypothesized to be smaller than the group on the right, but, as long as the pairings make sense, this doesn't have to be the case. Again, follow theory or prior research to determine your predictions.

What's also interesting is that as long as there is one pair of groups that are hypothesized to be statistically different, the other pairs can be hypothesized to be similar. This might look like:

1. $M_A < M_B$
2. $M_A = M_C$
3. $M_B = M_C$

For a *t*-test, the test statistic is a *t*. The test statistic for the ANOVA is called *F* for Ronald Fisher, the statistician who is credited with developing the ANOVA (Salsburg, 2002). It is a ratio of two estimates of the population variance based on the sample data. One estimate of the population variance is for the variability between groups, which is what *t*-tests only looked at. The other population variance is within groups**,** which is based on the differences among the scores within each group.

Again, the reason that *F* is useful is that we know how it is distributed when the null hypothesis is true. As shown in Figure **6.5.2**, this distribution is unimodal and positively skewed with values that cluster around 1. The precise shape of the distribution depends on both the number of groups and the sample size, and there are degrees of freedom values associated with each of these. Again, knowing the distribution of *F* when the null hypothesis is true allows us to find the *p*-value.

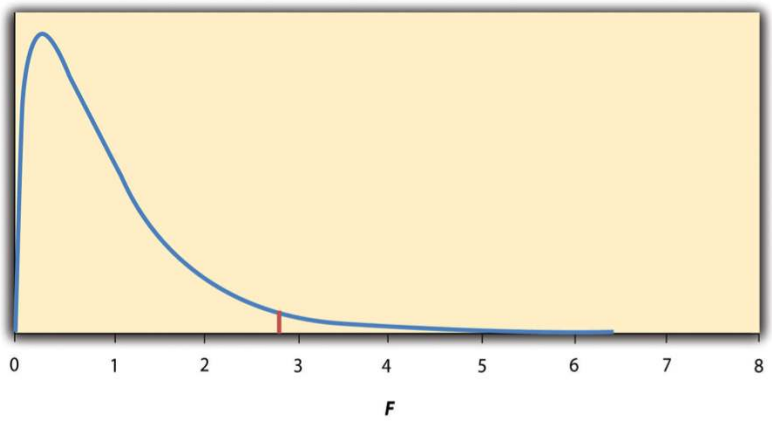


Figure **6.5.2**: Distribution of the F Ratio With 2 and 37 Degrees of Freedom When the Null Hypothesis Is True. The red vertical line represents the critical value when α is .05.

As alwys, if *p* is equal to or less than .05 ($p<.05$), then we reject the null hypothesis and conclude that there are differences among the group means in the population. If *p* is greater than .05 ($p>.05$), then we retain the null hypothesis (or fail to reject the null) and conclude that there is not enough evidence to say that there are differences.

Post Hoc Comparisons

You may have noticed that for directional research hypotheses, rejecting the null hypothesis doesn't really tell us much; when we reject the null hypothesis in a one-way ANOVA, we conclude that the group means are not *all* the same in the population. But this can indicate different things. With three groups, it can indicate that all three means are significantly different from each other. Or it can indicate that one of the means is significantly different from the other two, but the other two are not significantly different from each other. It could be, for example, that the mean calorie estimates of psychology majors, nutrition majors, and dieticians are all significantly different from each other. Or it could be that the mean for dieticians is significantly different from the means for psychology and nutrition majors, but the means for psychology and nutrition majors are not significantly different from each other. For this reason, statistically significant one-way ANOVA results are typically followed up with a series of post hoc comparisons of selected pairs of group means to determine which are different from which others.

One approach to post hoc comparisons would be to conduct a series of independent-samples *t*-tests comparing each group mean to each of the other group means. But there is a problem with this approach. In general, if we conduct a *t*-test when the null hypothesis is true, we have a 5% chance of mistakenly rejecting the null hypothesis for *each* comparison. If we conduct several *t*-tests when the null hypothesis is true, the chance of mistakenly rejecting *at least one* null hypothesis increases with each test we conduct. Researchers do not usually make post hoc comparisons using standard *t*-tests because there is too great a chance that they will mistakenly reject at least one null hypothesis. Instead, they use one of several modified *t*-test procedures—among them the Bonferonni procedure, Fisher's least significant difference (LSD) test, and Tukey's honestly significant difference (HSD) test. The details of these approaches are beyond the scope of this book, but it is important to understand their purpose: To test for differences between pairs of means while keeping the risk of mistakenly rejecting a true null hypothesis to an acceptable level (close to 5%).

Repeated-Measures ANOVA

Recall that the one-way ANOVA is appropriate for between-subjects designs in which the means being compared come from separate groups of participants. It is not appropriate for within-subjects designs in which the means being compared come from the same participants tested under different conditions or at different times. This requires a slightly different approach, called the repeated-measures ANOVA. The basics of the repeated-measures ANOVA are the same as for the one-way ANOVA. The main difference is that measuring the dependent variable multiple times for each participant allows for a more refined measure of the variation within each group. Imagine, for example, that the dependent variable in a study is a measure of reaction time. Some participants will be faster or slower than others because of stable individual differences in their nervous systems, muscles, and other factors. In a between-subjects design, these stable individual differences would simply add to the variability within the groups and increase the value of the within-groups variability (which would, in turn, decrease the value of F). In a within-subjects design, however, these stable individual differences can be measured and subtracted, lowering value of the within-groups variation; this will result in a higher value of *F* and a more sensitive test.

Factorial ANOVA

Finally, when *more than one independent variable*, not just one IV comparing different groups, we have factorial design. The appropriate approach is the **factorial ANOVA**. Again, the basics of the factorial ANOVA are the same as for the one-way and repeated-measures ANOVAs. The main difference is that it produces an *F* ratio and *p* value for *each* IV (main effect) and the *combination* of each IV (interaction). This will be detailed more in a later chapter, but returning to our calorie estimation example: Imagine that a health psychologist tests the effect of participant major (psychology vs. nutrition) and timing (morning vs. evening) in a factorial design. A factorial ANOVA would produce separate *F* ratios and *p* values for the main effect of major, the main effect of timing, and the interaction between major and timing. Appropriate modifications must be made depending on whether the design is between-subjects, within-subjects, or mixed.

References

Salsburg, D. (2002). The lady tasting tea: How statistics revolutionized science in the twentieth century. Macmillan.

This page titled 6.5: Comparing Means is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **13.3: Some Basic Null Hypothesis Tests** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

6.5.1: Effect Sizes

Learning Objectives

1. Interpret effects sizes.

Now that you've remembered a little bit about how to statistically test for differences between means, let's look at one number (an effect size) that can help you quantify these mean differences across studies.

Mean Differences Between Groups

Differences between groups or conditions are usually described in terms of the mean of each group or condition. For example, Ollendick et al. (2009) conducted a study in which they evaluated two one-session treatments for simple phobias in children. Ollendick et al. (2009) randomly assigned children with an intense fear (e.g., to dogs) to one of three conditions. In the exposure condition, the children actually confronted the object of their fear under the guidance of a trained therapist. In the education condition, they learned about phobias and some strategies for coping with them. In the wait-list control condition, they were waiting to receive a treatment after the study was over. There were several dependent variables, but we will look at the severity of each child's phobia. This DV was operationally defined as a rating on a 1-to-8 scale by a clinician who did not know which treatment the child had received. The average fear rating in the education condition was 4.83 with a standard deviation of 1.52, while the mean fear rating in the exposure condition was 3.47 with a standard deviation of 1.77. The mean fear rating in the control condition was 5.56 with a standard deviation of 1.21. Is seems like both treatments worked, but the exposure treatment worked better than the education treatment. Differences between group or condition means can be presented in a bar graph like that in Figure **6.5.1.1**, where the heights of the bars represent the group or condition means.

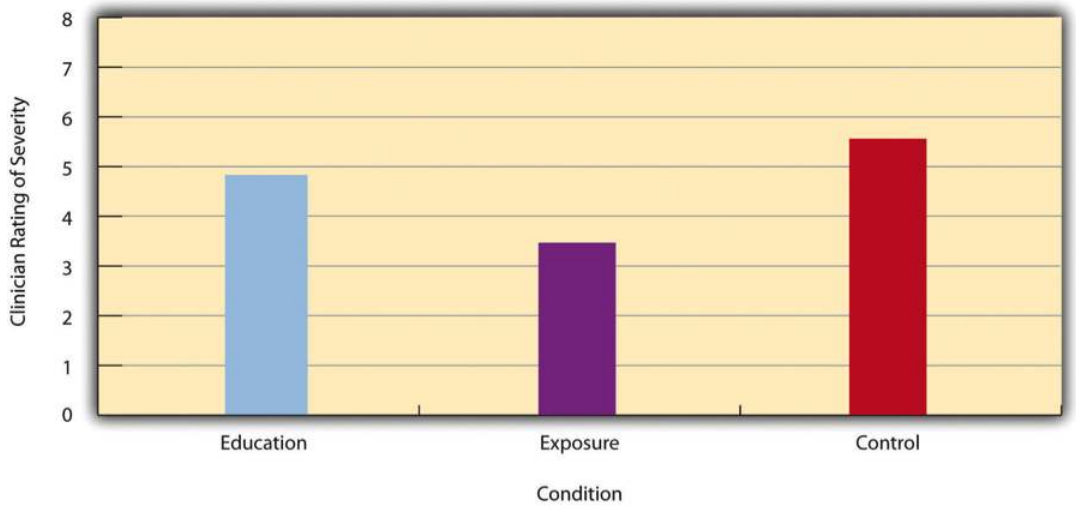


Figure **6.5.1.1**: Bar Graph Showing Mean Clinician Phobia Ratings for Children in Two Treatment Conditions

Effect Sizes

It is also important to be able to describe the strength of a statistical relationship, which is often referred to as the effect size. The most widely used measure of effect size for differences between group or condition means is called Cohen's *d*, which conceptually is the difference between the two means divided by a combined standard deviation (called the pooled-within groups standard deviation). Conceptually (but not quite mathematically), Cohen's *d* is the difference between the two means expressed in standard deviation units. A Cohen's *d* of 0.50 means that the two group means differ by 0.50 standard deviations (half a standard deviation). A Cohen's *d* of 1.20 means that they differ by 1.20 standard deviations.

But how should we interpret these values in terms of the strength of the relationship or the size of the difference between the means? Table **6.5.1.1** presents some guidelines for interpreting Cohen's *d* values in psychological research (Cohen, 1992). Values near 0.20 are considered small, values near 0.50 are considered medium, and values near 0.80 are considered large. Thus a Cohen's *d* value of 0.50 represents a medium-sized difference between two means, and a Cohen's *d* value of 1.20 represents a very large

difference in the context of psychological research. In Ollendick et al. (2009), there was a large difference (d = 0.82) between the exposure and education conditions.

Table **6.5.1.1**: Guidelines for Referring to Cohen's d and Pearson's r Values as "Strong," "Medium," or "Weak"

Relationship strength	Cohen's ^{d}	Pearson's ^{r}
Strong/large	0.80	± 0.50
Medium	0.50	± 0.30
Weak/small	0.20	± 0.10

Cohen's d is useful because it has the same meaning regardless of the variable being compared or the scale it was measured on. A Cohen's d of 0.20 means that the two group means differ by 0.20 standard deviations whether we are talking about scores on the Rosenberg Self-Esteem scale, reaction time measured in milliseconds, number of siblings, or diastolic blood pressure measured in millimeters of mercury. Not only does this make it easier for researchers to communicate with each other about their results, it also makes it possible to combine and compare results across different studies using different measures.

Be aware that the term *effect size* can be misleading because it suggests a causal relationship—that the difference between the two means is an "effect" of being in one group or condition as opposed to another. Imagine, for example, a study showing that a group of exercisers is happier on average than a group of non-exercisers, with an "effect size" of d = 0.35. If the study was an experiment—with participants randomly assigned to exercise and no-exercise conditions—then one could conclude that exercising caused a small to medium-sized increase in happiness. If the study was cross-sectional, however, then one could conclude only that the exercisers were happier than the nonexercisers by a small to medium-sized amount. In other words, simply calling the difference an "effect size" does not make the relationship a causal one.

Sex Differences Expressed as Cohen's d

Hyde (2007) looked at the results of numerous studies on gender differences and expressed the results in terms of Cohen's d . Following are a few of the values she has found, averaging across several studies in each case. (Note that because she always puts the mean for men first and subtracts the mean for women so positive values indicate that men score higher and negative values indicate that women score higher.)

Mathematical problem solving	+0.08
Reading comprehension	−0.09
Smiling	−0.40
Aggression	+0.50
Attitudes toward casual sex	+0.81
Leadership effectiveness	−0.02

Hyde (2007) points out that although men and women differ by a large amount on some variables (e.g., attitudes toward casual sex), they differ by only a small amount on the vast majority. In many cases, Cohen's d is less than 0.10, which she terms a "trivial" difference. (The difference in talkativeness discussed in Chapter 1 was also trivial: d = 0.06.) Although researchers and non-researchers alike often emphasize gender *differences*, Hyde has argued that it makes at least as much sense to think of men and women as fundamentally *similar*. She refers to this as the "gender similarities hypothesis."

We'll now move from comparing mean differences to testing for (linear) relationships with correlations.

References

Carlson, K. A., & Conard, J. M. (2011). The last name effect: How last name influences acquisition timing. *Journal of Consumer Research*, *38*(2), 300-307. doi: 10.1086/658470

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Hyde, J. S. (2007). New directions in the study of gender similarities and differences. *Current Directions in Psychological Science*, *16*, 259–263.

Ollendick, T. H., Öst, L.-G., Reuterskiöld, L., Costa, N., Cederlund, R., Sirbu, C.,…Jarrett, M. A. (2009). One-session treatments of specific phobias in youth: A randomized clinical trial in the United States and Sweden. *Journal of Consulting and Clinical Psychology*, *77*, 504–516.

This page titled 6.5.1: Effect Sizes is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **12.2: Describing Statistical Relationships** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

6.6: Introduction to Correlations

Learning Objectives

1. Interpret correlations to determine how strong the relatinship is.
2. Interpret correlations to determine how the variables vary.

The Correlation Coefficient

As a refresher, Pearson's correlation coefficient, which is traditionally denoted by r, shows the relationship between two variables, and is a measure that varies from −1.00 to 1.00. When r=−1.00 it means that we have a perfect negative relationship, and when r=1.00 it means we have a perfect positive relationship. When r=0, there's no relationship at all. If you look at Figure **6.6.1**, you can see several plots showing what different correlations look like.

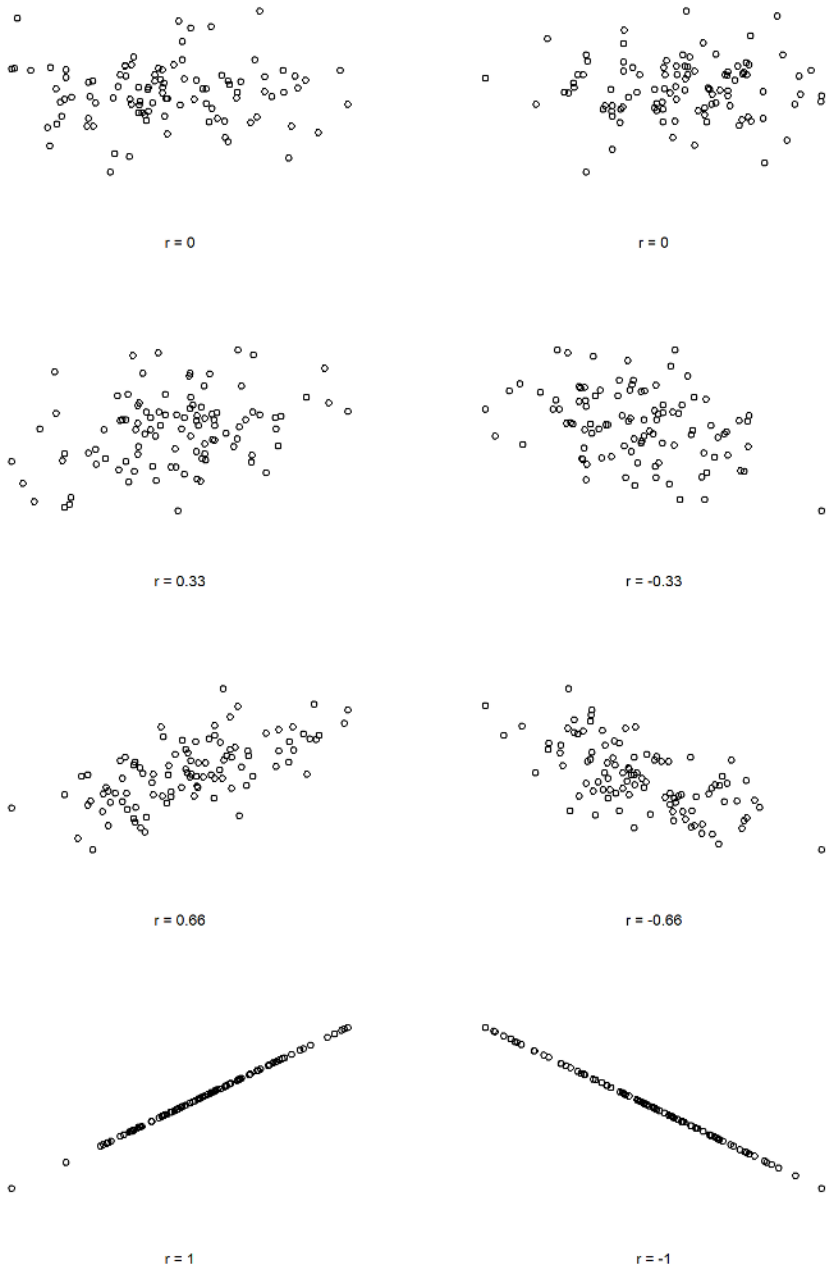


Figure **6.6.1-** Illustration of the effect of varying the strength and direction of a correlation (CC-BY-SA Danielle Navarro from Learning Statistics with R)

As you can see, strong correlations (shown on the bottom, r-values close to -1.00 or 1.00) are straight lines of dots. The closer to zero, the less linear the dots are, until you get to a correlation of zero (r=0.00) and there's just a random splatter of dots (top scatterplots). You can also see that positive correlations "point" up and to the right; as one variable increases, the other also increases. Negative correlations "point" down and to the right; as one variable increases, the other variable decreases.

Interpreting a Correlation

Naturally, in real life you don't see many correlations of 1.00. So how should you interpret a correlation of, say r = .40? The honest answer is that it really depends on what you want to use the data for, and on how strong the correlations in your field tend to be. A friend of mine in engineering once argued that any correlation less than .95 is completely useless (I think he was exaggerating, even for engineering). On the other hand there are real cases – even in psychology – where you should really expect correlations that strong. For instance, one of the benchmark data sets used to test theories of how people judge similarities is so clean that any theory that can't achieve a correlation of at least .90 really isn't deemed to be successful. However, when looking for (say)

elementary correlates of intelligence (e.g., inspection time, response time), if you get a correlation above .30 you're doing very very well. In short, the interpretation of a correlation depends a lot on the context. That said, the rough guide in Table **6.6.1** is pretty typical.

Table **6.6.1**- General Correlation Interpretations

Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive

However, something that can never be stressed enough is that you should *always* look at the scatterplot before attaching any interpretation to the data. As was said in prior sections, a correlation might not mean what you think it means. The classic illustration of this is "Anscombe's Quartet" (Anscombe, 1973), which is a collection of four data sets. Each data set has two variables, an X and a Y. For all four data sets the mean value for X is 9 ($\bar{X}_{x-axis} = 9.00$) and the mean for Y is 7.5 ($\bar{X}_{y-axis} = 7.50$). The, standard deviations for all X variables are almost identical, as are those for the the Y variables. And in each case the correlation between X and Y is r=0.816. You'd think that these four data sets would look pretty similar to one another. They do not. If we draw scatterplots of X against Y for all four variables, as shown in Figure **6.6.2** we see that all four of these are *spectacularly* different to each other.

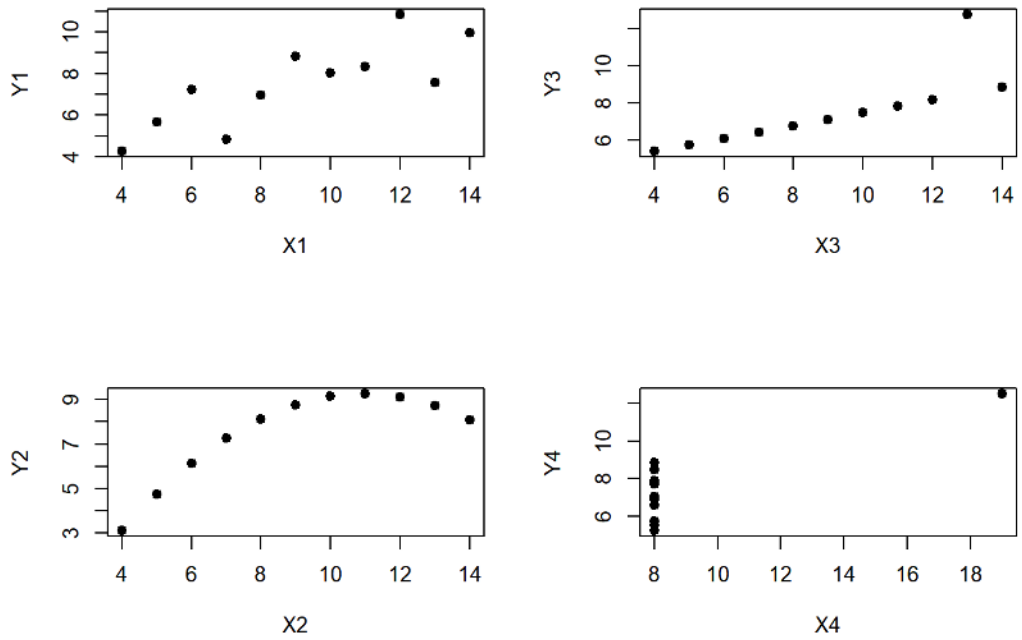


Figure **6.6.6**- Anscombe's quartet. All four of these data sets have a Pearson correlation of r=.816 (CC-BY-SA Danielle Navarro from Learning Statistics with R)

The lesson here, which so very many people seem to forget in real life is "*always graph your raw data*".

Now that you know a little more about what correlations can show, let's talk about what Pearson's correlation cannot show.

Reference

Anscombe, F. J. (1973). Graphs in statistical analysis, *The American Statistician*, *27*(1), 17-21).

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Template:ContribOja

This page titled 6.6: Introduction to Correlations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.6.1: Correlation versus Causation

We cover a great deal of material in introductory statistics, including how to interpret statistical information. Hopefully you've seen how the statistical results from quality studies can be used in your career, and day to day life, to help you make better decisions. We now come to what may be the most important lesson in introductory statistics: the difference between variables that are related (correlated) and a variable that causes a change in another variable (causation).

It is **very, very tempting** to look at variables that are correlated (have a significant Pearson's r) and assume that this means they are *causally* related; that is, it gives the impression that what we call the IV is *causing* changes in what we call the DV.

However, in reality, Pearson's correlational analysis doe not – and cannot – do this. Correlations DO NOT show causation. No matter how logical or how obvious or how convenient it may seem, no correlational analysis can demonstrate causality. The ONLY way to demonstrate a causal relation is with a properly designed and controlled experiment that rules out all of the other things that could have affected the DV.

Many times, we have prior information that suggests that one variable causes changes in the other. Thus, when we run our Pearson's r and find a strong, statistically significant results, it is very tempting to say that we found the causal relation that we are looking for. The reason we cannot do this is that, without an experimental design that includes random assignment and control variables, is that the relation we observe between the two variables may be caused by something else that we failed to measure. These "third variables" are lurking variables or confound variables, and they are impossible to detect and control for without an experiment.

> **Note: TW for drowning**
>
> A common example of this is the strong, positive correlation between ice cream sales and drowning; as ice cream sales increase, so does death by drowning. Does eating ice cream cause drowning? Probably not. Does drowning cause people to have eaten ice cream? Definitely not. Could there be a third variable? Yes! There is also a strong, positive correlation between outside temperatures and both ice cream sales and drowning. When it gets hotter outside, more people buy ice cream. And when it's hotter, more people go swimming (which just, statistically, leads to more drowning). Remember this example when you start thinking that a significant correlation means that one variable *caused* the change in the other!

Confound variables, which we will represent with C, can cause two variables (X and Y) to appear related when in fact they are not. They do this by being the hidden cause of each variable independently. That is, if C causes X and Z causes Y, the X and Y will appear to be related . However, if we control for the effect of C (the method for doing this is beyond the scope of this text), then the relation between X and Y will disappear. As another example, consider shoe size and spelling ability in elementary school children. Although there should clearly be no causal relation here (why would bigger feet lead to better spelling? Or why would better spelling lead to bigger feet?), the variables are nonetheless consistently correlated. The confound in this case? Age. Older children spell better than younger children and are also bigger, so they have larger shoes.

When there is the possibility of confounding variables being the hidden cause of our observed correlation, we will often collect data on C as well and control for it in our analysis. This is good practice and a wise thing for researchers to do. Thus, it would seem that it is easy to demonstrate causation with a correlation that controls for C. However, the number of variables that could potentially cause a correlation between X and Y is functionally limitless, so it would be impossible to control for everything. That is why we use experimental designs; by randomly assigning people to groups and manipulating variables in those groups, we can balance out individual differences in any variable that may be our cause.

It is not always possible to do an experiment, however, so there are certain situations in which we will have to be satisfied with our observed relation and do the best we can to control for known confounds. However, in these situations, even if we do an excellent job of controlling for many extraneous (a statistical and research term for "outside") variables, we must be very careful not to use causal language. That is because, even after controls, sometimes variables are related just by chance.

Sometimes, variables will end up being related simply due to random chance, and we call these correlation *spurious*. Spurious just means random, so what we are seeing is random correlations because, given enough time, enough variables, and enough data, sampling error will eventually cause some variables to be related when they should not. Sometimes, this even results in incredibly

strong, but completely nonsensical, correlations. This becomes more and more of a problem as our ability to collect massive datasets and dig through them improves, so it is very important to think critically about any significant correlation that you encounter.

The next section talks more about spurious correlations and other issues with trying to say that one variable causes changes in the other.

This page titled 6.6.1: Correlation versus Causation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **12.7: Correlation versus Causation** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

6.6.2: Factor Analysis

Learning Objectives

1. Explain some reasons that researchers use complex correlational designs.
2. Create and interpret a correlation matrix.
3. Describe how researchers can use partial correlation and multiple regression to statistically control for third variables.

As we have already seen, 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.

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)[1]. 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, Garrido, Rosenfield, & McDonald, 2009)[2]

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 (Caccioppo & Petty, 1982)[3]. The results of this study are summarized in Table **6.6.2.1**, 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 **6.6.2.1**: 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	—

Factor Analysis

When researchers study relationships among a large number of conceptually similar variables, they often use a complex statistical technique called **factor analysis**. In essence, factor analysis organizes the variables into a smaller number of clusters, such that they are strongly correlated within each cluster but weakly correlated between clusters. Each cluster is then interpreted as multiple measures of the same underlying construct. These underlying constructs are also called "factors." For example, when people perform a wide variety of mental tasks, factor analysis typically organizes them into two main factors—one that researchers interpret as mathematical intelligence (arithmetic, quantitative estimation, spatial reasoning, and so on) and another that they interpret as verbal intelligence (grammar, reading comprehension, vocabulary, and so on). The Big Five personality factors have been identified through factor analyses of people's scores on a large number of more specific traits. For example, measures of warmth, gregariousness, activity level, and positive emotions tend to be highly correlated with each other and are interpreted as representing the construct of extraversion. As a final example, researchers Peter Rentfrow and Samuel Gosling asked more than 1,700 university students to rate how much they liked 14 different popular genres of music (Rentfrow & Gosling, 2008)[4]. They then submitted these 14 variables to a factor analysis, which identified four distinct factors. The researchers called them *Reflective and Complex* (blues, jazz, classical, and folk), *Intense and Rebellious* (rock, alternative, and heavy metal), *Upbeat and Conventional* (country, soundtrack, religious, pop), and *Energetic and Rhythmic* (rap/hip-hop, soul/funk, and electronica); see Table 6.6.2.2.

Table **6.6.2.2**: 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.

Two additional points about factor analysis are worth making here. One is that factors are not categories. Factor analysis does not tell us that people are *either* extraverted *or* conscientious or that they like *either* "reflective and complex" music *or* "intense and rebellious" music. Instead, factors are constructs that operate independently of each other. So people who are high in extraversion might be high or low in conscientiousness, and people who like reflective and complex music might or might not also like intense and rebellious music. The second point is that factor analysis reveals only the underlying structure of the variables. It is up to researchers to interpret and label the factors and to explain the origin of that particular factor structure. For example, one reason that extraversion and the other Big Five operate as separate factors is that they appear to be controlled by different genes (Plomin, DeFries, McClean, & McGuffin, 2008)[5].

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." It is true that correlational research cannot unambiguously establish that one variable causes another. Complex correlational research, however, can often be used to rule out other plausible interpretations. The primary way of doing this is through the **statistical control** of potential third variables. Instead of controlling these variables through random assignment or by holding them constant as in an experiment, the researcher instead measures them and includes them in the statistical analysis called **partial correlation**. Using this technique, researchers can examine the relationship between two variables, while statistically controlling for one or more potential third variables.

For example, assume a researcher was interested in the relationship between watching violent television shows and aggressive behavior but she was concerned that socioeconomic status (SES) might represent a third variable that is driving this relationship. In this case, she could conduct a study in which she measures the amount of violent television that participants watch in their everyday life, the number of acts of aggression that they have engaged in, and their SES. She could first examine the correlation between violent television viewing and aggression. Let's say she found a correlation of +.35, which would be considered a

moderate sized positive correlation. Next, she could use partial correlation to reexamine this relationship after statistically controlling for SES. This technique would allow her to examine the relationship between the part of violent television viewing that is independent of SES and the part of aggressive behavior that is independent of SES. If she found that the partial correlation between violent television viewing and aggression while controlling for SES was +.34, that would suggest that the relationship between violent television viewing and aggression is largely independent of SES (i.e., SES is not a third variable driving this relationship). On the other hand, if she found that after statistically controlling for SES the correlation between violent television viewing and aggression dropped to +.03, then that would suggest that SES is indeed a third variable that is driving the relationship. If, however, she found that statistically controlling for SES reduced the magnitude of the correlation from +.35 to +.20, then this would suggest that SES accounts for some, but not all, of the relationship between television violence and aggression. It is important to note that while partial correlation provides an important tool for researchers to statistically control for third variables, researchers using this technique are still limited in their ability to arrive at causal conclusions because this technique does not take care of the directionality problem and there may be other third variables driving the relationship that the researcher did not consider and statistically control.

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:

$$Y = b_1X_1 \tag{6.6.2.1}$$

Y in this formula represents the person's predicted score on the outcome variable, b_1 represents the slope of the line depicting the relationship between two variables (or the regression weight), and X1 represents the person's score on the predictor variable. You can see that to predict a person's score on the outcome variable (Y), one simply needs to multiply their score on the predictor variable (X) by the regression weight (b_1)

While **simple regression** involves using one variable to predict another, **multiple regression** involves measuring several variables (X_1, X_2, X_3,…X_i), and using them to predict some outcome variable (Y). Multiple regression can also be used to simply describe the relationship between a single outcome variable (Y) and a set of predictor variables (X_1, X_2, X_3,…X_i). The result of a multiple regression analysis is an equation that expresses the outcome variable as an additive combination of the predictor variables. This regression equation has the following general form:

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + ... + b_iX_i \tag{6.6.2.2}$$

The regression weights (b_1, b_2, and so on) indicate how large a contribution a predictor variable makes, on average, to the prediction of the outcome variable. Specifically, they indicate how much the outcome variable changes for each one-unit change in the predictor variable.

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. Research like this, by

the way, has shown both income and health make extremely small contributions to happiness except in the case of severe poverty or illness (Diener, 2000[6]).

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.

References

1. Radcliffe, N. M., & Klein, W. M. P. (2002). Dispositional, unrealistic, and comparative optimism: Differential relations with knowledge and processing of risk information and beliefs about personal risk. *Personality and Social Psychology Bulletin*, *28*, 836–846. ↵
2. 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. ↵
3. Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116–131. ↵
4. Rentfrow, P. J., & Gosling, S. D. (2008). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, *84*, 1236–1256. ↵
5. Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral genetics* (5th ed.). New York, NY: Worth. ↵
6. Diener, E. (2000). Subjective well-being: The science of happiness, and a proposal for a national index. *American Psychologist*, *55*, 34–43. ↵

This page titled 6.6.2: Factor Analysis is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **6.4: Complex Correlation** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

6.7: The Qualitative Approach

Learning Objectives

1. Differentiate analysis of qualitative research from quantitative research.
2. Identify approaches to analyzing qualitative research.

While this section is titled “The Qualitative Approach,” it is actually inaccurate to suggest that there is just one overall approach to qualitative research. There are some core characteristics that qualitative approaches to research have in common, such as data that relies on words or images rather than numbers and a richer, more contextual understanding of the phenomena under study, but there are also many ways in which qualitative approaches to research vary. They use different methods of data collection. They take place within different paradigms and epistemologies. They focus their attention on emphasizing different standards for research quality. And, as the following sections will show, they utilize different methods for preparing and managing data, analyzing that data, and disseminating their findings.

Introduction to the Qualitative Approach

Your statistics class probably mentioned qualitative variables (sometimes called categorical variables or nominal variables) in contrast to quantitative variables. In fact, many statistics instructors go into detail about the different scales of measurement for quantitative variables.

Exercise 6.7.1

Can you describe the difference between qualitative variables and quantitative variables in terms of measuring behavior?

Answer

Qualitative variables are defined as a type of variable that has different values to represent different *categories* or kinds. This is the same as the nominal scale of measurement, while quantiative variables are defined as a type of variable that is measured with some sort of scale that uses numbers that measure something. This means, qualitative variables are about categorizing behavior into different kinds or categories, while quantitative variables are about measuring behavior through some sort of scale.

While there is not a one-to-one comparison, qualitative research mostly measures qualitative variables, and quantitative research mostly uses quantitative variables. However, both kinds of research can use either kind of variable. This section will orient you more to qualitative research, and then describe the extensive and time-consuming process for analyzing data gathered in qualitative research.

At the most basic level, qualitative research is research that emphasizes data that is not numerical in nature, data like words, pictures, and ideas. In contrast, quantitative research emphasizes measuring behavior with numerical scales, or at least variables that can relatively easily be translated into numerical terms. In other words, quantitative data is about quantities, while qualitative data is about qualities. Beyond this basic distinction, qualitative research can look very similar to quantitative research or it can take a very different approach; when people talk about qualitative approaches to research, however, they are often focused on those approaches that are distinct from what quantitative researchers do.

So what are some of the unique features of qualitative research? First of all, qualitative research tends to rely on the use of rich, thick description. In other words, qualitative research does not just provide summaries of data and findings, it really takes the reader or consumer of research there and lets them explore the situation and make conclusions for themselves by drawing on extended descriptions and excerpts from the data. Qualitative research also leaves room for focus on feelings and emotions, elements of the social world that can be harder to get at with quantitative data. For the qualitative researcher, data should not just depict specific actions or occurrences, but rather the contexts and backgrounds that lead up to what happened. More broadly, qualitative research tends to focus on a deep understanding of a specific place or organization or of a particular issue, rather than providing a wider but shallower understanding of an area of study.

Among the strengths of qualitative research are that it provides for the development of new theories and the exploration of issues that people do not know much about. It is very high in validity since it is so connected to real life. And it permits the collection and analysis of more detailed, contextual, and complex kinds of data and information. Of course, with strengths come limitations. The higher validity of qualitative data is matched with lower reliability due to the unique circumstances of data collection and the impact of interviewer effect. The greater ability to develop new theories is matched with a greater difficulty testing existing theories, especially causal ones, given the impossibility of eliminating alternative explanations for the phenomena under investigation. The ability to collect more detailed and complex information comes in large part due to the focus on a much smaller number of participants or cases, which in turn limits generalizability and in some cases can limit representativeness. And while there is no reason to conclude that any of these factors make qualitative research more prone to bias than quantitative research, which after all can be profoundly impacted by slight variations in survey question wording or sample design, those who are not well informed about research methodology may discount the strengths of qualitative research by suggesting that the lack of numbers or the close interaction between participants and researchers bias the results.

In their classic text on qualitative data analysis, Miles and Huberman (1994) present the following as among the key elements of qualitative data analysis:

- It involves more prolonged contact with more ordinary aspects of human life;
- It has a holistic rather than a particularistic focus, aiming to keep data and findings in context;
- Multiple interpretations and understandings of data are possible, and researchers should preserve respondents' own understandings of their worlds and lives;
- There is a lack of standardization and measurement, with the researcher themselves becoming the primary measurement instrument; and
- Analysis is done primarily with words.

However, other scholars would argue that qualitative analysis is not limited to words—it may also involve visuals ways of engaging with and presenting data.

Types of Qualitative Data

The data that we analyze in qualitative research consist primarily of words and images drawn from observation, interaction, interviewing, or existing documents. In particular, the types of data collection that tend to result in qualitative data include talking with people (interviews and focus groups) and the unobtrusive research methods discuss (observations, physical trace, archival data) which were discussed in the chapter on measuring behavior. These different data collection strategies imply a variety of analytical strategies as well, and indeed qualitative data analysis relies on a breadth of techniques. Thus, part of the process of formulating and selecting qualitative data is selecting the right kinds of strategies to apply to the particular data being utilized.

Paradigms of Research

Researchers approach their research from different perspectives or paradigms. A paradigm is a set of assumptions, values, and practices that shapes the way that people see, understand, and engage with the world, and thus the particular paradigm that a researcher inhabits shapes the fashion in which they carry out their research. Philosophers use the term epistemology to refer to the study of the nature of knowledge, and thus we can take an epistemological perspective to understanding how paradigms of research might vary.

Two paradigms that commentators often juxtapose are the positivist and interpretivist approaches. Positivism assumes that there is a real, verifiable reality and that the purpose of research is to come as close to it as possible. Thus, a positivist would argue that we can understand the world, subject it to prediction and control, and—through the processes of research and data analysis—empirically verify our claims. Positivist research projects can utilize a variety of methods, but experimental and quantitative survey data are especially likely. This might feel familiar as this is the paradigm of experimental research that most of this textbook covers. For qualitative research, positivism is often associated with the type of observational study once common in anthropology that aimed at uncovering the "real" social practices of a group. These methods tend to involve keeping some degree of distance between the researcher and the participants and positioning the researcher as the expert on both research methods and the participants' own lives. From a positivist perspective, standards of rigor like reliability and validity (including external validity), which we've discussed in some detail, are important and attainable markers of good research as they contribute to the likelihood that the research

arrives at the right answer. As this suggests, objectivity is an essential goal of positive research. Good research, to a positivist, is that which is valid, reliable, generalizable, and has strong, significant results.

In contrast, interpretivism suggests that our knowledge of the world is created by our own individual experiences and interactions, and thus that reality cannot be understood as existing on its own in a form separate from our distinct existences. Thus, an interpretivist would argue that understandings are always based in a particular time and on a particular interpreter and are always open to reinterpretation. Interpretivist research projects utilize naturalistic research methods that are rooted in real social contexts, especially in-depth interviewing and participant-observation. These methods tend to involve a closer and more reciprocal relationship between the researcher and the participants, which aligns well with qualitative reseach. There may also be greater concern for ethical treatment and, in some cases, an emphasis on possibilities for social change. Interpretivist researchers also value participants' expertise and their understandings of their own lives rather than assuming the researcher's perspective is necessarily more accurate. From an interpretivist perspective, validity may not be attainable due to the fact that truth is not certain, and in any case standards of rigor are far less important than considerations like ethics, morality, the degree to which biases are made clear, and what the world can learn from the research. As this might suggest, interpretivists would tend to believe that objectivity is probably not attainable, and that even it is, the pursuit of it may not be worthwhile. To an interpretivist, good research is that which is done in a careful, respectful manner, contributes to knowledge, is reflective, and takes appropriate political and ethical considerations into account.

Lisa Pearce (2012) has outlined a paradigm she calls pragmatist. This approach is sometimes understood as a kind of middle position between positivist and interpretivist ways of thinking. Thus, its proponents neither believe that strict objectivity is possible nor abandon efforts to seek objectivity at all, instead researchers reflect on their own perspectives and bias, as well as the experience and perspectives of the particiants. This process of reflecting on how the researchers' conscious and unconscious beliefs is called reflexivity. While pragmatist approaches can be used with various methods of data collection, they tend to be employed by those using mixed-methods approaches, especially those combining quantitative and quantitative strategies.

While the discussion of paradigms here is not exhaustive—there are many other approaches to research, many other epistemologies—it does provide an overview of some of the possible ways to think about research and data analysis. One important thing to remember is that while there *are* criteria for good research, there are no objective or empirical standards for which paradigm is "correct." In other words, researchers approach their research from the perspective or philosophy that makes sense to them, and while others may have reasons for disapproving, they cannot say that such a choice is right or wrong. Researchers must make these sorts of decisions for themselves.

Inductive and Deductive Approaches

Another question we might ask about the epistemology of research processes is whether our data *emerges from* our analysis or whether our data *generates* our analysis. If you argue that data emerges from analysis, you are suggesting that you begin the research process with a theory and then look to the data you have collected to see whether or not you can find support for your theory. This approach enables the testing of theories. It is typically understood as a deductive approach to research. In deductive approaches, researchers develop a theory, collect data, analyze the data, and use their analysis to test their theory. Positivist research is often deductive in its approach. Your experience with null hypothesis significance testing was deductive.

Instead, if you argue that data generates analysis, you are suggesting that you begin the research process by collecting data and you then look to see what you can find within it. This approach enables the building of theories. It is typically understood as an inductive approach. In inductive approaches, researchers begin by collecting data. Then they analyze that data and use that analysis to build new understandings. Interpretivist and feminist research are often inductive in their approach.

While qualitative research can be conducted using both deductive and inductive approaches, it is a bit more common for qualitative researchers to use inductive approaches. Such approaches are far less possible in quantitative analysis because of the need for more precisely-designed data collection techniques. Thus, one advantage of qualitative research is that it permits for an inductive approach and is thus especially useful in contexts in which very little is already know or where new explanations need to be uncovered. It is also possible to conduct research using what some call abduction, or an interplay between deductive and inductive approaches (Pearce 2012). Such an approach may also be found in mixed-methods research. The next sections will focus primarily on inductive approaches to qualitative data analysis, given that they are far more common. But deductive approaches do exist. For example, consider a researcher who is interested in what sorts of circumstances give rise to nonprofit organization boards deciding

to replace the organization's director. More typically, a qualitative researcher with this question would interview a wide variety of non-profit board members and, based on the responses, would build a theory—an inductive approach. In contrast, the researcher could choose to conduct her study deductively. Then, she would read the prior literature on management and organizational decision-making and develop one or more hypotheses about the circumstances that give rise to leadership changes. She would then interview board members looking specifically for the constellation of circumstances she hypothesized to test whether these circumstances were associated with the decision to replace the director.

The Process of Qualitative Research

To analyze qualitative data, it's important to get a basic idea about how to conduct an inductive qualitative research project. While there are a series of steps researchers follow, it is important to note that, unlike quantitatie research, qualitative research and data analysis involve a high degree of fluidity and are typically iterative, meaning that they involve repeatedly returning to prior steps in the process.

First, researchers design their data collection process, which includes developing any data collection instruments such as interview guides and locating participants. Then, they collect their data. To collect data researchers might conduct interviews, observations, or they might locate documents or other sources of textual or visual data. While deductive quantitative approaches require researchers collect all their data and only then analyze it, inductive qualitative approaches provide the opportunity for more of a cyclical process in which researchers collect data, begin to analyze it, and then use what they have found so far to reshape their further data collection. Once data is collected, researchers need to ensure that their data is usable. This may require the transcription of audio or video recordings, the scanning or photocopying of documents, typing up handwritten field notes, or other processes designed to move raw data into a more manipulable form.

Next, researchers engage in data reduction. Research projects typically entail the collection of really large quantities of data, more data that can possibly be managed or utilized in the context of one paper. This is especially likely in the case of qualitative research because of the richness and complexity of the data that is collected. Therefore, once data collection is completed, researchers use strategies and techniques to reduce the hundreds or thousands of pages of transcripts, observation notes, or documents into a manageable form. Activities involved in data reduction include coding, summarization, the development of data displays, and categorization.

Once data reduction has made data more usable, researchers develop conclusions based on their data. Remember, however, that this process is iterative, which means that it is a continuing cycle. So, when researchers make conclusions, they also go back to earlier stages to refine their approaches. For example, imagine a researcher interviewed twelve African American teenage girls about their experiences in school. At the beginning, the researcher may develop codes related to academic achievement and romantic relationships. However, as the researcher keeps reading the interview transcripts she may realize that relationships with friends and family are also mentioned and impactful. The researcher would then have to review all of the interviews that they've already analyzed to look for information on friendships and family relationships.

The process of developing conclusions also requires careful consideration to limitations of the data and analytical approaches, but finally, researchers will present their findings. During each project, researchers must determine how best to disseminate results. Factors influencing this determination include the research topic, the audience, and the intended use of the results—for instance, are these the results of basic research, designed to increase knowledge about the phenomena under study, or are they the results of applied research, conducted for a specific audience to inform the administration of a policy or program? Findings might be disseminated in a graphical form like an infographic or a series of charts, a visual form like a video or animation, an oral form like a lecture, or a written form like a scholarly article or a report. Of course, many projects incorporate multiple forms of dissemination.

In sum, the qualitative research cycle looks like:

1. Design
2. Collect data
3. Data reduction and analysis (some researchers may consider these two separate steps)
4. Conclusions and presenting (again, some researchers may consider these separate steps)

If you remember all the way back to Figure 2.3.1, this qualitative approach more or less covers all of the steps of the scientific method other than deciding what topic to research (Research Question).

References

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications, Inc.

Pearce, L. D. (2012). Mixed methods inquirty in sociology. *American Behavioral Scientitist*, *56*(6), 829-848. https://doi.org/10.1177/0002764211433798

This page titled 6.7: The Qualitative Approach is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Mikaila Mariel Lemonik Arthur via source content that was edited to the style and standards of the LibreTexts platform.

- **4.1: The Qualitative Approach** by Mikaila Mariel Lemonik Arthur is licensed CC BY-NC-SA 4.0. Original source: https://pressbooks.ric.edu/socialdataanalysis/.

6.7.1: Qualitative Analysis- Preparing and Managing Qualitative Data

When you have completed data collection for a qualitative research project, you will likely have voluminous quantities of data—thousands of pages of fieldnotes, hundreds of hours of interview recordings, many gigabytes of images or documents—and these quantities of data can seem overwhelming at first. Therefore, preparing and managing your data is an essential part of the qualitative research process. Researchers must find ways to organize the voluminous quantities of data into a form that is useful and workable. This chapter will explore data management and data preparation as steps in the research process, steps that help facilitate data analysis. It will also review methods for data reduction, a step designed to help researchers get a handle on the volumes of data they have collected and coalesce the data into a more manageable form. Finally, it will discuss the use of computer software in qualitative data analysis.

Data Management

Even before the first piece of data is collected, a data management system is a necessity for researchers. Data management helps to ensure that data remain safe, organized, and accessible throughout the research process and that data will be ready for analysis when that part of the project begins. Miles and Huberman (1994) outline a series of processes and procedures that are important parts of data management.

First, researchers must attend to the formatting and layout of their data. Developing a consistent template for storing fieldnotes, interview transcripts, documents, and other materials, and including consistent metadata (data about your data) such as time, date, pseudonym of interviewee, source of document, person who interacted with the data, and other details will be of much use later in the research process.

Similarly, it is essential to keep detailed records of the research process and all research decisions that are made. Storing these inside one's head is insufficient. Researchers should keep a digital file or a paper notebook in which all details and decisions are recorded. For instance, how was the sample conducted? Which potential respondents never ended up going through with the interview? What software decisions were made? When did the digital voice recorder fail, and for how long? What day did the researcher miss going into the field because they were ill? And, going forward, what decisions were made about each step in the analytical process?

As data begin to be collected, it is necessary to have appropriate, well-developed physical and/or digital filing systems to ensure that data are safely stored, well-organized, and easy to retrieve when needed. For paper storage, it is typical to use a set of file folders organized chronologically, by respondent, or by some other meaningful system. For digital storage, researchers might use a similar set of folders or might keep all data in a single folder but use careful file naming conventions (e.g. RespondentPseudonym_Date_Transcript) to make it easy to find each piece of data. Some researchers will keep duplicate copies of all data and use these copies to begin to sort, mark, and organize data in ways that enable the presence of relationships and themes to emerge. For instance, researchers might sort interview transcripts by the way respondents answered a particular key question. Or they might sort fieldnotes by the central activities that took place in the field that day. Activities such as these can be facilitated by the use of index cards, color-coding systems, sticky notes, marginal annotations, or even just piles. Cross-referencing systems may be useful to ensure that thematic files can be connected to respondent-based files or to other relevant thematic files. Finally, it is essential that researchers develop a system of backups to ensure that data is not lost in the event of a catastrophic hard drive failure, a house fire, lack of access to the office for an extended period, or some other type of disaster.

One more issue to attend to in data management is research ethics. It is essential to ensure that confidential data is protected from disclosure; that identifying information (including signed consent forms) are not kept with or linkable to data; and that all researchers, analysts, interns, and administrative personnel involved in a study sign statements of confidentiality to ensure they understand the importance of nondisclosure (Berg 2009). Note that such documents will not protect researchers and research personnel from subpoena by the courts—if research documents will contain information that could expose participants to criminal or legal liability, there are additional concerns to consider and researchers should do due diligence to protect themselves and their respondents (see, e.g., Khan 2019), though the methods and mechanisms for doing so are beyond the scope of this text. Researchers must attend to data security protocols, many of which were likely agreed to in the IRB submission process. For example, paper research records should be locked securely where they cannot be seen by visitors or by personnel or accessed by accident. Digital records should be securely stored in password protected files that meet current standards for strong passwords. Cloud storage or backups should have similar protections, and researchers should carefully review the terms of service to ensure that they continue to own their data and that the data are protected from disclosure.

Preparing Data

In most cases, data are not entirely ready for analysis at the moment at which they are collected. Additional steps must be taken to prepare data for analysis, and these steps are somewhat different depending on the form in which the data exists and the approach to data collection that was used: fieldnotes from observation or ethnography, interviews and other recorded data, or documentary data like texts and images.

Fieldnotes

When researchers conduct ethnographic or observational research, they typically do not have the ability to maintain verbatim recordings. Instead, they maintain fieldnotes. Maintaining fieldnotes is a tricky and time-consuming process! In most instances, researchers cannot take notes—at least not too many—while present in the research site without making themselves conspicuous. Therefore, they need to limit themselves to a couple of jotted words or sentences to help jog their memories later on, though the quantity of notes that can be taken in the field is higher these days because of the possibility of taking notes via smartphone, a notetaking process largely indistinguishable from the socially-ubiquitous practices of text messaging and social media posts. Immediately after leaving the site, researchers use the skeleton of notes they have taken to write up full notes recording everything that happened. And later, within a day or so, many researchers go back over the fieldnotes to edit and refine the fieldnotes into a useful document for later analysis. As this process suggests, analysis is already beginning even while the research is ongoing, as researchers make notes and annotations about theoretical ideas, connections to explore, potential answers to their research questions, and other things in the process of refining their fieldnotes.

When fleshing out fieldnotes, researchers should be attentive to the distinctions between recollections they believe are accurate, interpretations and reflections they have made, and analytical thoughts that develop later through the process of refining the fieldnotes. It is surprisingly easy for a slight mistake in recording,

say, which people did what, or in what sequence a series of events occurred, to entirely change the interpretation of circumstances observed in the field. To demonstrate how such issues can arise, consider the following two hypothetical fieldnote excerpts:

Excerpt A	Excerpt B
Sarah walked into the living room and before she knew what happened, she found Marisol on the floor in tears, surrounded by broken bits of glass. "What did you do?" Sarah said, her voice thick with emotion. Marisol covered her face and cried louder.	Her voice thick with emotion, Sarah said, "What did you do?" Before she knew what happened, she found Marisol on the floor in tears, surrounded by bits of broken glass. Sarah walked into the living room. Marisol covered her face and cried louder.

In Excerpt A, the most reasonable interpretation of events is probably that Sarah walked into the room and found Marisol, the victim of an accident, and was concerned about her. In Excerpt B, in contrast, Sarah probably caused the accident herself. Yet the words are exactly the same in both excerpts—they have just been slightly rearranged. This example highlights how important careful attention to detail is in recording, refining, and analyzing fieldnotes (and other forms of qualitative data, for that matter).

Fieldnotes contain within them a vast array of different types of data: records of verbal interactions between people, observations about social practices and interactions, researchers' inferences and interpretations of social meanings and understandings, and other thoughts (Berg 2009). Therefore, as researchers work to prepare their fieldnotes for analysis, they may need to work through them again to organize and categorize different types of notes for different uses during analysis. The data collected from ethnographic or observational research can also include documents, maps, images, and recordings, which then need to be prepared and managed alongside the fieldnotes.

Interviews & Other Recordings

First of all, interview researchers need to think carefully about the form in which they will obtain their data. While most researchers audio- or video-record their interviews, it is useful to keep additional information alongside the recordings. Typically, this might include a form for keeping track of themes and data from each interview, including details of the context in which the interview took place, such as the location and who was present; biographical information about the participant; notes about theoretical ideas, questions, or themes that occur to the researcher during the interview; and reminders of particularly notable or valuable points during the interview. These information sheets should also contain the same pseudonym or respondent number that is used during the interview recording, and thus can be helpful in matching biographical details to participant quotes at the time of ultimate writeup. Interviewers may also want to consider taking notes throughout the interview, as notes can highlight elements of body language, facial expression, or more subtle comments that might not be picked up on audio recordings. While video recordings can pick up such details, they tend to make participants more self-conscious than do audio recordings.

Once the interview has concluded, recordings need to be transcribed. While automated transcription has improved in recent years, it still falls far short of what is needed to make an accurate transcript. Transcription quality is typically assessed using a metric called the Word Error Rate—basically, dividing the number of incorrect words by the number of words that should appear in the passage—there are other, more complex assessment metrics that take into consideration individual words' importance to meaning. As of 2020, automated transcription services still tended to have Word Error Rates of over 10%, which may be sufficient for general understanding (such as in the case of apps that convert voicemails to text) but which is definitely too high of an error rate for use in data analysis. And error rates increase when audio recordings contain background noise, accented speech, or the use of dialects other than Standard American English (SAE). There can also be ethical concerns about data privacy when automated services are used (Khamsi 2019). However, automated services can be cost-effective, with a typical cost of about 25 cents per minute of audio (Brewster 2020). For a typical study involving 40 interviews averaging 90 minutes each, this would come to a total cost of about $900, far less than the cost of human transcription, which averages about $1 per minute these days. Human transcription is far more accurate, with extremely low Word Error Rates, especially for words essential to meaning. But human transcribers also suffer from increased error when transcribing audio with noisy backgrounds, where multiple speakers may be interrupting one another (for instance in recordings of focus groups), or in cases where speakers have stronger accents or speak in dialects other than Standard American English. For example, a study examining court reporters—professional transcribers with special experience and training at transcribing speech in legal contexts—working in Philadelphia who were assigned to transcribe African American English had average Word Error Rates of above 15%, and these errors were significant enough to fundamentally alter meaning in over 30% of the speech segments they transcribed (Jones *et al.* 2019).

Researchers can, of course, transcribe their recordings themselves, an option that vastly reduces cost but adds an enormous amount of time to the data preparation process. The use of specialized software or devices like foot-pedal controlled playback can facilitate the ease of transcription, but it can easily take up to four hours to complete the transcription of one hour of recordings. This is because people speak far faster than they type—a typical person speaks at a rate of about 150 words per minute and types at a rate more like 30-60 words per minute. Another possibility is to use a kind of hybrid approach in which the researcher uses automated transcription or voice recognition to get a basic—if error-laden—transcript and then corrects it by hand. Given the time that will be invested in correcting the transcript by listening to the recording while reviewing the transcript, even lower-quality transcription services may be acceptable, such as the automated captioning video services like YouTube offer, though of course these services also present data privacy concerns. Alternatively, researchers might use voice-recognition software. The accuracy of such software can typically be improved by training it on the user's voice. This approach can be especially helpful when interview respondents speak with accents, as the researcher can re-record the interview in their own voice and feed it into software that is already trained to understand the researcher's voice.

Table 1 below compares different approaches to transcription in terms of financial cost, time, error rate, and ethical concerns. Costs for transcription by the researcher and hybrid approaches are typically limited to the acquisition of software and hardware to aid the transcription process. For a new researcher, this might entail several hundred dollars of cost for a foot pedal, a good headset with microphone, and software, though these costs are often one-time costs not repeated with each project. In contrast, even automated transcription can cost nearly a thousand dollars per project, with costs far higher for the hired human transcriptionsts who have much better accuracy. In terms of time, though, automated and hired services require far less of the researchers' time. Hired services

will require some time for turnaround, more if the volume of data is high, but the researcher can work on other things during that time. For self and hybrid transcription approaches, researchers can expect to put in much more time on transcription than they did conducting interviews. For a typical project involving 40 interviews averaging 90 minutes each, the time required to conduct the interviews and transcribe them—not including time spent preparing for interviews, recruiting participants, traveling, analyzing data, or any other task—can easily exceed 300 hours. If you assume a researcher has 10 hours per week to devote to their project, that would mean it would take over 30 weeks just to collect and transcribe the data before analysis could begin. And after transcription is complete, most researchers find it useful to listen to the recordings again, transcript in hand, to correct any lingering errors and make notes about avenues for exploration during data analysis.

Table 1. Comparing Transcription Approaches for a Typical Interview-Based Research Project

	Cost	Time	Error Rate	Ethical Concerns
Automated	$900	A few hours turnaround	High	High
Hired	$3,600	At least several days turnaround	Low for SAE	Probably Low
Self	Minimal	About 240 hours active	Varies	Low
Hybrid	Minimal	Varies, likely at least 120 hours active	Low for SAE	Varies

Note: this table assumes a project involving 40 interviews, all conducted by the main researcher, averaging 90 minutes in length. Time costs do not include interviewing itself, which would add an additional 60 hours to the time required to complete the project.

Documents and Images

Data preparation is far different when data consists of documents and images, as these already exist in textual form. Here, concerns are more likely to revolve around storage, filing, and organization, which will be discussed later in this chapter. However, it can be important to conduct a preliminary review of the data to better understand what is there. And for visual data, it may be especially useful to take notes on the content in and the researcher's impressions of each visual as a starting point to thinking about how to further work with the materials (Saldaña 2016).

There are special concerns about research involving documents and images that are worth noting here. First all, it is important to remember the importance of sampling issues in relation to the use of documents. Sampling is not always a concern—for instance, research involving newspaper articles may involve a well-conducted random sample, or photographs may have been taken by the researcher themselves according to a clear purposive sampling process—but many projects involving textual data have used sampling procedures where it remains unclear how representative the sample is of the universe of data. Researchers must keep careful notes on where the documents and images included in their data came from and what sorts of limitations may exist in the data and include a discussion of these issues in any reporting on their research.

When writing about interview data, it is typical to include excerpts from the interview transcripts. Similarly, when using documents or visual materials, it is preferable to include some of the original data. However, this can be more complex due to copyright concerns. When using published works, there are real legal limits on the quantity of text that you can include without getting permission from the copyright owner, who may make you pay for the privilege. This is not an issue for works that were created or published more than 95 years ago, as their copyrights have expired. For works more recent than that, the use of more than a small portion of the work typically violates copyright, and the use of an image is almost never permitted unless it has been specifically released from copyright (or created by the researcher themselves). Archival data may be subject to specific usage restrictions imposed by the archive or donor. Copyright can make the goal of providing the data in a form useful to the reader very difficult, so you might need to get the copyright clearance or find other creative ways of providing the data.

Data Reduction

In qualitative data analysis, data collection and data analysis are often not two distinct research phases. Rather, as researchers collect data, they begin to develop themes, ask analytical questions, write theoretical memos, and otherwise begin the work of analysis. And when researchers are analyzing data, they may find they need to go back and collect more to flesh out certain areas that need further elaboration (Taylor, Bogdan, and DeVault 2016). But as researchers move further towards analysis, one of the first steps is reading through all of the data they have collected. Many qualitative researchers recommend taking notes on the data and/or annotating it with simple notations like circles or highlighting to focus your attention on those passages that seem especially fruitful for later focus (Saldaña 2016). This is often called "pre-coding." Other approaches to pre-coding include noting hypotheses about what might emerge elsewhere in the data, summarizing the main ideas of each piece of data and annotating it with details about the respondent or circumstances of its creation, and taking preliminary notes about concepts or ideas that emerge.

This sort of work is often called "preliminary analysis," as it enables researchers to start making connections and working with themes and theoretical ideas, but before you get to the point of making actual conclusions. It is also a form of data reduction. In qualitative analysis, the volume of data collected in any given research project is often enormous, far more than can be productively dealt with in any particular project or publication. Thus, data reduction refers to the process of reducing large volumes of data such that the more meaningful or important parts are accessible. As sociologist Kristen Luker points out in her text *Salsa Dancing into the Social Sciences* (2008), what we are really trying to do is recognize patterns, and data reduction is a process of sifting through, digesting, and thinking about our data until we can see the patterns we might not have seen before. Luker argues that one important way to help ourselves see patterns is to talk about our data with others—lots of others, and not just other social scientists—until what we are explaining starts to make sense.

There are a variety of approaches to data reduction. Which of these are useful for a particular project depends on the type and form of data, the priorities of the researcher, and the goals of the research project, and so each researcher must decide for themselves how to proceed. One approach is summarization. Here,

researchers write short summaries of the data—summaries of individual interview transcripts, of particular days or weeks of fieldnotes, or of documents. Then, these summaries can be used for preliminary analysis rather than requiring full engagement with the larger body of data. Another approach involves writing memos about the data in which connections, patterns, or theoretical ideas can be laid out with reference to particular segments of the data. A third approach is annotation, in which marginal notes are used to highlight or draw attention to particularly important or noteworthy segments of the data. And Luker's suggestion of conversations about our data with others can be understood as a form of data reduction, especially if we record notes about our conversations.

One of the approaches to data reduction which many analysts find most useful is the creation of typologies, or systems by which objects, events, people, or ideas can be classified into categories. In constructing typologies, researchers develop a set of mutually-exclusive categories—no one can be placed into more than one category of the typology (Berg 2009)—that are, ideally, also exhaustive, so that no one is left out of the set of categories (an "other" category can always be used for those hard to classify). They then go through all their pieces of data or data elements, be they interview participants, events recorded in fieldnotes, photographs, tweets, or something else, and place each one into a category. Then, they examine the contents of each category to see what common elements and analytical ideas emerge and write notes about these elements and ideas.

One approach to data reduction which qualitative researchers often fall back on but which they should be extremely careful with is quantification. Quantification involves the transformation of non-numerical data into numerical data. For example, if a researcher counts the number of interview respondents who talk about a particular issue, that is a form of quantification. Some limited quantification is common in qualitative analysis, though its use should be particularly rare in ethnographic research given the fact that ethnographic research typically relies on one or a very small number of cases. However, the use of quantification should be constrained to those circumstances where it provides particularly useful or illuminating descriptive information about the data, and not as a core analytical tool. In addition, given that it is exceptionally uncommon for qualitative research projects to produce generalizable findings, any discussion of quantified data should focus on numbers rather than percents. Numbers are descriptive—"35 out of 40 interview respondents said they had argued with housemates over chores in the past week"—while percents suggest broader and more generalizable claims ("87.5% of respondents said they had argued with housemates over chores in the past week").

Qualitative Data Analysis Software

As part of the process of preparing data for analysis and planning an analysis strategy, many—though not all—qualitative researchers today use software applications to facilitate their work. The use of such technologies has had a profound impact on the way research is carried out, as have many technological changes over history. Take a much older example: the development of technology permitting for the audio recording of interviews. This technology made it possible to develop verbatim transcripts, whereas prior interview-based research had to rely on handwritten notes conveying the interview content—or, if the interviewer had significant financial resources, perhaps a stenographer. Recordings and verbatim transcripts also made it possible for researchers to minutely analyze speech patterns, specific word choices, tones of voice, and other elements that would not previously have been able to be preserved.

Today's technologies make it easier to store and retrieve data, make it faster to process and analyze data, and provide access to new analytical possibilities. On a basic level, software can allow for more sophisticated possibilities for linking data to memos and other documents. And there are a variety of other benefits (Adler and Clark 2008) to the use of software-aided analysis (often referred to as CAQDAS, or computer-aided qualitative data analysis software). It can allow for more attention to detail, more systematic analysis, and the use of more cases, especially when dealing with large data sets or in circumstances where some quantification is desirable. The use of CAQDAS can enhance the perception of rigor, which can be useful when bringing qualitative data to bear in settings where those using data are more used to quantitative analysis. When coding (to be discussed further in the chapter on qualitative coding), software enhances flexibility and complexity, and may enliven the coding process. And software can provide complex relational analysis tools that go well beyond what would be possible by hand.

However, there are limitations to the use of CAQDAS as well (Adler and Clark 2008). Software can promote ways of thinking about data that are disconnected from qualitative ideals, whether through reductions in the connection between data and context or the increased pressure to quantify. Each individual software application creates a specific model of the architecture of data and knowledge, and analysis may become shaped or constrained by this architecture. Coding schemes, taxonomies, and strategies may reflect the capacities available in and the structures prioritized by the software rather than reflecting what is actually happening in the data itself, and this can further homogenize research, as researchers draw from a few common software applications rather than from a wide variety of personal approaches to analysis. Software can also increase the psychic distance between the researcher or analyst and their data and reduce the likelihood of researchers understanding the limitations of their data. The tools available in CAQDAS applications tend to emphasize typical data rather than unusual data, and so outliers or negative cases may be missed. Finally, CAQDAS does not always reduce the amount of time that a research project takes, especially for newer users and in cases with smaller sets of data. This is because there can be very steep learning curves and prolonged set-up procedures.

The fact that this list of limitations is somewhat longer than the list of positives should not be understood as suggesting that researchers avoid CAQDAS-based approaches. Software truly does make forms of research possible that would not have been without it, speeds data processing tasks, and makes a variety of analytical tasks much easier to do, especially when they require attention to detail. And digital technologies, including both software applications and hardware devices, facilitate so much about how qualitative researchers work today. There are a wide variety of types of technological aids to the qualitative research purpose, each with different functions.

First of all, digital technologies can be used for capturing qualitative data. This may seem obvious, but as the example of audio recording above suggests, the development of technologies like audio and film recording, especially via cellphone or other small personal devices, led to profound changes in the way qualitative research is carried out as well as an expansion in the types of research that are possible. Other technologies that have had similar impacts include the photocopier and scanner, and more recently the possibility to use a cell phone to capture photographs of documents in archives (without the flash on to avoid damaging delicate items). Finally, videoconferencing software makes it possible to interview people who are halfway around the world, and most videoconferencing platforms have a built-in option to save a video record of the conversation, and potentially autocaption it. It's also worth noting that digital technologies provide access to sources of data that simply did not exist in the past, whether interviewing via videoconferencing, content analysis of social media, or ethnography of massively-multiplayer online games or worlds.

Software applications are very useful for data management tasks. The ability to store, file, and search electronic documents makes the management of huge quantities of data much more feasible. Storing metadata with files can help enormously with the management of visual data and other files. Word processing programs are also relevant here. They help us produce and revise text and reports, compile and edit our fieldnotes and transcriptions, write memos, make tables, count words, and search for and count specific words and phrases. Graphics programs can also facilitate the creation of graphs, charts, infographics, and other data displays. Finally, speech recognition programs aid our transcription process and, for some of us, our writing process.

Coding programs fall somewhere between data reduction and data analysis in their functions. Such software applications typically provide researchers with the ability to apply one or more codes to specific segments of text, search for and retrieve all segments that have had particular codes applied to them, and look at relationships between different codes. Some also provide data management features, allowing researchers to store memos, documents, and other materials alongside the coded text, and allow for interrater reliability testing (to be discussed in another chapter). Finally, there are a variety of data analysis tools. These tools allow researchers to carry out functions like organizing coded data into maps or diagrams, testing hypotheses, merging work carried out by different researchers, building theory, utilizing formal comparative methods, creating diagrams of networks, and others. Many of these features will be discussed in subsequent chapters.

Choosing the Right Software

There are so many programs out there that carry out each of the functions discussed above, with new ones appearing constantly. Because the state of technology changes all the time, it is outside the scope of this chapter to detail specific options for software applications, though online resources can be helpful in this regard (see, e.g., University of Surrey n.d.). But researchers still need to make decisions about which software to use. So, how do researchers choose the right qualitative software application or applications for their projects? There are four primary sets of questions researchers should ask themselves to help with this decision.

First, what functions does this research need and does this project require? As discussed above, programs have very different functions. In many cases, researchers may need to combine multiple programs to get access to all the functions they need. In other cases, researchers may need only a simple software application already available on their computers.

Second, researchers should consider how they use technology. There are a variety of questions that are relevant here. For example, what kind of device will be used, a desktop computer, laptop, tablet, or phone? What operating system, Windows, Mac/iOS, Chrome, or Android? How much experience and skill do researchers have with computers—do they need software applications that are very easy to use, or can they handle command-line interfaces that require some programming skills? Do they prefer software that is installed on their devices or a cloud-based approach? And will the researcher be working alone or as part of a team where multiple people need to contribute and share access to the same materials?

What type of data will be used? Will it be textual, visual, audio, or video? Will data come from multiple sources and styles or will it all be consistent? Is the data organized or free-form? What is the magnitude of the data that will be analyzed?

Finally, what resources does the researcher already have available? What software can they access, whether already available on their personal computing devices or via licenses provided by their employer or college/university? What degree of technical support can they access, and are technical support personnel familiar with CAQDAS? And how much money do they have available to pay for software on a one-time or ongoing basis? Note that some software can be purchased, while other software is provided as a service with a monthly subscription fee. And even when software is purchased, licenses may only provide access for a limited time period such as a year. Thus, both short-term and long-term financial costs and resource availability should be assessed prior to committing to a software package.

Exercises

1. Transcribe about 10 minutes of an audio interview—one good source might be your local NPR station's website. Be sure that your transcription is an *exact record* of what was said, including any pauses, laughter, vulgarities, or other kinds of things you might not typically write in an academic context, and that you transcribe both questions and responses. What was it like to complete this exercise?
2. Use the course listings at your college or university as a set of data. Develop a typology of different types of courses—*not* based on the department or school offering them or the course number alone—and classify courses within this typology. What does this exercise tell you about the curriculum at your college or university?
3. Review the notes, documents, and other materials you have already collected from this course and develop a new system of file management for them, with digital or physical folders, subfolders, and labels or file names that make items easy to locate.

This page titled 6.7.1: Qualitative Analysis- Preparing and Managing Qualitative Data is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Mikaila Mariel Lemonik Arthur via source content that was edited to the style and standards of the LibreTexts platform.

- **4.2: Preparing and Managing Qualitative Data** by Mikaila Mariel Lemonik Arthur is licensed CC BY-NC-SA 4.0. Original source: https://pressbooks.ric.edu/socialdataanalysis/.

6.7.2: Qualitative Coding

Codes are words or phrases that capture a central or notable attribute of a particular segment of text or visual data (Saldaña 2016). Coding, then, is the process of applying codes to texts or visuals. It is one of the most common strategies for data reduction and analysis of qualitative data, though many qualitative projects do not require or use coding. This chapter will provide an overview of approaches based in coding, including how to develop codes and how to go through the coding process.

In order to understand coding, it is essential to think about what it means for something to be a code. To analogize to social media, codes might function a bit like tags or hashtags. They are words or phrases that convey content, ideas, perspectives, or other key elements of segments of text. Codes are not the same as themes. Themes are broader than codes—they are concepts or topics around which a discussion, analysis, or text focuses. Themes are more general and more explanatory—often, once we code, we find themes emerge as ideas to explore in our further analysis (Saldaña 2016). Codes are also different from descriptors. Descriptors are words or phrases that describe characteristics of the entire text and/or the person who created it. For example, if we note the profession of an interview respondent, whether an article is news or opinion, or the type of camera used to take a photograph, those would be descriptors. Saldaña (2016) instead calls these attributes. The term attributes more typically refers to the possible answer choices or options for a variable, so it is possible to think about descriptors as variables (or perhaps their attributes) as well.

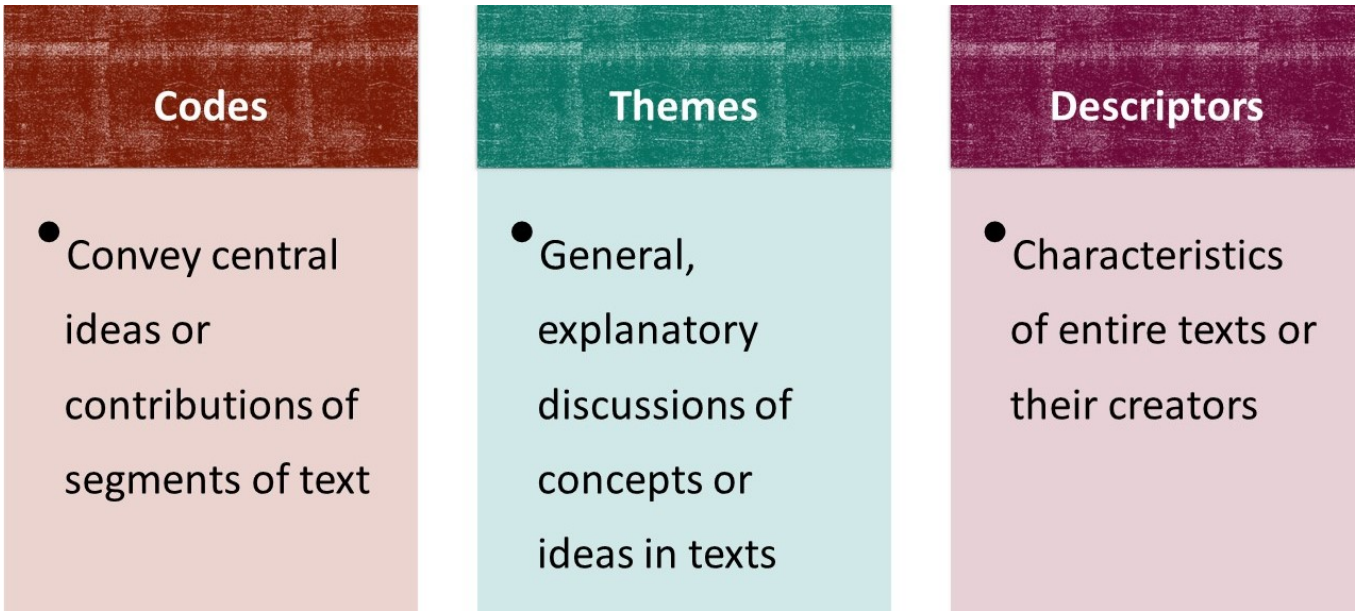


Figure 1. Codes vs. Themes v.s. Descriptors

Let's consider an example. Imagine that you were conducting an interview-based study looking at minor-league athletes' workplace experiences and later-life career plans. In this study, themes might be broad ideas like "aspirations" or "work experiences." There would be a vast array of codes, but they might include things like "short-term goals," "educational plans," "pay," "team bonding," "travel," "treatment by managers," "family demands," and many more. Descriptors might include the athlete's gender and what sport they play.

Developing a Coding System

While all approaches to coding have in common the idea that codes are applied to segments of text or visuals, there are many different ways to go about coding. These approaches differ in terms of when they occur during the research process and how codes are developed. First of all, there is a distinction between first- and second-cycle coding approaches (Saldaña 2016). First-cycle coding happens early in the research process and is really a bridge from data reduction to data analysis, while second-cycle coding occurs later in the research process and is more analytical in nature. Another version of this distinction is the comparison between rough, analytic, and focused coding. Rough coding is really part of the process of data reduction. It often involves little more than putting a few words near each segment of text to make clear what is important in that segment, with the approach being further refined as coding continues. In contrast, analytic coding involves more detailed techniques designed to move towards the

development of themes and findings. Finally, focused coding involves selecting ideas of interest and going back and re-coding your texts to orient your approach more specifically around these ideas (Bergin 2018).

A second set of distinctions concerns whether the data drives the development of codes or whether codes are instead developed in advance. If codes are determined in advance, or predetermined, researchers develop a set of codes based on their theory, hypothesis, or research question. This sort of coding is typically called deductive coding or closed coding. In contrast, open coding or inductive coding refers to a process in which researchers develop codes based on what they observe in their data, grounding their codes in the texts. This second approach is more common, though by no means universal, in qualitative data analysis. In both types of coding, however, researcher may rely upon ideas generated by writing theoretical memos as they work through the connections between concepts, theory, and data (Saldaña 2016).

Finally, a third set of distinctions focuses on *what* is coded. Manifest coding refers to the coding of surface-level and easily observable elements of texts (Berg 2009). In contrast, latent coding is a more interpretive approach based on looking deeply into texts for the meanings that are encoded within or symbolized by them (Berg 2009). For example, consider a research project focused on gender in car advertisements. A manifest approach might count the number of men versus women who appear in the ads. A latent approach would instead focus on the use of gendered language and the extent to which men and women are depicted in gender-stereotyped ways.

Researchers need to answer two more questions as they develop their coding systems. First, *what* to code, and second, *how many* codes. When thinking about what to code, researchers can look at the level of individual words, characters or actors in the text, paragraphs, entire textual items (like complete books or articles), or really any unit of text (Berg 2009), but the most useful procedure is to look for chunks of words that together express a thought or idea, here referred to as “segments of text” or “textual segments,” and then code to represent the ideas, concepts, emotions, or other relevant thoughts expressed in those chunks.

How many codes should a particular coding system have? There is no simple answer to this question. Some researchers develop complex coding systems with many codes and may have over a hundred different codes. Others may use no more than 25, perhaps fewer, even for the same size project (Saldaña 2016). Some researchers nest codes into code trees, with several related “child” codes (or subcodes) under a single “parent” code. For example, a code “negative emotions” could be the parent code for a series of codes like “anger,” “frustration,” “sadness,” and “fear.” This approach enables researcher to use a smaller or larger number of codes in their analysis as seems fit after coding is complete. While there is no formula for determining the right number of codes for a particular project, researchers should be attentive to overgrowth in the number of codes. Codes have limited analytical value if they are used only once or twice—if a coding system includes many codes that are applied only a small number of times, consider whether there are larger categories of codes that might be more useful. Occasionally, there are codes worth keeping but applying rarely, for example when there is a rare but important phenomenon that arises in the data. But for the most part, codes should be used with some degree of frequency in order for them to be useful for uncovering themes and patterns.

Types of Codes

A wide variety of different types of codes can be used in coding systems. The discussion below, which draws heavily on the work of Saldaña (2016), details a variety of different approaches to coding and code development. Researchers do not need to choose just one of these approaches—most researchers combine multiple coding approaches to create an overall system that is right for the texts they are coding and the project they are conducting. The approaches detailed here are presented roughly in order of the degree of complexity they represent.

At the most basic level is descriptive coding. Descriptive codes are nouns or phrases describing the content covered in a segment of text or the topic the segment of text focuses on. All studies can use descriptive coding, but it often is less productive of rich data for analysis than other approaches might be. Descriptive coding is often used as part of rough coding and data reduction to prepare for later iterations of coding that delve more deeply into the texts. So, for instance, that study of sexism in advertisements might involve some rough coding in which the researcher notes what type of product or service is being advertised in each advertisement.

Structural coding, in contrast, attends more closely to the research question rather than to the ideas in the text. In structural coding, codes indicate which specific research question, part of a research question, or hypothesis is being addressed by a particular segment of text. This may be most useful as part of rough coding to help researchers ensure that their data addresses the questions and foci central to their project.

In vivo coding captures short phrases derived from participants' own language, typically action-oriented. This is particularly important when researchers are studying subcultural groups that use language in different ways than researchers are accustomed to and where this language is important for subsequent analysis (Manning 2017). In this approach, researchers choose actual portions of respondents' words and use those as codes. In vivo coding can be used as part of both rough and analytical coding processes.

A related approach is process coding, which involves "the use of gerunds to label actual or conceptual actions relayed by participants" (Saldaña 2016:77). (Gerunds are verb forms that end in -ing and can function grammatically as if they are nouns when used in sentences). Process coding draws researchers' attention to actions, but in contrast to in vivo coding it uses the researcher's vocabulary to build the coding system. So, for instance, in the study of minor league athletes discussed earlier in the chapter, process codes might include "traveling," "planning," "exercising," "competing," and "socializing."

Concept coding involves codes consisting of words or short phrases that represent broader concepts or ideas rather than tangible objects or actions. Sticking with the minor league athletes example, concept codes might include "for the love of the game," "youth," and "exploitation." A combination of concept, process, and descriptive coding may be useful if researchers want their coding system to result in an inventory of the ideas, objects, and actions discussed in the texts.



Figure 2. A Selection of "Emoticodes"

Emotion codes are codes indicating the emotions participants discuss in or that are evoked by a segment of text. A more contemporary version of emotion codes relies on "emoticodes" or the emoji that express specific kinds of emotions, as shown in Figure 2.

Values coding involves the use of codes designed to represent the "perspectives or worldview" of a respondent by conveying participants' "values, attitudes, and beliefs" (Saldaña 2016:131). For example, a project on elementary school teachers' workplace satisfaction might include values codes like "equity," "learning," "commitment," and "the pursuit of excellence." Do note that choices made in values coding are, even more so than in other forms of coding, likely to reflect the values and worldviews of the coder. Thus, it can be essential to use a team of multiple coders with different backgrounds and perspectives in order to ensure a values coding approach that reflects the contents of the texts rather than the ideas of the coders.

Versus coding requires the construction of a series of binary oppositions and then the application of one or the other of the items in the binary as a code to each relevant segment of text. This may be a particularly useful approach for deductive coding, as the researcher can set out a series of hypothesized binaries to use as the basis for coding. For example, the project on elementary school teachers' workplace satisfaction might use binaries like feeling supported vs. feeling unsupported, energized vs. tired, unfulfilled needs vs. fulfilled needs, kids ready to learn vs. kids needing services, academic vs non-academic concerns, and so on.

Evaluation coding is used to signify what is and is not working in the policy, program, or endeavor that respondents are discussing or that the research focuses on. This approach is obviously especially useful in evaluation research designed to assess the merit or functioning of particular policies or programs. For example, if the project about elementary school teachers was part of a mentoring program designed to keep new teachers in the education profession, codes might include "future orientation" to flag portions of the text in which teachers discuss their longer-term plans and "mentor/mentee match" to flag portions in which they explore how they feel about their mentors, both key elements of the program and its goals.

There are a variety of other approaches more common outside of sociology, such as dramaturgical coding, which is a coding approach that treats interview transcripts or fieldnotes as if they are scripts for a play, coding such things as actors, attitudes, conflicts, and subtexts; coding approaches relying on terms and ideas from literary analysis; and those drawn from communications studies, which focus on facets of verbal exchange. Finally, some researchers have outlined very specific coding strategies and

procedures such that someone else could pick up their methods and apply them exactly. This sort of approach is typically deductive, as it requires the advance specification of the decisions that will be made about coding.

Some coding strategies incorporate measures of weight or intensity, and this can be combined with many of the approaches detailed above. For example, consider a project collecting narratives of people's experiences with losing their jobs. Respondents might include a variety of emotional content in their narratives, whether sadness, fear, stress, relief, or something else. But the emotions they discuss will vary not only in type, they will also vary in extent. A worker who is fired from a job they liked well enough but who knows they will be able to find another job soon may express sadness while a worker whose company closed after she worked there for 20 years and who has few other equivalent employment opportunities in the region may express devastation. Code weights help account for these differences.

A final question researchers must consider is whether they will apply only one code per segment of text or will permit overlapping codes. Overlapping codes make data analysis more complex but can facilitate the process of looking for relationships between different concepts or ideas in the data.

Codebooks

As a coding system is developed and certainly upon its completion, researchers create documents known as codebooks. As is the case with survey research, codebooks lay out the details of how the measurement instrument works to capture data and measure it. For surveys, a codebook tells researchers how to transform the multiple-choice and short-answer responses to survey questions into the numerical data used for quantitative analysis. For qualitative coding, codebooks instead explain when and how to use each of the codes included in the project. Codebooks are an important part of the coding process because they remind the researcher, and any other coders working on the project, what each code means, what types of data it is meant to apply to, and when it should and should not be used (Luker 2008). Even if a researcher is coding without others, it is easy to lose sight of what you were thinking when you initially developed your coding system, and so the codebook serves as an important reminder.

For each code, the codebook should state the name of the code, include a couple of sentences describing the code and what it should be used for, any information about when the code should *not* be used, examples of both typical and atypical conditions under which the code would be used, and a discussion of the role the code plays in analysis (Saldaña 2016). Codebooks thus serve as instruction manuals for when and how to apply codes. They can also help researchers think about taxonomies of codes as they organize the code book, with higher-level ideas serving as categories for groups of child, or more precise, codes.

The Process of Coding

So, what does the process of coding look like? While qualitative research can and does involve deductive approaches, the process that will be detailed here is an inductive approach, as this is more common in qualitative research. This discussion will lay out a series of steps in the coding process as well as some additional questions researchers and analysts must consider as they develop and carry out their coding.

The first step in inductive coding is to completely and thoroughly read through the data several times while taking detailed notes. To Saldaña (2016), the most important question to ask during this initial read is what is especially interesting or surprising or otherwise stands out. In addition, researchers might contemplate the actions people take, how people go about accomplishing things, how people use language or understand the world, and what people seem to be thinking. The notes should include anything and everything—objects, people, emotions, actions, theoretical ideas, questions—really anything, whether it comes up again and again in the data or only once, though it is useful to flag or highlight those concepts that seem to recur frequently in the data.

Next, researchers need to organize these notes into a coding system. This involves deciding which coding approach(es) to incorporate, whether or not to use parent and child codes, and what sort of vocabulary to use for codes. Remember that readers will not see the coding system except insofar as the researcher chooses to convey it, so vocabulary and terms should be chosen based on the extent to which they make sense to the research team. Once a coding system has been developed, the researcher must create a codebook. If paper coding will be used, a paper codebook should be created. If researchers will be using CAQDAS, or computer-aided qualitative data analysis software, to do their coding, it is often the case that the codebook can be built into the software itself.

Next, the researcher or research team should rough code, applying codes to the text while taking notes to reflect upon missing pieces in the coding system, ways to reorganize the codes or combine them to enhance meaning, and relevant theoretical ideas and

insights. Upon completing the rough coding process, researchers should revise the coding system and codebook to fully reflect the data and the project's needs.

At this point, researchers are ready to engage in coding using the revised codebook. They should always have someone else code a portion of the texts—usually a minimum of 10%—for interrater reliability checks, and if a larger research team is used, 10% of the texts should be coded in common by all coders who are part of the research team. Even in cases where researchers are working alone, it truly strengthens data analysis to be able to check for interrater reliability, so most analysts suggest having a portion of the data coded by another coder, using the codebook. If at all possible, additional coding staff should not be told what the hypothesis or research question is, as one of the strengths of this approach is that additional coding staff will be less likely to be influenced by preexisting ideas about what the data should show (Luker 2008). There are various quantitative measures, such as Chronbach's alpha and Kappa, that researchers use to calculate interrater reliability, the measure of how closely the ratings of multiple coders correspond. All coders should keep detailed notes about their coding process and any obstacles or difficulties they encounter.

How do researchers know they are done coding? Not just because they have gone through each text once or twice! Researchers may need to continue repeating this process of revision and re-coding until additional coding does not reveal anything more. This repetition is an essential part of coding, as coding always requires refinement and rethinking (Saldaña 2016). In Berg's (2009:354-55) words, it is essential to "code minutely," beginning with a rough view of the entire text and then refining as you go until you are examining each detail of a text. Then, researchers think about why and how they developed their codes and what jumps out at them as important from the research as they delve into findings, making sure that nothing has been left out of the coding process before they move towards data analysis.

One interesting question is whether the identities and standpoints (as discussed in the chapter "The Qualitative Approach") of coders matter to the coding process. Eduardo Bonila-Silva (Zuberi and Bonilla-Silva 2008:17) has described how, after a presentation discussing his research on racism, a colleague asked whether the coders were White or Black—and he responded by asking the colleague "if he asked such questions across the board or only to researchers saying race matters." As Bonilla-Silva's question suggests, race (like other aspects of identity and experience, such as gender, immigration status, disability status, age, and social class, just to name a few) very well might shape the way coders see and understand data, functioning as part of a particular coding filter (Saldaña 2016). But that shaping extends broadly across all issues, not just those we might assume are particularly salient in relationship to identities. Thus, it is best for research teams to be diverse so as to ensure that a variety of perspectives are brought to bear on the data and that the findings reflect more than just a narrow set of ideas about how the world works.

Coding and What Comes After

If researchers will code by hand, they will need multiple copies of their data, one for reference and one for writing on (Luker 2008). On the copy that will be written on, researchers use a note-taking system that makes sense to them—whether different-colored markers, Roman numerals in the margins, a complex series of sticky notes, or whatever—to mark the application of various codes to sections of your data. You can see an example of what hand coding might look like in Figure 3 below, which is taken from a study of the comments faculty members make on student writing. Segments of text are highlighted in different colors, with codes noted in the margins next to the text. You can see how codes are repeated but in different combinations. Once the initial coding process is complete, researchers often cut apart the pieces of paper to make chunks of text with individual codes and sort the pieces of paper by code (if multiple codes appear in individual chunks of text, additional copies might be needed). Then, each pile is organized and used as the basis for writing theoretical memos. Another option for coding by hand is to use an index sheet (Berg 2009). This approach entails developing a set of codes and categories, arranging them on paper, and entering transcript, page, and paragraph information to identify where relevant quotes can be found.

For more complex analytical processes, researchers will likely want to use software, though there are limitations to software. Luker (2008), for instance, argues that when coding manually, she tends to start with big themes and only breaks them into their constituent parts later, while coding using software leads her to start with the smallest possible codes. (One solution to this, offered by some software packages, is upcoding, where a so-called "parent" code is simultaneously applied to all of the "child" codes under it. For instance, you might have a parent code of "activism" and then child codes that you apply to different kinds of activism, whether protest, legislative advocacy, community organizing, or whatever.)

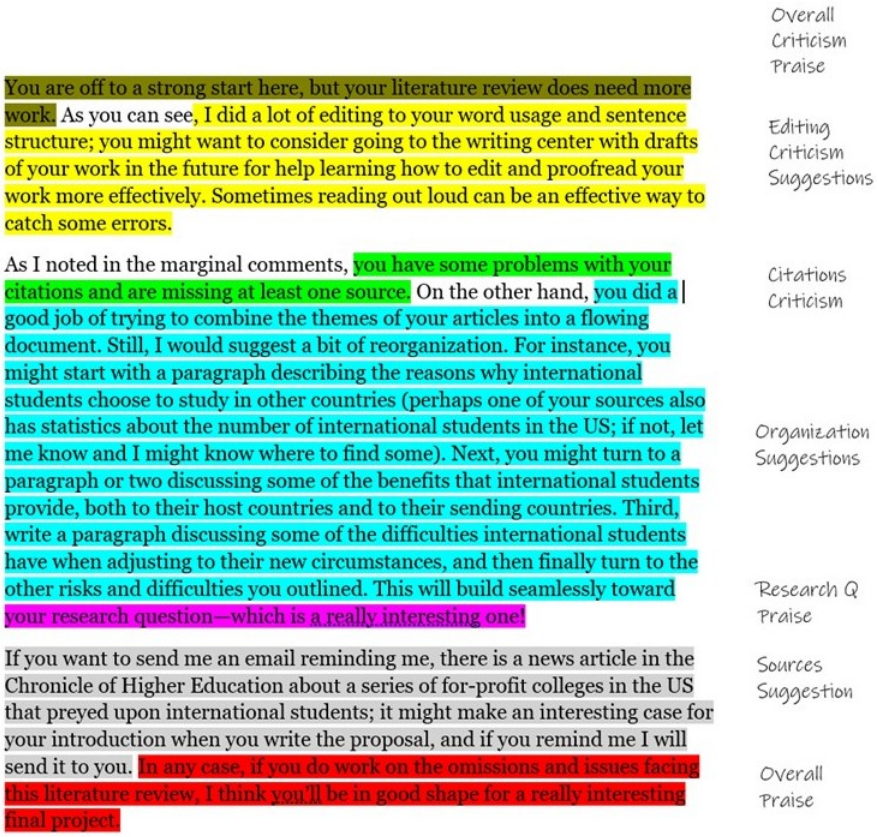


Figure 3. An Example of Hand Coding

Coding does not stand on its own, and thus simply completing the coding process does not move a research project from data to analysis. While the analysis process will be discussed in more detail in a subsequent chapter, there are several steps researchers take alongside coding or immediately after completing coding that facilitate analysis and are thus useful to discuss in the context of coding. Many of these are best understood as part of the process of data reduction. One of the most important of these is categorizing codes into larger groupings, a step that helps to enable the development of themes. These larger groupings, sometimes called "parent" codes, can collapse related but not identical ideas. This is always useful, but it is especially useful in cases where researchers have used a large number of codes and each one is applied only a few times. Once parent codes have been created, researchers then go back and ensure that the appropriate parent code is assigned to all segments of text that were initially coded with the relevant "child" codes (a step that can be automated in CAQDAS). If appropriate, researchers may repeat this process to see if parent codes can be further grouped. An alternative approach to this grouping process is to wait until coding is complete, and then create more analytical categories that make sense as thematic groupings for the codes that have been utilized in the project so far (Saldaña 2016).

There are a variety of other approaches researchers may take as part of data reduction or preliminary analysis after completing coding. They may outline the codes that have occurred most frequently for specific participants or texts, or for the entire body of data, or the codes that are most likely to co-occur in the same segment of text or in the same document. They may print out or photocopy documents or segments of text and rearrange them on a surface until the arrangement is analytically meaningful. They may develop diagrams or models of the relationships between codes. In doing this, it is especially helpful to focus on the use of verbs or other action words to specify the nature of these relationships—not just stating that relationships exist, but exploring what the relationships do and how they work.

In inductive coding especially, it is often useful to write theoretical and analytical memos while coding occurs, and after coding is completed it is a good time to go back and review and refine these memos. Here, researchers both clearly articulate to themselves how the coding process occurred and what methodological choices they made as well as what preliminary ideas they have about analysis and potential findings. It can be very useful to summarize one's thinking and any patterns that might have been observed so far as a step in moving towards analysis. However, it is extremely important to remember the data and not just the codes. Qualitative researchers always go back to the actual text and not just the summaries or categories. So a final step in the process of

moving toward analysis might be to flag quotes or data excerpts that seem particularly noteworthy, meaningful, or analytically useful, as researchers need these examples to make their data come alive during analysis and when they ultimately present their results.

Becoming a Coder

This chapter has provided an overview of how to develop a coding system and apply that system to the task of conducting qualitative coding as part of a research project. Many new researchers find it easy—if sometimes time-consuming and not always fascinating—to get engaged with the coding process. But what does it take to become an effective coder? Saldaña (2016) emphasizes personality attributes and skills that can help. Some of these are attributes and skills that are important for anyone who is involved in any aspect of research and data analysis: organization, to keep track of data, ideas, and procedures; perseverance, to ensure that one keeps going even when the going is tough, as is often the case in research; and ethics, to ensure proper treatment of research participants, appropriate data security behaviors, and integrity in the use of sources. In most aspects of data analysis, creativity is also important, though there are some roles in quantitative data analysis that require more in the way of technical skills and ability to follow directions. In qualitative data analysis, creativity remains important because of the need to think deeply and differently about the data as analysis continues. Flexibility and the ability to deal with ambiguity are much more important in qualitative research, as the data itself is more variable and less concrete; quantitative research tends to place more emphasis on rules and procedures. A final strength that is particularly important for those working in qualitative coding is having a strong vocabulary, as vocabulary both helps researchers understand the data and enhances their ability to create effective and useful coding systems. The best way to develop a stronger vocabulary is to read more, especially within your discipline or field but broadly as well, so researchers should be sure to stay engaged with reading, learning, and growing.

Reading, learning, and growing, along with a lot of practice, is of course how researchers enhance their data collection, coding, and data analysis skills, so keep working at it. Qualitative research can indeed be easy to get started with, but it takes time to become an expert. Put in the time, and you, too, can become a skilled qualitative data analyst.

Exercises

1. For each of the following words or phrases, consider whether it is most likely to represent a code, a theme, or a descriptor. Explain your response.
   - Female respondent
   - Energized
   - The relationship between poverty and social control
   - Creative
   - A teacher
   - The process of divorce
   - Social hierarchies
   - Grief
2. Pick a research topic you find interesting and determine which of the approaches to coding detailed in this chapter might be most appropriate for your topic, then write a paragraph about why this approach is the best.
3. Sticking with the same topic you used to respond to Exercise 2, brainstorm some codes that might be useful for coding texts related to this topic. Then, write appropriate text for a codebook for each of those codes.
4. Select a hashtag of interest on a particular social media site and randomly sample every other post using that hashtag until you have selected 15 tweets. Then inductively code those posts and engage in summarization or classification to determine what the most important themes they express might be.
5. Create a codebook based on what you did in Exercise 4. Exchange codebooks and tweets with a classmate and code each other's tweets according to the instructions in the codebook. Compare your results—how often did your coding decisions agree and how often did they disagree? What does this tell you about interrater reliability, codebook construction, and coder training?

Media Attributions

- codes themes descriptors © Mikaila Mariel Lemonik Arthur is licensed under a CC BY-NC (Attribution NonCommercial) license
- Emoticodes © AnnaliseArt is licensed under a CC BY (Attribution) license
- Hand Coding Example © Mikaila Mariel Lemonik Arthur is licensed under a CC BY-NC-ND (Attribution NonCommercial NoDerivatives) license

---

This page titled 6.7.2: Qualitative Coding is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Mikaila Mariel Lemonik Arthur via source content that was edited to the style and standards of the LibreTexts platform.

- **4.3: Qualitative Coding** by Mikaila Mariel Lemonik Arthur is licensed CC BY-NC-SA 4.0. Original source: https://pressbooks.ric.edu/socialdataanalysis/.

6.7.3: From Qualitative Data to Findings

Learning Objectives

1. Identify some options for making conclusions and presenting findings of qualitative research.
2. Using information from previous chaptesr, compare and contrast making conlusions and presenting your findings between qualitative and quantitative research.

So far in this section on qualitative research, we have covered various approaches to managing, preparing, reducing, and otherwise interacting with qualitative data. Because of the iterative and cyclical nature of the qualitative research process, it is not accurate to say that these steps come *before* analysis. Rather, they are an integral *part of* analysis. Yet there are procedures and methods for moving from the data to findings that are essential to completing a qualitative data analysis project. This section will discuss how to move towards conclusions, and suggest approaches for testing these conclusions to ensure that they hold up to scrutiny.

But before the final stages of analysis occur, researchers do need to take a step back and ensure that data collection really is finished—or at least finished enough for the particular phase of analysis and publication the researcher is working on, in the case of a very long-term project. How do qualitative researchers know their data collection has run its course? Well, in some cases they know because they have exhausted their sample. If a project was designed to include interviews of forty respondents or the collection of 500 social media posts, then it is complete when those interviews have been conducted or those social media posts have been saved. In other cases, researchers know that data collection is complete when they reach saturation, or the point in the research process where continuing to engage in data collection no longer yields any new insights. This way of concluding data collection is more common in ethnographic work or work with archival documents.

In addition, since qualitative research often results in a truly enormous amount of data, one of the key tasks of analysis is finding ways to select the most central or important ideas for a given project. Keep in mind that doing so does not mean dismissing other ideas as unimportant. Rather, these other ideas may become the basis for another analysis drawing on the same data in the future. But one project, even an entire book, often cannot contend with the full body of data that a researcher or research team has collected. That is why it is important to engage in data reduction before or alongside the analysis process.

As researchers move from data towards findings, it is essential that they remember that, unlike much quantitative research, most qualitative research draws on small or otherwise unrepresentative samples, the findings also cannot be generalized. Thus, while the findings of qualitative research may be suggestive of general patterns, they must be regarded as just that: only suggestive.

Similarly, qualitative research cannot demonstrate causation. Demonstrating causation requires three elements:

- Association, or the ability to show a clear (statistical) relationship between the phenomena, concepts, or processes in question,
- Temporal order, or the ability to show that the supposed cause came earlier in time than the supposed effect, and
- Elimination of alternatives, or the ability to show that there is *no possible alternative explanation* that could account for the phenomena in question.

While qualitative research *can* demonstrate association and temporal order, it cannot eliminate all alternative explanations—only well-designed and properly-controlled laboratory experiments can do so. This will be discussed in detail in later sections of this textbook. What this means now is that qualitative researchers and quantitative researchers who are not relying on data from controlled laboratory experiments need to take care to stay away from arguments suggesting that their analysis has proven anything or shown a causal relationship. However, qualitative researchers can locate evidence that *supports* the argument that a relationship is causal, leading to sentences like “This provides evidence suggestive of a causal relationship between X and Y.”

Organizing Data for Conclusions

Theoretical Memos

Memo-writing is as a strategy for data reduction, but memos (or theoretical notes) can be a key element in the process of moving from data to findings. Memos and theoretical notes are what the researcher writes for themselves in which they work through ideas from the data, connect examples and excerpts to themes and theories, pose new questions, and illuminate potential findings. Memos can serve as a way to move from the rich, contextual detail of qualitative data—detail that can sometimes be overwhelming—

towards the broader issues and questions that motivate a study. Initial memos are often drafted while data collection is still going on. For instance, a researcher might write reflective memos that integrate preliminary thoughts and ideas about data, help clarify concepts central to the research project, or pull together disparate chunks of data. But additional, and more specifically analytical, memos come later in the process.

There are techniques and prompts for this writing that won't be covered in this textbook, but you can learn more about the options when you're ready and then try out different suggestions to see what works for you. The memos may focus on a variety of topics and ideas, including reflecting on the researcher's role and thought processes; contemplating the research question, potential answers, and shifts in the research focus; noting choices about coding strategies, the coding process, and choices made during coding; clarifying ethical or methodological issues that have arisen in the course of the research; considering what further research may need to be done in the future; and trying out ideas that may become part of the final analysis or write-up (Saldaña, 2016). What is integral to the memo-writing approach is that in drafting memos, researchers can come to better understand the ideas shaping their study and what their data is saying in response to their research question.

Data Displays

Theoretical memos are good for interrogating your data and yourself. There are additional techniques for organizing the data to identify patterns and themes. analyzing One such tool for moving from data to findings is called a data display. Data displays are diagrams, tables, and other items that enable researchers to visualize and organize data so that it is possible to clearly see the patterns, comparisons, processes, and themes that emerge. These patterns, comparisons, processes, and themes then enable the researcher to articulate conclusions and findings. This textbook will not discuss data displays, but researchers interested in learning more about the development and use of data displays in qualitative research can consult Miles and Huberman's (1994) thorough and comprehensive sourcebook. It is important to note that even as the displays enable the drawing of conclusions through looking for patterns and themes, this is not sufficient to support analysis and writeup. The display is simply a tool for analysis and understanding and cannot fully encapsulate the findings or the richness of the data. Thus, the analytical process always needs to return to the data itself, and whatever researchers write up or present needs to include stories, quotes, or excerpts of data to bring the concepts and ideas that are part of the study alive. The display is just that and does not itself contain or encapsulate the conclusions/analysis.

Narrative Approach

In many cases, researchers rely on the crafting of narratives as a key part of the analytical process. Such narratives may be used only for analysis or they may be integrated into the ultimate write-up of the project. There are a variety of different narrative approaches (Grbich, 2007). Narratives can be descriptive—simply using the data to tell a story—or theoretical and analytical, where data is used to illustrate concepts and ideas (Taylor et al., 2016).

There are a variety of other cautions and concerns that researchers should keep in mind as they build and evaluate their conclusions and findings. The term anecdotalism refers to the practice of treating anecdotes, or individual stories or events, as if they themselves are sufficient data upon which to base conclusions. In other words, researchers who are engaging in anecdotalism present snippets of data to illustrate or demonstrate a phenomenon without any evidence that these particular snippets are representative. While it is natural for researchers to include their favorite anecdotes in the presentation of their results, this needs to be done with attention to whether the anecdote illustrates a broader theme expressed throughout the data or whether it is an outlier. Without this attention, the use of anecdotes can quickly mislead researchers to misplace their focus and build unsupported conclusions. One of the most problematic aspects of anecdotalism is that it can enable researchers to focus on particular data because it supports the research hypothesis, is aligned with researchers' political ideals, or is exotic and attention-getting, rather than focusing on data that is representative of the results. The practice of anecdotalism is at the foundation of some people's perceptions that qualitative research is not rigorous or methodologically sound. In reality, bad or sloppy qualitative research, including that suffering from anecdotalism, is not rigorous, just as bad quantitative research is also not rigorous.

Which Technique?

Whether to choose memo-writing, data display, or narrative approaches to analysis, or some combination of two or more of these approaches, is determined both by researchers' personal styles of research and analysis and by the nature, type, and complexity of the data. For instance, a multifaceted case study of an organization with multiple departments may have a story which is too complex for a narrative approach and thus requires the researcher to find other ways of simplifying and organizing the data. An

observation of a girls' soccer team over the course of a single season, though, might lend itself well to a narrative approach utilizing thick description. And interviews with people recovering from surgery about their experiences might best be captured through a narrative approach focusing on quotes from participants.

Making Conclusions and Testing Findings

Theoretical memos, data displays, and narratives are not themselves conclusions or findings. Rather, they are tools and strategies that help researchers move from data towards findings. So how do researchers use these tools and strategies in service of their ultimate goal of making conclusions? Most centrally, this occurs by looking for patterns, comparisons, associations, or categories. And patterns are probably the most common and useful of these. There are a variety of types of patterns that researchers might encounter or look for. These include finding patterns of similarities, patterns of predictable differences, patterns of sequence or order, patterns of relationship or association, and patterns that appear to be causal (Saldaña, 2016). If you are interested in qualitative research, the resources in this page's References are a great place to start learning about analyzing the qualitative data that you have collected, organized, and thought about.

Settling in on a set of findings does not mean a research project has been completed. Rather, researchers need to go through a process of testing, cross-checking, and verifying their conclusions to be sure they stand up to scrutiny. Researchers use a variety of approaches to accomplish this task, usually in combination. One of the most important is consulting with others. Researchers discuss their findings with other researchers and other professional colleagues as well as with participants or other people similar to the participants (Warren & Karner, 2015). These conversations give researchers the opportunity to test their logic, learn about questions others may have in regards to the research, refine their explanations, and be sure they have not missed obvious limitations or errors in their analysis. Presenting preliminary versions of the project to classmates, at conferences or workshops, or to colleagues can be particularly helpful, as can sharing preliminary drafts of the research write-up. Talking to participants or people like the participants can be especially important. While it is always possible that a research project will develop findings that are valid but that do not square with the lived experiences of participants, researchers should take care in such circumstances to responsibly address objections in ways that uphold the validity of both the research and the participants' experiences and to listen carefully to criticisms to be sure all potential errors or omissions in the analysis have been addressed.

But researchers need not rely only on others to help them assess their work. There are a variety of steps researchers can take as part of the research process to test and evaluate their findings. For instance, researchers can critically re-examine their data, being sure their findings are firmly based on stronger data: data that was collected later in the research process after early weaknesses in collection methods were corrected, first-hand observations rather than occurrences the researcher only heard about later, and data what was collected in conditions with higher trust. They should also attend to the issue of face validity, the type of validity concerned with whether the measures used in a study are a good fit for the concepts. Sometimes, in the course of analysis, face validity falls away as researchers focus on exciting and new ideas, so returning to the core concepts of a study and ensuring the ultimate conclusions are based on measures that fit those concepts can help ensure solid conclusions.

Qualitative researchers must always look for evidence that is surprising or that does not fit the model, theory, or predictions. If no such evidence appears—if it seems like all of the data conforms to the same general perspective—remember that the absence of such evidence does not mean the researcher's assumptions are correct. For example, imagine a research project designed to study factors that help students learn more math in introductory math classes. The researcher might interview students about their classes and find that the students who report that there were more visual aids used in their class all say that they learned a lot, while the students who report that their classes were conducted without visual aids say they did not learn so much. In this analysis, there may not have been any responses that did not fit this overall pattern. Clearly, then, something is going on within this population of students. But it is not necessarily the case that the use of visual aids impacts learning. Rather, the difference could be due to some other factor. Students might have inaccurate perception of how much they have learned and the use or lack of use of visual aids could impact these perceptions. Or visual aids might have a spurious relationship with students' perceptions of learning given some other variable, like the helpfulness of the instructor, that correlates with both perception and use of visual aids. Remember that a spurious relationship is a relationship in which two phenomena seem to vary in association with one another, but the observed association is not due to any causal connection between the two phenomena. Rather, the association is due to some other factor that is related to both phenomena but that has not been included in the analysis. If an association is observed, then the researcher should consider whether there might be alternative explanations that make more sense, turning back to the data as necessary. Indeed, researchers should always consider the possibility of alternative explanations, and it can be very helpful to ask others to suggest

alternative explanations that have not yet been considered in the analysis. Not only does doing this increase the odds that a project's conclusions will be reliable and valid, it also staves off potential criticism from others who may otherwise remain convinced that their explanations are more correct.

Presenting Findings

Researchers should always make clear to others how they carried out their research. Providing sufficient detail about the research design and analytical strategy makes it possible for other researchers to replicate the study, or carry out a repeat of the research designed to be as similar as possible to the initial project. Complete, accurate replications are possible for some qualitative projects, such as an analysis of historical newspaper articles or of children's videos, and thus providing the level of detail and clarity necessary for replication is a strength for such projects. It is far less possible for in-depth interviewing or observations to be replicated given the importance of specific contextual factors as well as the impact of interviewer effect. However, providing as much detail about methodological choices and strategies as possible, along with why these choices and strategies were the right ones for a given project, keeps the researcher and the project more honest and makes the approach more clear, as the goals of good research should include transparency.

Additionally, research projects involving multiple coders should have already undergone inter-rater reliability checks including at least 10% of the texts or visuals to be coded, and, if possible, even projects with only one coder should have conducted some inter-rater reliability testing. A discussion of the results of inter-rater reliability testing should be included in any publication or presentation drawing on the analysis, and if inter-rater reliability was *not* conducted for some reason this should be explicitly discussed as a limitation of the project. There are other types of limitations researchers must also clearly acknowledge, such as a lack of representativeness among respondents, small sample size, any issues that might suggest stronger-than-usual interviewer or Hawthorne effects, and other issues that might shape the reliability and validity of the findings.

Qualitative researchers must take care to present excerpts and examples from their data. When researchers do not do this and instead focus their write-up on summaries (or even numbers), readers are not able to draw their own conclusions about whether the data supports the findings. Respondents' actual words, words or images from documents, and ethnographers' first-hand observations are the true strength of qualitative research and thus it is essential that these things come through in the final presentation. Plus, if researchers focus on summaries or numbers, they may miss important nuances in their data that could more accurately shape the findings. On the other hand, researchers also must take care to avoid making overconclusions, or conclusions that go beyond what the data can support. Researchers risk making overconclusions when they assume data are representative of a broader or more diverse population than that which was included in the study, when they assume a pattern or phenomenon they have observed occurs in other types of contexts, and in similar circumstances when limited data cannot necessarily be extended to apply to events or experiences beyond the parameters of the study.

Another risk in qualitative research is that researchers might underemphasize theory. The role of theory marks one of the biggest differences between social science research and journalism. By connecting data to theory, social scientists have the ability to make broader arguments about social process, mechanisms, and structures, rather than to simply tell stories. Remember that one common goal in social science research is to focus on ordinary and everyday life and people, showing how—for instance—social inequality and social organizations structure people's lives, while journalism seeks stories that will draw attention.

Thinking Like a Researcher

So what ties all of these approaches to qualitative data analysis together? Among the most important characteristics is that the data needs to speak for itself. Yes, qualitative researchers may engage in data reduction due to the volume and complexity of the data they have collected, but they need to stay close enough to the data that it continues to shape the analysis and come alive in the write up.

Another very important element of qualitative research is reflexivity. Reflexivity, in the context of social science research, refers to the process of reflecting on one's own perspective and positionality and how this perspective and positionality shape "research design, data collection, analysis, and knowledge production" (Hsiung, 2008, 212). The practice of reflexivity is one of the essential habits of mind for qualitative researchers, and should be incorporated by quantitative researchers. While researchers should engage in reflexivity throughout the research process, it is important to engage in a specifically reflexive thought process as the research moves towards conclusions. Here, qualitative researchers consider what they were thinking about their project, methodology,

theoretical approach, topic, question, and participants when they began the research process, how these thoughts and ideas have or have not shifted, and how these thoughts and ideas—along with shifts in them—might have impacted the findings (Taylor et al., 2016). They do this by "turn[ing] the investigative lens away from others and toward themselves" (Hsiung, 2008, 213), taking care to remember that the data they have collected and the data reduction and analysis strategies they have pursued result in records of interpretations, not clear, objective facts. Reflexive research may also involve having additional researchers serve as a kind of check on the research processes to ensure they comport with researchers' goals and ethical priorities (Grbich, 2007).

Adjusting to the qualitative way of thinking can be challenging. In fact, doing research is hard by definition, and when it is done right, researchers are inevitably going to hit many obstacles and will frequently feel like they do not know what they are doing. This is not going to be because there is something wrong with the researcher! Rather, this is because that is how research works; this is particularly true for doing qualitative research as it is a totally different kind of research and analysis than what you may have learned in your science and psychology courses. Research involves trying to answer a question no one has answered before by collecting and analyzing data in a way no one has tried before.

This section of the chapter on analysis has highlighted a variety of strategies for moving from data to conclusions. In the quantitative research process, moving from data to conclusions really is the analysis stage of research. But in qualitative research, especially inductive qualitative research, the process is more iterative, and researchers move back and forth between data collection, data management, data reduction, and analysis. It is also important to note that the strategies and tools outlined here are only a small sampling of the possible analytical techniques qualitative researchers use—but they provide a solid introduction to the qualitative research process. As you practice qualitative research and develop your expertise, you will continue to find new approaches that better fit your data and your research style.

References

Grbich, C. (2013). *Qualitative data analysis: An introduction.* Sage. https://doi.org/10.4135/9781529799606

Hsiung, P.-C. (2008). Teaching reflexivity in qualitative interviewing. *Teaching Sociology, 36*(3), 211-226. https://doi.org/10.1177/0092055X0803600302

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications, Inc.

Saldaña, J. (2016). *The Coding Manual for Qualitative Researchers (3rd ed.).* Sage.

Taylor, S. J., Bogdan, R., & DeVault, M. L. (2016). *Introduction to qualitative research methods: A guidebook and resource* (4th ed.). John Wiley & Songs, Inc.

Warren, C. A. B. & Karner, T. X. (2015). *Discovering qualitative methods: Ethnography, interviews, documents, and images*. Oxford University Press.

This page titled 6.7.3: From Qualitative Data to Findings is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Mikaila Mariel Lemonik Arthur via source content that was edited to the style and standards of the LibreTexts platform.

- **4.4: From Qualitative Data to Findings** by Mikaila Mariel Lemonik Arthur is licensed CC BY-NC-SA 4.0. Original source: https://pressbooks.ric.edu/socialdataanalysis/.

6.8: Summary of Analyzing

Key Takeaways

- Every variable has a distribution—a way that the scores are distributed across the levels. The distribution can be described using a frequency table and histogram. It can also be described in words in terms of its shape, including whether it is unimodal or bimodal, and whether it is symmetrical or skewed.
- The central tendency, or middle, of a distribution can be described precisely using three statistics—the mean, median, and mode. The mean is the sum of the scores divided by the number of scores, the median is the middle score, and the mode is the most common score.
- The variability, or spread, of a distribution can be described precisely using the range and standard deviation. The range is the difference between the highest and lowest scores, and the standard deviation is the average amount by which the scores differ from the mean.
- These descriptive statistics tell the story of what happened in a study. Although inferential statistics are also important, it is essential to understand the descriptive statistics first.
- Graphs are a great way to display frequency distributions, and can be interpreted to help identify important characteristics of the distribution.
- There are different kinds of graphs; choose the type that fits the type of data that you're displaying and the purpose of the graph.
- Null Hypothesis Significance Testing (NHST) is the approach to analyzing quantitative research. To show that the means are different or that there is relationship, you want a small probability ($p<.05$) that the null hypothesis is true in the population based on this sample.
- NHST for comparing means generally uses one type of t-test or one type of ANOVA.
- Differences between groups or conditions are typically described in terms of the means and standard deviations of the groups or conditions or in terms of Cohen's *d* and are presented in bar graphs.
- Cohen's *d* is a measure of relationship strength (or effect size) for differences between two group or condition means. It is the difference of the means divided by the standard deviation. In general, values of ±0.20, ±0.50, and ±0.80 can be considered small, medium, and large, respectively.
- Correlations between quantitative variables are typically described in terms of Pearson's *r* and presented in line graphs or scatterplots.
- Correlations between variables can be used to identify what factors (IVs) are most related to a specific outcome (DV); this is called factor analysis.
- Just become two variables are correlated, does NOT mean that one variable *causes* the other variable!
- Qualitative research tends to collect and analyze data this primarily of words and images drawn from observation, interaction, interviewing, or existing documents. Because of this, qualitative research requires additional considerations for managing, organizing, analyzing, and interpreting this rich data, as well as additional considerations for presenting the findings.
- Analyzing quantitative or qualitative data requires designing high-quality research studies, integrity, and thinking like a researcher.

What's Next?

Now that you have some familiarity with finding research articles, as well as reading research articles and interpreting their results, the next chapter will focus on writing in APA Style.

If you'd like a more in-depth refresher of quantitative analyses, LibreTexts offers a selection of openly-licensed textbooks on statistics (https://stats.libretexts.org/), including textbooks for different social sciences (https://stats.libretexts.org/Bookshe...ied_Statistics).

Exercises

- Practice: Make a frequency table and line graph for each country's data in Table **6.8.1**, which represents scores on the Rosenberg Self-Esteem Scale for a sample of 10 Japanese university students and 10 university students in the United States. (Although hypothetical, these data are consistent with empirical findings, such as Schmitt and Allik (2005).)

Table 1.

Self-Esteem Scores by Country

Japan	United States
25	27
20	30
24	34
28	37
30	26
32	24
21	28
24	35
20	33
26	36

Table 6.8.1: Self-Esteem Scores by Country. (Copyright CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton)

- Practice: Answer the following questions based on the frequency line graph for the data above:
  - What does the x-axis measure?
  - What does the y-axis measure?
  - Is the line graph skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
  - What do you notice from the line graph? What pops out to you?
  - What does the figure make you wonder about?
  - What is a catchy headline?
  - How could you summarize the info from the figure in one sentence?
  - Who might want to know the information?
- Practice: The hypothetical data in Table 6.8.2 are extraversion scores and the number of Instagram followers for 15 university students. Make a scatterplot for these data.

Table 2.

Example Data of Extraversion and Instagram Followers

Extraversion	Instagram Followers
8	75
10	315
4	28
6	214
12	176
14	95
10	120
11	150
4	32
13	250
5	99
7	136
8	185
11	88
10	144

Table **6.8.2**: Extraversion and Instagram Followers. (Copyright CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton)

- Practice: Interpret the results from the data in Table **6.8.2** if $r(13)=0.42$, $p=0.12$
- Practice: Utilize either of the data sets here to practice analysing data from one of the linked websites that will analyze statistics for you from: https://statpages.info/
- Discussion: Locate a recent scholarly journal article in your field of study and read it. Do you think this article used a more positivist or more interpretivist paradigm of knowledge? Explain how you know, drawing on the key elements of these paradigms.
- Discussion: What do you think it means to do good research? Which of the various standards for good research do you think are most important to the topics or issues you are interested in? And what are some of the strategies you might employ to be sure your research lives up to these standards?
- Locate a work of long-form journalism about a topic of social science interest. Good publications to explore for this purpose include The Atlantic, The New Yorker, The New York Times Magazine, Vanity Fair, Slate, and longreads.com, among others. Summarize how the article might be different if it were an example of social science rather than journalism—what theory or theories might it draw on? What types of scholarly sources might it cite? How might its data collection have been different? How might data analysis have been conducted? What social science conclusions might it have reached?

References

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology, 89*, 623–642.

This page titled 6.8: Summary of Analyzing is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **12.5: Descriptive Statistics (Summary)** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

7: Formatting in APA Style

7.1: Prelude to Presenting Your Research

7.2: Introduction to American Psychological Association (APA) Style

7.3: General Paper Format

7.4: Formatting by Section

7.5: Presenting Results

7.5.1: Tables and Figures

7.6: Research Proposals

7.7: Summary and APA Style Resources

7: Formatting in APA Style is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

7.1: Prelude to Presenting Your Research

> *Research is complete only when the results are shared with the scientific community.*
> *-American Psychological Association*

Imagine that you have identified an interesting research question, reviewed the relevant literature, designed and conducted an empirical study, analyzed the data, and drawn your conclusions. There is still one more step in the process of conducting scientific research. It is time to add your research to the literature so that others can learn from it and build on it. Remember that science is a social and cumulative process—a large-scale collaboration among many researchers distributed across space and time. For this reason, it could be argued that unless you make your research public in some form, you are not really engaged in science at all.

7.1: Prelude to Presenting Your Research is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **11.1: Prelude to Presenting Your Research** has no license indicated.

7.2: Introduction to American Psychological Association (APA) Style

Learning Objectives

1. Define APA style and list several of its most important characteristics.
2. Identify three levels of APA style and give examples of each.
3. Identify multiple sources of information about APA style.

What Is APA Style?

APA Style is a set of guidelines for writing in psychology and related fields. You probably learned MLA (Modernal Language Association) Style in your English classes, and maybe other styles in your history, sociology, or biology courses. However, many of the social sciencies and medical field use APA Style, so it's great to learn for everyone. The APA Style guidelines are set down in the *Publication Manual of the American Psychological Association, 7th edition* (APA, 2020). The *Publication Manual* originated in 1929 as a short journal article that provided basic standards for preparing manuscripts to be submitted for publication (Bentley et al., 1929). It was later expanded and published as a book by the association and is now in its seventh edition. TThe 7th edition has a Professional style and a Student style. **We will be discussing the Professional style**. Many resources on the internet use the Student style, or even are still using the 6th edition, so don't trust them!

The primary purpose of APA style is to facilitate scientific communication by promoting clarity of expression and by standardizing the organization and content of research articles and book chapters. It is easier to write about research when you know what information to present, the order in which to present it, and even the style in which to present it. Likewise, it is easier to read about research when it is presented in familiar and expected ways.

APA style is best thought of as a "genre" of writing that is appropriate for presenting the results of psychological and social science research—especially in academic and professional contexts. There are formatting requirements, and writing expectations best described as technical writing (in comparison to creative writing or persuasive writing). This type of technical writing is not considered "good writing" in non-experimental fields. A literary analysis for an English class is not technical writing and would not use APA Style, even if it were based on psychoanalytic concepts; you would use MLA Style instead. And you would not write a newspaper article, even if it were about a new breakthrough in behavioral neuroscience, in APA style. You would write it in Associated Press (AP) style instead. At the same time, you would not write an empirical research report in MLA style, in AP style, or in the style of a romance novel, an email to a friend, or a shopping list. You would write it in APA style. Part of being a good writer in general is adopting a style that is appropriate to the writing task at hand, and for writing about psychological research, this is APA style.

The Levels of APA Style

Because APA style consists of a large number and variety of guidelines—the *Publication Manual* is nearly 300 pages long—it can be useful to think about it in terms of three basic levels. These levels can help you think about how you can use APA Style in your own writing:

- Organization: Order of sections of empirical reports
- High-Level Style: Writing and communication
- Low-Level Style: Mechanics and formatting

Organization

The first level is the overall organization of an article (which is covered in Chapter 2 "Manuscript Structure and Content" of the *Publication Manual*). As discussed in the chapter on literature reviews, empirical research reports, in particular, have several distinct sections that always appear in the same order:

- Title page
- Abstract
- Main body, which includes:
  - Introduction (unlabeled)

  - Methods (includes sub-sections)
  - Results
  - Discussion/Conclusion
- Reference page
- Appendices (optional)

High-Level Style

The second level of APA style can be referred to as high-level style (covered in Chapter 3 “Writing Clearly and Concisely” of the *Publication Manual*), which includes guidelines for the clear expression of ideas. There are two important themes here. One is that APA-style writing is formal rather than informal. It adopts a tone that is appropriate for communicating with professional colleagues—other researchers and practitioners—who share an interest in the topic. Beyond this shared interest, however, these colleagues are not necessarily similar to the writer or to each other. A graduate student in British Columbia might be writing an article that will be read by a young psychotherapist in Syria and a respected professor of psychology in Tokyo. Thus formal writing avoids slang, contractions, pop culture references, humor, and other elements that would be acceptable in talking with a friend or in writing informally.

The second theme of high-level APA style is that it is straightforward. This means that it communicates ideas as simply and clearly as possible, putting the focus on the ideas themselves and not on how they are communicated. Thus APA-style writing minimizes literary devices such as metaphor, imagery, irony, suspense, and so on. Again, humor is kept to a minimum. Sentences are short and direct. Technical terms must be used, but they are used to improve communication, not simply to make the writing sound more “scientific.” For example, if participants immersed their hands in a bucket of ice water, it is better just to write this than to write that they “were subjected to a pain-inducement apparatus.”

APA Style and the Values of Psychology

Robert Madigan and his colleagues have argued that APA style has a purpose that often goes unrecognized (Madigan et al., 1995). Specifically, it promotes psychologists’ scientific values and assumptions. From this perspective, many features of APA style that at first seem arbitrary actually make good sense. Following are several features of APA-style writing and the scientific values or assumptions they reflect.

APA style feature	Scientific value or assumption
There are very few direct quotations of other researchers.	The phenomena and theories of psychology are objective and do not depend on the specific words a particular researcher used to describe them.
Criticisms are directed at other researchers’ work but not at them personally.	The focus of scientific research is on drawing general conclusions about the world, not on the personalities of particular researchers.
There are many references and reference citations.	Scientific research is a large-scale collaboration among many researchers.
Empirical research reports are organized with specific sections in a fixed order.	There is an ideal approach to conducting empirical research in psychology (even if this ideal is not always achieved in actual research).
Researchers tend to “hedge” their conclusions, e.g., “The results *suggest* that…”	Scientific knowledge is tentative and always subject to revision based on new empirical results.

Another important element of high-level APA style is the avoidance of language that is biased against particular groups. This is not only to avoid offending people—why would you want to offend people who are interested in your work?—but also for the sake of scientific objectivity and accuracy. For example, the term *sexual orientation* should be used instead of *sexual preference* because people do not generally experience their orientation as a “preference,” nor is it as easily changeable as this term suggests (APA

Committee on Lesbian, Gay, and Bisexual Concerns Joint Task Force on Guidelines for Psychotherapy With Lesbian, Gay, and Bisexual Clients, 2000).

The general principles for avoiding biased language are fairly simple. First, be sensitive to labels by avoiding terms that are offensive or have negative connotations. This includes avoiding terms that identify people with a disorder or other problem they happen to have. Instead, refer to the individual, what the APA *Publication Manual* refers to as putting the "person first." For example, *people diagnosed with schizophrenia* is better than *schizophrenics*. Second, use more specific terms rather than more general ones. For example, *Chinese Americans* is better than *Asian Americans* if everyone in the group is, in fact, Chinese American. Third, avoid objectifying research participants. Instead, acknowledge their active contribution to the research. For example, "The *students completed* the questionnaire" is better than "The *subjects were administered* the questionnaire." Note that this principle also makes for clearer, more engaging writing. Table **7.2.1** shows several more examples that follow these general principles. For more information, the APA created the Equity, Diversity, and Inclusion: Inclusive Language Guide (APA, 2023), with an introductory page to identify additions in the second edition (Akbar, 2022) that links to the new guide.

Table **7.2.1**: Examples of Avoiding Biased Language

Instead of…	Use…
man, men	men and women, or people
firemen	firefighters
homosexuals, gays, bisexuals	lesbians, gay men, bisexual men, or bisexual women
minority	specific group label (e.g., African American)
neurotics	people scoring high in neuroticism
special children	children with learning disabilities

Low-Level Style

The third level of APA style can be referred to as low-level style (which is covered in Chapter 4 "The Mechanics of Style" through Chapter 7 "Reference Examples" of the *Publication Manual*). Low-level style includes all the specific guidelines pertaining to spelling, grammar, references and reference citations, numbers and statistics, figures and tables, and so on. Much of the rest of this chapter will focus on these formatting issues.

There are so many low-level guidelines that even experienced professionals need to consult the *Publication Manual* from time to time. Table **7.2.2** contains some of the most common types of APA style errors based on an analysis of manuscripts submitted to one professional journal over a 6-year period (Onwuegbuzie, et al., 2010). These errors were committed by professional researchers but are probably similar to those that students commit the most too.

Table **7.2.2**: Top 10 APA Style Errors

Error type	Example
1. Use of numbers	Failing to use numerals for 10 and above
2. Hyphenation	Failing to hyphenate compound adjectives that precede a noun (e.g., “role playing technique” should be “role-playing technique”)
3. Use of *et al.*	Failing to use it after a reference is cited for the first time
4. Headings	Not capitalizing headings correctly
5. Use of *since*	Using *since* to mean *because*
6. Tables and figures	Not formatting them in APA style; repeating information that is already given in the text
7. Use of commas	Failing to use a comma before *and* or *or* in a series of three or more elements
8. Use of abbreviations	Failing to spell out a term completely before introducing an abbreviation for it
9. Spacing	Not consistently double-spacing between lines
10. Use of “&” in references	Using *&* in the text or *and* in parentheses

This chapter is a resource for learning about APA Style, and how to write and format your reporting based on these guidelines. The last section of the chapter (Summary and APA Style Resources) provides additional resources from the APA and others to help!

References

Akbar, M. (2022). Three key things you should know about APA’s new inclusive language guidelines. *APA Style*. https://apastyle.apa.org/blog/inclus...age-guidelines

American Psychological Association. (2020). *Publication Manual of the American Psychological Association* (7th ed.). Washington, D.C.: American Psychological Association.

American Psychological Association. (2023). *Inclusive language guide* (2nd ed.). https://pages.apa.org/edi-guidelines/ or https://www.apa.org/about/apa/equity-diversity-inclusion/language-guidelines.pdf

American Psychological Association, Committee on Lesbian, Gay, and Bisexual Concerns Joint Task Force on Guidelines for Psychotherapy With Lesbian, Gay, and Bisexual Clients. (2000). *Guidelines for psychotherapy with lesbian, gay, and bisexual clients*. Retrieved from https://www.apa.org/pi/lgbt/resources/guidelines

Bentley, M., Peerenboom, C. A., Hodge, F. W., Passano, E. B., Warren, H. C., & Washburn, M. F. (1929). Instructions in regard to preparation of manuscript. *Psychological Bulletin, 26*, 57–63.

Madigan, R., Johnson, S., & Linton, P. (1995). The language of psychology: APA style as epistemology. *American Psychologist, 50*, 428–436.

Onwuegbuzie, A. J., Combs, J. P., Slate, J. R., & Frels, R. K. (2010). Editorial: Evidence-based guidelines for avoiding the most common APA errors in journal article submissions. *Research in the Schools, 16*, ix–xxxvi.

This page titled 7.2: Introduction to American Psychological Association (APA) Style is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **11.2: American Psychological Association (APA) Style** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

7.3: General Paper Format

APA papers have different sections that have different slightly different formatting styles, however, the **entire** paper should:

The ENTIRE paper should:

- Be double-spaced (except Tables and Figures)
- Be 12 pt. Times New Roman font (or similar)
- Have 1" margins on all sides
- Not have any extra spacing before or after each paragraph.
    - Check **Paragraph** settings for this; the default in Word is to add 10 points of space after each paragraph (regardless of line spacing).
- Have a page number in the upper right-hand side of every page that is the same font as the rest of the paper.
- Have the paper's title, all capitalize, as a page header in the upper left-hand of every page. When all papers were submitted by mailing hard-copies to journal publishers, the running head was used to keep the papers together and organized, but anonymous. Dr. MO has also used this when she has literally dropped a stack of printed-but-not-yet-stapled student papers.

There should be a page break (starting on a new page) *between* each of these sections. There should not be a page break *within* any of these sections. Even if you have part of a sentence dangling on a new page. *Insert a Page Break* to start a new page, don't just hit Enter a bunch of times. Doing that messes up other formatting.

In-Text Citations

If you cite any sources, there is a particular way to format them in your sentences. Surprised?

In-text citations help readers locate the source of your information in the References section of the paper. Whenever you use a source, provide the author's last name(s), *in the order that they appear on the article*, and the year of publication. You can use the authors' names in the sentence, or only include their names in the reference at the end of the sentence.

Example with One Author

- Using their name in the sentence: Wilson (2014) developed mechanical wings.
- Not using their name in the sentence: Mechanical wings can be used as shielding or for flight (Wilson, 2014).

Example with Two Authors

- Using their names in the sentence: According to Prince and Trevor (2021), you must sometimes sacrifice to live your values.
- Not using their name in the sentence: There is still controversy over the cause of World War I, also known as the Great War (Prince & Trevor, 2017).

Notice that everything cited is in the *past tense*. It was written before today, so it is in the past.

> **Note**
>
> What is different with two authors between using the authors' names in the sentence, compared to at the end of the sentence?

Example with Three or More Authors

When citing a work with three or more authors, you can use the first author's last name followed by "et al." in the sentence or in parentheses:

- Using their names in the sentence: Kirk et al. (1968) found many similarities between their different cultures.
- Not using their name in the sentence: Exploration shows that there is infinite diversity in infinite combinations (Kirk et al., 1968).

Students often want to cite once at the end of a whole paragraph. Resist this urge. In APA Style, each sentence with information from an outside source must be cited. Citing at the end of the sentence makes it difficult to recognize which parts of the paragraph are the author's thoughts, and which portions are the ideas or findings for another author.

This page titled 7.3: General Paper Format is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

7.4: Formatting by Section

Learning Objectives

1. Identify how formatting differs between the major sections of an APA Style paper.
2. Differentiate between a heading and a header.
3. Identify when to use each heading level.

APA Style Formatting by Paper Section

APA papers generally include the following sections:

1. Cover page (sometimes called a title page)
2. Abstract
3. Main body
4. References
5. Appendices

If you forgot what each of these sections cover, go back to the section How to Read a Research Article. This section will focus on the specific formatting of each section. The formatting discussed in the last section that should be throughout the paper will *not* be repeated here, so make sure to format your whole paper that way first, then move on to making sure each section is correctly formatted. For a sample paper, check out sample papers from the APA or OWL Purdue.

First Section: Title Page

In the upper half of the page, centered:

- The title of your paper, in **bold**.
- Your name
- Affiliation (your college/university)
- Author note (optional)

Remember, this text will focus on the Professional Style described in the 7th edition of the APA Style manual. The Student Style will have slightly different information on the title page, and the header on the first page will differ.

Insert a Page Break after the affiliation so that the first page of the main body starts on the next page. [Don't just hit Enter a bunch of times. Doing that messes up other formatting.]

Second Section: Abstract

An abstract is a 150- to 250- word summary of your paper. The abstract page should include:

- Page header and page number (2, because this will always be on the second page)
- The abstract title (which should be the word **Abstract**), centered and **bold**, at the top of the page.
- The abstract paragraph is **not indented**.

Third Section: Main Body

This is the section of the paper where most of your writing will be; this is where you describe what was measured, on whom, and what was found.

The first page of the main body includes the **title in the header *and* at the top of the page (as a heading).** This title is in the style of a Level 1 heading (see below). Every page of your main body should include:

- Page header and Page number
- Indented paragraphs (**with no extra space between the paragraphs**).

There should *not* be a page break *within* the main body; Just let the paragraphs break where they naturally do.

Headings

A heading is a title for different sections of your paper. It is not a header, which is the title in the top margin of each page (that's all capitalized). APA uses a system of five heading levels, as shown in the Note below. Start with the Level 1 style for your titles and headings, then move down for lower level sections.

Note

Level 1: Centered, Bold, Title Case (first letter of important words is capitalized)

Level 2: Flush Left, Bold, Title Case

Level 3: Flush Left, Bold & Italicized, Title Case

It is exceedingly unlikely that you will need a Level 4 or Level 5 heading, but here they are:

Level 4: Indented, Bold, Title Case, Ending with a Period.

Level 5: Indented, Bold and Italicized, Title Case, Ending with a Period.

Fourth Section: Reference Page

Your reference page is where each of you cited sources is listed. You need a reference page when you cite someone else's work. Like every other page, the reference page should include:

- Running head (only the paper's title, all capitalized)
- The page number
- Double-spaced, with no extra space between paragraphs

It should also include a title (**References**) at the top of the page in bold (Level 1 heading). APA uses References**, *not* works cited *or* bibliography**.

Many of the pages in this textbook provide examples of references of articles, books, or webpages. Some of the pages in this text show examples of footnotes, but APA Style uses references. The references (not footnotes) are correctly formatted, except that the they do not have hanging indents. Hanging indents of each reference is the opposite of the indentation paragraphs in the main body. Each reference should start all of the way to the left, and each subsequent line should be indented. This is called a *hanging indent*, and Word will do this automatically for you through the paragraph settings. OWL Purdue has several pages dedicated to formatting references of different types of sources. You will most likely be using the research articles, so you would look at the Reference List: Articles and Periodicals page.

While your authors for each reference should be kept in the order that they appear on the source, the order of the references themselves (each reference "paragraph") should be in alphabetical order by the first author's last name. If you are referencing the same author or set of authors (in the same order) but from different sources, you would order the reference "paragraph" by the year.

Fifth Section: Appendix

Depending on the journal publisher, an appendix could be used to present tables, figures, lists of stimulus words, questionnaire items, detailed descriptions of special equipment or unusual statistical analyses, or picture or video screenshots. Basically, nn appendix is where material that would interrupt the flow of the research report if it were presented within any of the major sections, but is still important for understanding the study. Like the abstract page and the references, each appendix should include

- Page header and page number (like all pages)
- The title, centered and **bold**, at the top of the page.

If you only have one appendixe, then the title should be the word **Appendix**. For example, researchers who use visual prompts (like pictures of faces) would put the pictures in one appendix. If you have multiple appendices, then each appendix is identified by a letter; **Appendix A** would inclue the information discussed first in the paper, **Appendix B** would include information mentioned next in the body of the paper, etc. For example, a research might mention the outcome variable (DV) of a survey in the Measures sub-section of the Method section, so that could be Appendix A. Then, the visual prompts could be mentioned in the Procedure sub-section of the Method section, so those could be in Appendix B.

In the 7th edition of the APA Style manual, tables and figures can be in appendixes or in the paper around where each is mentioned. The formatting for these will be discussed next.

This page titled 7.4: Formatting by Section is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

7.5: Presenting Results

Learning Objectives

1. Recognize the importance of tables and graphs in presenting information.
2. Identify how to format tables, graphs, and their titles in APA Style.

Tables are where you put rows or columns of numbers, like a frequency table or your calculation results. A figure is a chart/graph or other picture. Often, these are used to present statistical information, but they could also be used to present lists (like for words that participants were asked to memorize) or pictures that were presented to participants.

Presenting Descriptive Statistics in Writing

APA Style includes several rules for presenting numerical results in the text (see 4.31–4.34 in the APA *Publication Manual*) . These include using words only for numbers less than 10 that do not represent precise statistical results and using numerals for numbers 10 and higher. However, statistical results are always presented in the form of numerals rather than words and are usually rounded to two decimal places (e.g., “2.00” rather than “two” or “2”). When you have a small number of results to report, it is often most efficient to write them out.

Examples of Statistical Results in the Narrative

- The mean age of the participants was 22.43 years with a standard deviation of 2.34.
- Among the participants with low self-esteem, those in a negative mood expressed stronger intentions to have unprotected sex (M = 4.05, SD = 2.32) than those in a positive mood (M = 2.15, SD = 2.27).
- The treatment group had a mean of 23.40 (SD = 9.33), while the control group had a mean of 20.87 (SD = 8.45).
- The test-retest correlation was .96.
- There was a moderate negative correlation between the alphabetical position of respondents’ last names and their response time (r = −.27).

Notice that when presented in the narrative, the terms *mean* and *standard deviation* are written out, but when presented parenthetically, the symbols M and SD are used instead. Notice also that it is especially important to use parallel construction to express similar or comparable results in similar ways. The third example is *much* better than the following nonparallel alternative: The treatment group had a mean of 23.40 (SD = 9.33), while 20.87 was the mean of the control group, which had a standard deviation of 8.45.

An additional consideration is needed when writing about comparisons between social groups. As noted in Russel et al. (2022), researchers often write about differences between social identity groups as if the difference is caused by the group membership itself. As an example from Russell et al. (2022), educational research on the achievement differences between White and Black elementary students seem to imply that it's the fact of being Black that affects achievement while ignoring policies and institutional practices that contribute to educational outcomes. For example, public school funding is based based on property taxes, but red-lining practices through the early 1900's have excluded from upper and middle class neighbors, resulting in many African American families being forced to live in poorer neighborhoods that do not have the tax base to fully support their community schools. Another example is the use of testing to track students into remedial or advanced coursework, even while we know that academic tests are culturally biased (Russell et al., 2022). To counter this "deficit narrative," Russell et al. (2022) urged researchers to frame any differences between social identity groups to be caused by institutional influences (racism, biased policies), rather than imply that the difference is caused by any inadequacies of groups disadvantaged by institutional policies, or imply the superiority of groups already advantaged by the past and current systems. In terms fo race, "one should consider race a proxy for the impacts of racism," (Russell et al., 2022, 5).

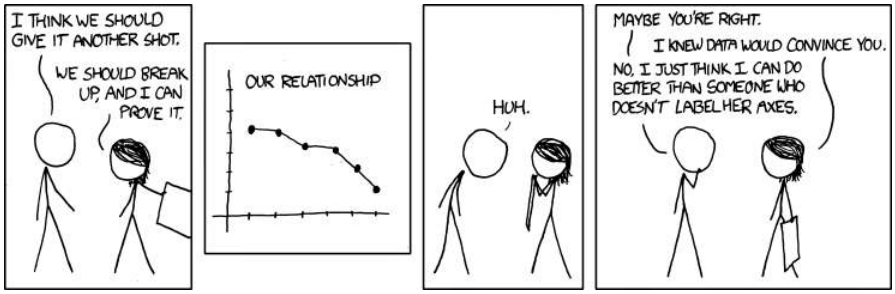
Presenting Descriptive Statistics in Figures

When you have a large number of results to report, you can often do it more clearly and efficiently with a graphical depiction of the data, such as frequency line graphs, bar graphs, or scatterplots. In an APA style research report, these graphs are presented as

figures. Figures also include pictures. When you prepare figures for an APA Style research report, there are some general guidelines that you should keep in mind. First, the figure should always *add* important information rather than repeat information that already appears in the text or in a table (if a figure presents information more clearly or efficiently, then you should keep the figure and eliminate the text or table.) Second, figures should be as simple as possible. For example, the *Publication Manual* discourages the use of color in manuscripts unless it is absolutely necessary (although color can still be an effective element in posters, slide show presentations, or textbooks.) Third, figures should be interpretable on their own. A reader should be able to understand the basic result based only on the figure and its caption and should not have to refer to the text for an explanation.

There are also several more technical guidelines for presentation of figures that include the following (see the APA Publication Manual section 5.20 through 5.30):

- Layout of graphs
  - In general, scatterplots, bar graphs, and line graphs should be slightly wider than they are tall.
  - The independent variable should be plotted on the *x*-axis and the dependent variable on the *y*-axis.
  - Values should increase from left to right on the *x*-axis and from bottom to top on the *y*-axis.
  - The *x*-axis and *y*-axis should begin with the value zero.
- Axis Labels and Legends
  - Axis labels should be clear and concise and include the units of measurement if they do not appear in the caption.
  - Axis labels should be parallel to the axis.
  - Legends should appear within the figure.
  - Text should be in the same simple font throughout and no smaller than 8 point and no larger than 14 point.
- Captions
  - Captions are titled with the word “Figure”, followed by the figure number in the order in which it appears in the text, and terminated with a period. This title is italicized.
  - After the title is a brief description of the figure terminated with a period (e.g., “Reaction times of the control versus experimental group.”)
  - Following the description, include any information needed to interpret the figure, such as any abbreviations, units of measurement (if not in the axis label), units of error bars, etc.



“Convincing” retrieved from http://imgs.xkcd.com/comics/convincing.png (CC-BY-NC 2.5)

Expressing Descriptive Statistics in Tables

Like graphs, tables can be used to present large amounts of information clearly and efficiently. The same general principles apply to tables as apply to graphs. They should add important information to the presentation of your results, be as simple as possible, and be interpretable on their own. Again, we focus here on tables for an APA Style manuscript.

The most common use of tables is to present several means and standard deviations—usually for complex research designs with multiple independent and dependent variables. Table **7.5.1**, for example, shows the results of a hypothetical study similar to the one by MacDonald and Martineau (2002). MacDonald and Martineau (2002) categorized participants as having low or high self-esteem, put them into a negative or positive mood, and measured their intentions to have unprotected sex. MacDonald and Martineau (2002) also measured participants’ attitudes toward unprotected sex. Notice that the table includes horizontal lines spanning the entire table at the top and bottom, and just beneath the column headings. Furthermore, every column has a heading—including the leftmost column—and there are additional headings that span two or more columns that help to organize the

information and present it more efficiently. Finally, notice that APA Style tables are numbered consecutively starting at 1 and given a brief but clear and descriptive title.

	Negative mood		Positive mood	
Self-Esteem	*M*	*SD*	*M*	*SD*
		Intentions		
High	2.46	1.97	2.45	2.00
Low	4.05	2.32	2.15	2.27
		Attitudes		
High	1.65	2.23	1.82	2.32
Low	1.95	2.01	1.23	1.75

Table 7.5.1: Sample APA Style Table Presenting Means and Standard Deviations. Means and Standard Deviations of Intentions to Have Unprotected Sex and Attitudes Toward Unprotected Sex as a Function of Both Mood and Self-Esteem.

Another common use of tables is to present correlations—usually measured by Pearson's *r*—among several variables. This kind of table is called a correlation matrix. Table 7.5.2 is a correlation matrix based on a study by McCabe et al. (2010). McCabe et al. (2010) were interested in the relationships between working memory and several other variables. We can see from the table that the correlation between working memory and executive function, for example, was an extremely strong .96, that the correlation between working memory and vocabulary was a medium 0.27, and that all the measures except vocabulary tend to decline with age. Notice here that only half the table is filled in because the other half would have identical values. For example, the Pearson's *r* value in the upper right corner (working memory and age) would be the same as the one in the lower left corner (age and working memory). The correlation of a variable with itself is always 1.00, so these values are replaced by dashes to make the table easier to read.

Measure	1	2	3	4	5
1. Working memory	—				
2. Executive function	.96	—			
3. Processing speed	.78	.78	—		
4. Vocabulary	.27	.45	.08	—	
5. Episodic memory	.73	.75	.52	.38	—
6. Age	–.59	–.56	–.82	.22	–.41

Table 7.5.2: Sample APA-Style Table (Correlation Matrix) Based on Research by McCabe and Colleagues. Correlations Between Five Cognitive Variables and Age.

As with graphs, precise statistical results that appear in a table do not need to be repeated in the text. Instead, the writer can note major trends and alert the reader to details (e.g., specific correlations) that are of particular interest.

Formatting Tables and Figures

That introduction gave you some indication of the formatting of tables and figures. And as previously mentioned, the 7th Edition allows you to put the tables and figures in the main body of the paper, but you can also put them on their own pages at the end of your paper. Here are some more specific information on now to format tables and figures:

- Tables and Figures can be single-spaced.
- Tables and Figures must be numbered, in **bold** (Example: **Figure 1** or **Table 1**). The table or figure number should be on a line above the title.
- Tables and Figures must have a title (*in italics*) on the line below the number.
- In a Table, the column titles should be underlined.
- There should be a line across the whole bottom edge of each table.

If you are putting a table or figure in an appendix, then each set of tables or figures should be started on a new page [Insert a Page Break, don't hit enter a bunch of times].

Example Table Formatting

Table 3

Measures of Central Tendency and Standard Deviation

Table **7.5.3**- Example of APA Style Table Formatting

Groups	N	Mean	Median	Mode	SD
Group 1	34	9	2	6.50	2.51
Group 2	550	246.99	45.5	22.22	5.54
Note: "SD" is "standard deviation.					

Table **7.5.3**: Example of number, title, and table formatted in APA Style.. (CC-BY-NC-SA; Michelle Oja)

Example Chart Formatting

Figure 1

Frequency of Final Exam Scores

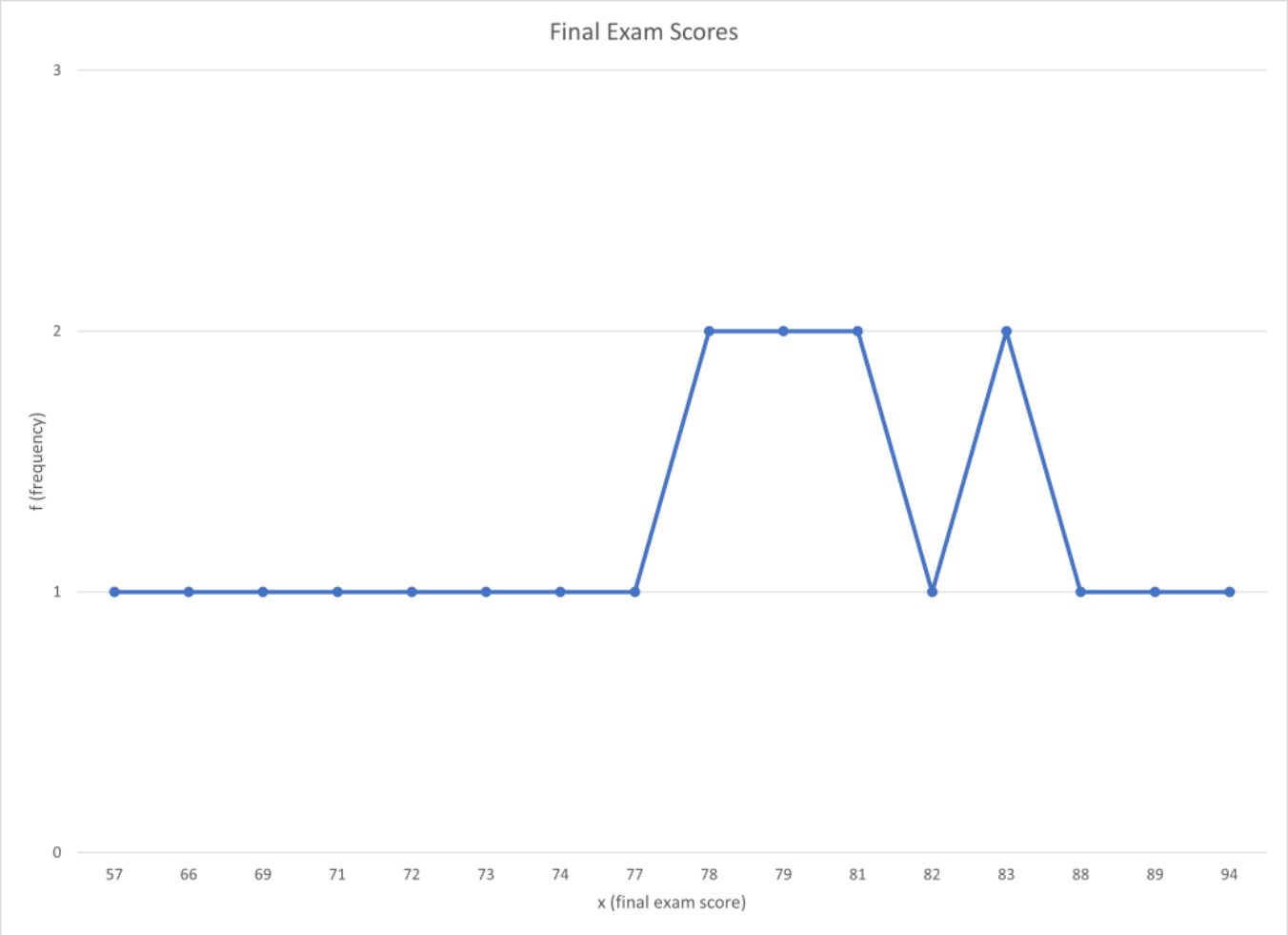


Figure **7.5.1**: Example of number, title, and chart formatted in APA Style. (CC-BY-NC-SA; created by Michelle Oja with data from OpenIntro.org)

References

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.

McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning. *Neuropsychology, 24*(2), 222–243. doi:10.1037/a0017619

Russell, M., Oddleifson, C., Russell Kish, M., & Kaplan, L. (2022). Countering deficit narratives in quantitative educational research. Practical Assessment, Research, and Evaluation, 27(14), 1-18. DOI: https://doi.org/10.7275/k44e-sp84 Available at: https://openpublishing.library.umass...ticle/id/1328/

This page titled 7.5: Presenting Results is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **12.3: Expressing Your Results** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

7.5.1: Tables and Figures

Tables are where you put rows or columns of numbers, like a frequency table or your calculation results. A figure is a chart/graph or other picture. The 7th Edition allows you to put the tables and figures in the main body of the paper, but you can also put them on their own pages at the end of your paper.

- Tables and Figures can be single-spaced.
- Tables and Figures must be numbered, in **bold** (Example: **Figure 1** or **Table 1**). The table or figure number should be on a line above the title.
- Tables and Figures must have a title (*in italics*) on the line below the number.
- In a Table, the column titles should be underlined.
- There should be a line across the whole bottom edge of each table.

If you are putting a table or figure in an appendix, then each set of tables or figures should be started on a new page [Insert a Page Break, don't hit enter a bunch of times].

Example Table Formatting

Table 1

Measures of Central Tendency and Standard Deviation

Table **7.5.1.1**- Example of APA Style Table Formatting

Groups	N	Mean	Median	Mode	SD
Group 1	34	9	2	6.50	2.51
Group 2	550	246.99	45.5	22.22	5.54
Note: "SD" is "standard deviation.					

Table **7.5.1.1**: Example of number, title, and table formatted in APA Style.. (CC-BY-NC-SA; Michelle Oja)

Example Chart Formatting

Figure 1

Frequency of Final Exam Scores

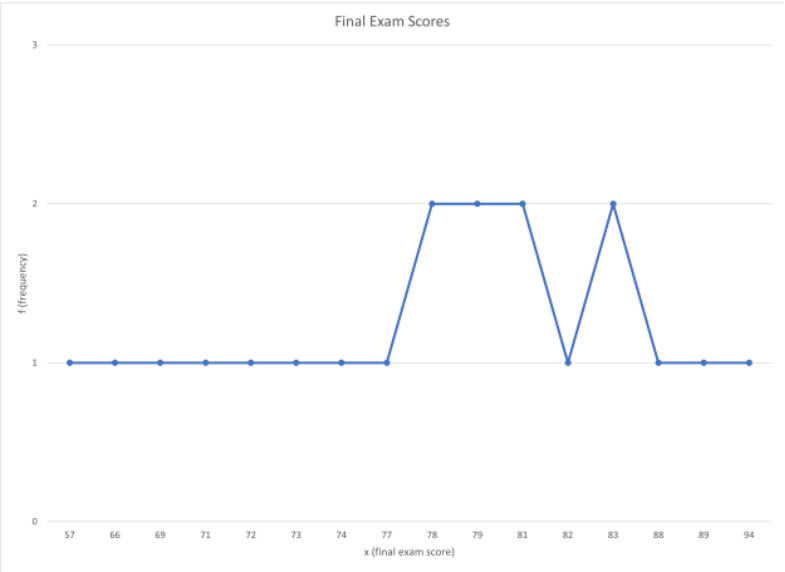


Figure **7.5.1.1**: Example of number, title, and chart formatted in APA Style. (CC-BY-NC-SA; created by Michelle Oja with data from OpenIntro.org)

A summary of formatting in APA Style is up next!

This page titled 7.5.1: Tables and Figures is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

7.6: Research Proposals

Learning Objectives

1. Identify how a research proposal differs from a research report.
2. Describe how sections of a research proposal differ from similar sections of a research manuscript.

A research proposal is often submitted to a course professor or a research advisor as the last step before you start actually collecting data. The proposal shows that you've read previous research, that the previous research or a theory has led you to identify needed information or clarification, a description of how you think it would be best to collect and analyze the data that you collect, and a conclusion about the impact and application of your expected findings. This is packaged in the same format as a manuscript or APA Style research report:

- Title page
- Abstract
- Main body, which includes:
  - Introduction
  - Methods, which includes sub-sections:
    - Participants
    - Materials
    - Design
    - Procedure
  - Results
  - Discussion/Conclusion
- Reference page
- Appendices (optional)

In addition to those sections, research proposals frequently include a research timeline. We discuss each of these eight sections as well as producing a research timeline below.

These sections are generally formatted the same, other than the verb tense. In manuscripts, everything is written in past tense. Others have already conducted their research (in your literature review), and you have already collected and analyzed your data. In research proposals, the literature review is still in past tense (other have already completed their research), but your Method, Result, and Conclusion sections are in future tense because you have not collected or analyzed your data yet.

Introduction Section

Just like a research report, the research proposal should begin with a hook to entice your readers. This can be done in many ways, just remember that the key objective is to convince your reader that the issue is problematic and is worth investigating. After the problem and rationale are introduced, the next step is to frame the problem within the academic discourse through summarizing others' prior research. Again, this is the same as in any other type of research report. By foregrounding the previous findings, you will then set the stage for your own methodology. Part of outlining the scholarly discussion should also focus on clarifying the boundaries of your topic. While making the significance concrete, try to hone in on select variables that your research should include.

Tips for A Research Proposal's Literature Review

- ***Summarize***: The literature in your literature review is not going to be exhaustive but it should demonstrate that you have a good grasp on key debates and trends in the field
- ***Quality not quantity***: Despite the fact that this is non-exhaustive, there is no magic number of sources that you need. Do not think in terms of how many sources are sufficient. Think about presenting a decent representation of key themes in the literature.

- ***Highlight theory and methodology*** of your sources (if they are significant). Doing so could help justify your theoretical and methodological decisions, whether you are departing from previous approaches or whether you are adopting them.
- ***Synthesize*** your results. Do not simply state "According to Robinson (2021)….According to Wilson (2021)… etc". Instead, find common grounds between sources and summarize the point e.g., "Researchers argue that we should not list our literature (Bartolic, 2021; Robinson, 2020; Wilson 2021).
- ***Justify*** methodological choice
- ***Assess and Evaluate:*** After assessing the literature in your field, you should be able to answer the following questions: Why should we study (further) this research topic/problem?
- ***Contribution***: At the end of the literature, you should be able to determine contributions will my study make to the existing literature?

As you briefly discuss the key literature concerning your topic of interest, it is important that you allude to gaps. Gaps are ambiguities, faults, and missing aspects of previous studies. Think about questions that you have which are not answered by existing literature. Specifically, think about how the literature might insufficiently address the following, and locate your research as filling those gaps (UNE, 2021):

- ***Population or sample***: size, type, location, demography etc. [Are there specific populations that are understudied e.g., Indigenous people, transgender youth, the elderly, etc.]
- ***Research methods***: qualitative, quantitative, or mixed [Has the research in the area been limited to just a few methods e.g., all surveys? How is yours different?]
- ***Variables or conditions*** [Are you examining a new or different set of variables than those previously studied? Are the conditions under which your study is being conducted unique e.g., under pandemic conditions]
- **Theory** [Are you employing a theory in a new way?]

Identifying a Gap

To indicate the usefulness and originality of your research, you should be conscious of how your research is both unique from previous studies in the field and how its findings will be useful. When you write your research report, you will expound on these gaps some more. However, in the body of your proposal, it is important that you explicitly highlight the insufficiency of existing literature (i.e. gaps). Below are some phrases that you can use to indicate gaps:

- …has not been clarified, studied, reported, or elucidated
- further research is required or needed
- …is not well reported
- key question(s) remains unanswered
- it is important to address …
- …poorly understood or known
- Few studies have (UNE, 2021)

The gaps and literature you outline should set the context for your research questions, and lead to your research hypothesis. After indicating what your research intends to study, formulate this gap into a research hypothesis which make investigating the gap tangible.

Method Section

The Method section of a research proposal is the place in which you explain, in specific details, how you plan to collect data to test your research hypothesis. By the time you begin explaining your methodology to test your research hypothesis, you would have already introduced your topic and its significance, and have provided a brief account of its scholarly history (literature review) and the gaps you will be filling. The methods section allows you to discuss *how* you intend to fulfill said gap. This section is in future tense ("Data will be collected..."), but the sub-sections of the Method section are still relevant. You would propose who your target sample could be, and how many participants you would like to collect data from. You would detail the materials that you plan to use, including the specific questions or prompts or other measures. You would name the type of design that your study will be, and describe the procedure (as you expect the participants to experience it). Your descriptions must be specific enough so that other

researchers can repeat them. It is important in this section that you also justify why you believe this specific methodology is the most effective for answering the research question. This does not need to be extensive, but you should at least briefly note why you think, for instance, qualitative, quantitative, or mixed methods (and your specific proposed approach) are appropriate for answering your research question.

Results Section

The Results section maybe the most different in a research proposal compared to a research report. You haven't collected data yet, so how can you have results? Instead of analyzing your data and interpreting your results, this section in a research proposals discusses how the data will be analyzed. For quantitative research, this should include the specific statistical analyses that are most appropriate based on the IV and DV you explained in the Method section. You would refer back to your research hypothesis, and may include the operational definitions of your IV and DV so that it is clear that the statistical analyses that you are proposing fit with the data that you will have. For qualitative research, this section could describe the qualitative analysis approach that you plan to take. This section also does not have to be extensive. Examples of analysis methods for both types of research are show in Table **7.6.1**. This demonstrates that you have given thought to what you intend to do with the data once you have collected it.

Table **7.6.1**- Example Analysis Methods

Qualitative Analysis Methods	Quantitative Analysis Methods
Qualitative Content Analysis	
Discourse Analysis	Descriptive Statistics
Thematic Analysis	Mean, median, standard deviation, skewness
Grounded Theory (Inductive)	Inferential Statistics
Narrative Analysis	T-tests, ANOVA, Correlation, regression, chi-square
Interpretive Phenomenological Analysis	

Table **7.6.1**: Example Analysis Methods. (CC-BY-NC-SA; Robinson & Wilson via 2.4: The Body of the Research Proposal)

This demonstrates that you have given thought to what you intend to do with the data once you have collected it.

Conclusion Section

After you have outlined the literature, the gaps in the literature, how you intend to investigate that gap, and how you intend to analyze what you have found, it is important to again reiterate the significance of your study. You still do not have data, but you can still summarize what you've read and what you expect to find, note any limitations of your proposed study, and discuss implications of the expected findings. Allude to what your study *could find* and what this *would mean*. This requires returning to the significant territory that began your proposal and linking it to how your study could help to explain/change this understanding or circumstance. Report on the possible beneficial outcomes of your study. For instance, say you study the impact of welfare checks on homelessness. Then you could respond to the following question: How could my findings improve our responses to homelessness? How could it make welfare policies more effective? Remember you must explain the usefulness or benefits of the study to both the outside world and the research community.

In addition to noting your strengths, also reflect on the weaknesses. All research has limitations but you need to demonstrate that you have taken steps to mitigate those that can be mitigated and that the research is valuable despite the weaknesses. Be straightforward about the things your study will not be able to find, and the potential obstacles that will be presented to you in conducting your study (in research that is conducted with a population, be sure to note harms/benefits that might come to them). It is tempting to ask why you would propose a study with limitations. The answer is that the real world is messy, and sometimes you can't get the information that you want from the people you want it from. Sometimes, you cannot randomly assign people into the

groups that you want to investigate. Sometimes the most scientifically effective study is unethical. It is important for your professor or advisor to be aware of your understanding the practical and ethical barriers to conducting research on your topic. Try to address these obstacles to the best of your ability while continuing to argue for the value of your study despite inevitable tradeoffs. However, do not finish with a long list of inadequacies. End with a crescendo –with the impression that despite the trials and limitations of research, you are prepared for the challenge and the challenge is well worth overcoming. This means reiterating the significance, potential uses, and implications of the findings.

References

UNE (2021). Gaps in the literature. UNE Library services. https://library.une.edu/wp-content/uploads/2020/07/Gaps-in-the-Literature.pdf

This page titled 7.6: Research Proposals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Oral Robinson and Alexander Wilson via source content that was edited to the style and standards of the LibreTexts platform.

- **2.4: The Body of the Research Proposal** by Oral Robinson and Alexander Wilson is licensed CC BY-NC 4.0. Original source: https://pressbooks.bccampus.ca/undergradresearch.

7.7: Summary and APA Style Resources

This summary page will include more than key takeaways and exercises. This is also where you'll find additional resources on APA Style!

Key Takeaways

- Many fields in the social sciences and medicine utilize APA Style, a formatting and writing style to encourage technical writing that follows a typical structure.
- APA Style has very specific requirements for formatting, but also is a guide for writing style.
- There are specific formats for citing sources in the text, and for formatting the references of these sources at the and of the manuscript. There are also specific formatting requirements for each section of a research report.
- Research proposals are designed to present researchers' ideas on how they would design and analyze their next research study. While the formatting is similar to a typical APA Style research report, the content focuses on justifying the research study rather than describing a study that the researcher has already completed.

Resources

Dr. MO encourages you to purchase the APA Style manual if you are a psychology or sociology major. However, your school's library has the *Publication Manual of the American Psychological Association, 7th ed.* Note that it is typically considered a reference (like a dictionary or encyclopedia), so you probably can't take it out of the library. Your library probably has a webpage devoted to APA Style, as well. Your library probably also has a Library Guide or additional resources for APA Style. The APA has developed a website (https://apastyle.apa.org/) to help you find how to format in APA Style (APA, 2025a). This website is a good resource for formatting, but you need the actual book to learn about the high-level style of writing and communication (discussed in the Introduction to APA Style section of this textbook). When using these, make sure to use the resources for the Professional type of paper, not the Student version (unless your instructor prefers the Student version). The APA also provides a variety of writing guides; the box below shows some of the newer guides and introductions. As mentioned in the section on How to Read a Research Article, there is also a Journal Article Reporting Standards (JARS) webpage (APA, 2025b). Specific guides are listed in the box below. For more information on qualitative research, review The Qualitative Approach section in the chapter on Analyzing.

Writing Guides

Primary Guide: *Publication Manual of the American Psychological Association* (7th ed.) (2020); more information can be found at https://apastyle.apa.org/

These writing guides developed by the APA provide guidelines for specific types of research and topics:

- JARS-Quant (2018) standards (includes additional guides for many different types of designs and methodologies): https://apastyle.apa.org/jars/quantitative
- JARS-Qual (2018) standards: https://apastyle.apa.org/jars/qualitative
- JARS-Mixed (2018) standards: https://apastyle.apa.org/jars/mixed-methods
- Three key things you should know about APA's new inclusive language guidelines (Akbar, 2022) *APA Style*. https://apastyle.apa.org/blog/inclus...age-guidelines (which is supposed to link to this document: Equity, Diversity, and Inclusion: Inclusive Language Guidelines (APA, 2023): https://pages.apa.org/edi-guidelines/)
- JARS- Race, Ethnicity, and Culture (2023) (includes video towards the bottom of the page): https://apastyle.apa.org/jars/race-ethnicity-culture
- Sexual Orientation and Gender Diversity Research Manuscript Writing Guide (Veldhuis et al., 2024): https://psycnet.apa.org/fulltext/2024-79040-001.pdf

Another great resource for formatting is Purdue University's Online Writing Lab (website address: http://owl.english.purdue.edu) has an immense collection of webpages dedicated to APA Style, including a whole set of pages on references. There's a PDF of a

sample Professional Style paper with explanations of formatting. There's also a sample paper for students, but we are using the Professional Style version.

This APA Checklist is a PDF developed by Dr. MO to go through before you turn in any project that is supposed to be formatted in APA Style.

Many people learn best by watching (and re-watching) videos, so here's a good video in Figure **7.7.1** showing how to format a Word document into an APA Style paper. Although sometimes the features are in different places in a Google doc or on a Mac, all word processing software will allow you to make these modifications.

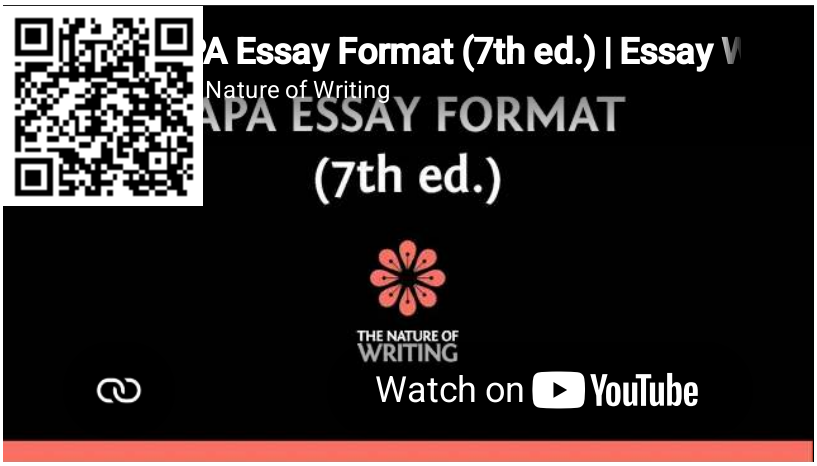


Figure **7.7.1**: APA Essay Format (7th Edition). (The Nature of Writing via YouTube)

DO NOT USE THE APA STYLE TEMPLATE PROVIDED BY WORD. I know that this seems like an easy way to format you paper, but the template is really difficult to modify. For example, if you don't need, say, an abstract or a reference page for your paper, it is nearly impossible to delete those from your paper when created from the template without messing up the formatting for the rest of your paper.

Here is a sample paper as a Word document (.docx). Save this to your own drive or computer so that you can save and modify it. This is an example paper to show APA Style formatting, and the basic information in a descriptive statistics paper. Careful reading shows that the topic and numbers change throughout the paper because it is a combination of several different students' papers. This is just a guide; do NOT copy the words, just the format.

What's Next

If you can believe it, you are now done learning the foundational concepts of research methods and you should have more familiarity in reading and writing research reports. The next unit focuses on different kinds of research designs.

Exercises

- Practice: Using a research article, find all of the standard APA Style sections detailed in this chapter
- Practice: Using a research article, note how that article's published formatting differs from APA Style formatting for a standard research report.
- Practice: Correctly cite a research article by using the authors' names in the sentence, then write another sentence with a citation of a second research article at the end of the sentence.
- Practice: Correctly format a reference for each of the articles for each of the citations that you just completed.
- Practice: Draft a paper in APA Style, then use the APA Checklist to check-off what you did correctly and what you missed.
- Discussion: Read Russel et al. (2022), then see if you can:
  - Summarize the three sections of the introduction section in 3-5 sentences for each section for a classmate who has not read the article,
  - Summarize the method section in 4-8 sentences for a classmate who has not read the article,
  - Summarize the results section (called Findings) in 3-5 sentences for a classmate who has not read the article, then

- Describe what you learned from this article.
- Discussion: Work with your classmates to find what resources your library or professor has provided related to APA Style.

References

Akbar, M. (2022). Three key things you should know about APA's new inclusive language guidelines. *APA Style*. https://apastyle.apa.org/blog/inclus...age-guidelines

American Psychological Association. (2020). *Publication Manual of the American Psychological Association* (7th ed.). Washington, D.C.: American Psychological Association.

American Psychological Association. (2023). *Inclusive language guide* (2nd ed.). https://pages.apa.org/edi-guidelines/ or https://www.apa.org/about/apa/equity-diversity-inclusion/language-guidelines.pdf

American Psychological Association. (2025a). *APA Style*. https://apastyle.apa.org/

American Psychological Association (2025b). Journal article reporting standards (JARS). https://apastyle.apa.org/jars

Nature of Writing (2019, November 7). *APA Essay Format (7th ed)* [Video]. Youtube. https://www.youtube.com/watch?v=Mrh5OC3T6dc

Veldhuis, C. B., Cascalheira, C. J., Delucia, K., Budge, S. L., Matsuno, E., Huynh, K., Pickett, J. A., Balsom, K. F., Velez, B. L., & Paz Galupo, M. (2024). Sexual orientation and gender diversity research manuscript writing guide. *Psychology of Sexual Orientation and Gender Diversity, 11*(3), 365-396. https://doi.org/10.1037/sgd0000722

This page titled 7.7: Summary and APA Style Resources is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **11.2: American Psychological Association (APA) Style** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

8: Experimental Research

8.1: Prelude to Experimental Research

8.2: Initial Equivalence

8.2.1: Natural Groups Research

8.3: Between Groups Designs and Matched Group Designs

8.4: Repeated Measures Design

8.5: Practical Considerations

8.6: Summary of Experimental Research

8: Experimental Research is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

8.1: Prelude to Experimental Research

In the late 1960s social psychologists John Darley and Bibb Latané proposed a counter-intuitive 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)[1].

They also suggested the theory that this phenomenon occurs because each witness feels less responsible for helping—a process referred to as the "diffusion of responsibility." Darley and Latané noted that their ideas were consistent with many real-world cases. For example, a New York woman named Catherine "Kitty" Genovese was assaulted and murdered while several witnesses evidently failed to help. But Darley and Latané also understood that such isolated cases did not provide convincing evidence for their hypothesized "bystander effect." There was no way to know, for example, whether any of the witnesses to Kitty Genovese's murder would have helped had there been fewer of them.

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 he or she 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)[2].

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 Parable of the 38 Witnesses

The story of Kitty Genovese has been told and retold in numerous psychology textbooks. The standard version is that there were 38 witnesses to the crime, that all of them watched (or listened) for an extended period of time, and that none of them did anything to help. However, recent scholarship suggests that the standard story is inaccurate in many ways (Manning, Levine, & Collins, 2007)[3]. For example, only six eyewitnesses testified at the trial, none of them was aware that he or she was witnessing a lethal assault, and there have been several reports of witnesses calling the police or even coming to the aid of Kitty Genovese. Although the standard story inspired a long line of research on the bystander effect and the diffusion of responsibility, it may also have directed researchers' and students' attention away from other equally interesting and important issues in the psychology of helping—including the conditions in which people do in fact respond collectively to emergency situations.

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.

References

1. Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 4*, 377–383. ↵
2. Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 4*, 377–383. ↵

3. Manning, R., Levine, M., & Collins, A. (2007). The Kitty Genovese murder and the social psychology of helping: The parable of the 38 witnesses. *American Psychologist, 62*, 555–562. ↵

8.1: Prelude to Experimental Research is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **5.1: Prelude to Experimental Research** has no license indicated.

8.2: Initial Equivalence

Learning Objectives

1. Recognize examples of studies that are experiments and studies that are not experiments.
2. Explain how random assignment and initial equivlance of groups leads to the ability to show cause-and-effect relationships.

What Is an Experiment?

A true 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 (the independent variable) cause a change in another variable (the dependent variable). Experiments have two fundamental features: Initial equivalence and ongoing equivalence.

Initial Equivalence

As you might guess, initial equivalence means that groups start (are initially) similar (equivalent). What this means in practical terms is usually that the researcher has control over the IV, and randomly assigns participants into the different IV groups; these different groups, or levels of the independent variable, are called conditions. For example, in Darley and Latané's (1968) experiment, the independent variable was the number of witnesses that participants believed to be present. Darley and Latané (1968) 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).

While Darley and Latané (1968) did not use the phrase "random assignment," the data provided and their interpretation of the results suggest that suggest that they used random assignment to put participants into the IV conditions. Random assignment is putting participants in the different IV conditions randomly. This can be done through any computerized randomizer, or flipping a coin, rolling a die, picking a condition out of a hat, or any method that randomly distributes participants into the conditions. If an assignment method probably leads to higher proportions of certain kinds of people in any one condition, then it is not random. For example, you could alternate assigning participants into conditions as they enter such that the first person to arrive for the study is assigned into the one-person group, the second person is assigned into the two-person group, the third person is assigned into the five-person group, the fourth person to arrive is assigned to the one-person group, and so on. What wouldn't be random is assigning the first 10 people to arrive for the study are assigned to the one-person group, the next 10 to arrive to the two-person group,a nd the last 10 people to arrive to the five-person group because there may be something different about the type of people who arrive early compared to the type of people to arrive late.

There are two things to understand about random assignment and initial equivalence. First, random assignment does not ensure that all relevant characteristics are randomly distributed across the IV conditions. Randomization relies on probability, which is, by definition, not a sure thing. While it's a great method to try to control for variables that are not the IV and might affect the DV, we could always somehow still end up with a bunch of really helpful people in the one-person condition. That's why replication of research findings is so important.

The second important idea to note is that there may be other ways to reach initial equivalence other than random assignment. You could use the same participants in all IV conditions or use pretest scores; this would results in initial equivalence between the two groups if you assume that the participants are similar to themselves at the beginning of each part of the study. For example, you could have participants complete a survey on digital stress at the beginning of the study, then limit participants' access to social media, then have the participants re-take the survey on digital stress. Analysis would compare each individuals' first score (pretest) to their second score (posttest) to see if there were any changes in digital stress. If nothing else changed in the participants' lives, we could be fairly confident that it was the social media restrictions and nothing else that could have reduced their digital stress.

The thing to remember is that we want initial equivalence so that we can argue that any changes in the DV **must** be due to the IV since everything else started out similar between the IV conditions. Random assignment is the most trusted and most commonly used method to try to attain initially equivalence IV groups.

Random Assignment [1]

The primary way that researchers attempt to reach initial equivalence is through 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 (sometimes called random selection). Random sampling, discussed in the chapter on the scientific method, is a method for selecting a sample from a population based on probability to represent the population. Random assignment is a method for assigning participants in a sample to the different IV conditions.

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. 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.

Random assignment is not guarantee of initial equivalence. 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, is usually considered a strength of a research designs that attempt to show cause-and-effect relationships.

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 limiting social media affect digital stress, a researcher might instruct some participants to limit their social media use to less than 1 hour a day and others to limit their social media use to 1-2 hours a day. 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 “1 hour a day” and the “2 hours a day.”

Notice that the manipulation of an independent variable must involve the active intervention of the researcher. This is why (random) assignment to reach initial equivalence is so connected to the researcher being the one who is responsible for experiencing the IV conditions.

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 digital stress of people who already spend less than an hour a day of social media with the digital stress of people who spend 1-2 hours on social media has not manipulated this variable and therefore has not conducted an experiment; this is a natural groups design, which will be discussed in more detail next. 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 already spend less than an hour a day on social media might not respond as strongly to fear of missing out or other indicators of digital stress compared to people who spend 1-2 hours a day on social media. Therefore, any observed difference between the two groups in terms of their digital stress might have been caused by the time spend on social media, or it might have been caused by any of the other differences between people who spend different amounts of time on social media. Thus the active

manipulation of the independent variable is crucial for eliminating potential alternative explanations for the results. Random assignment is the method that researchers most often use to reach initial equivalence; with initial equivalence between the groups, there are no other alternative explanations for why the DV changed other than the IV condition.

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 hypochondriasis. This caveat does not mean it is impossible to study the relationship between early illness experiences and hypochondriasis—only that it must be done using nonexperimental approaches. We will discuss this type of methodology in detail later in the book.

The "Other" IV Condition

In these kinds of experiments, there are often at least two groups to compare. The group that the researcher is most interested in is the one that they think causes changes in the DV. This group is often called the treatment group or an intervention group. To determine whether an intervention works, though, there needs to be at least one other group that is initially equivalent to the intervention group as a comparison. This is the "other" IV condition, the one that's not the one being tested. Students new to research first think about the control group (or control condition), in which participants do not receive the treatment. If participants in the intervention condition end up with statistically significantly different scores on the DV than participants in the control condition, then the researcher can conclude that the intervention works. 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 similar participants who are not currently receiving it but who still expect to improve (eventually). If participants were randomly assigned to the intervention condition or the wait-list control condition, then the two groups would (probably) have initial equivalence.

But does it really make sense to compare the intervention group with a group who received nothing at all? If a psychiatrist thinks that a new drug would work better for their clients, would it really make sense to compare that to clients who aren't taking any medication? No! We would want to compare the mental health outcome (DV) with the new IV condition (new drug) and an established treatment (commonly-used drug). In these studies, participants are randomly assigned into the intervention condition and to another treatment condition. The other treatment conditions could be a placebo condition (which is really not a treatment, but the participants think that it is) or a different treatment (comparison group). In research on the effectiveness of psychotherapies and medical treatments, this type of experiment is often called a randomized clinical trial (or RCT), and is often considered the gold standard of research.

Placebo

Let's discuss the comparison with placebos first. A placebo is a simulated treatment that lacks any active ingredient or element that should make it effective, and a placebo effect is changes in the DV based on such a treatment (compared to no real or fake treatment at all). In medical studies, participants would 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.

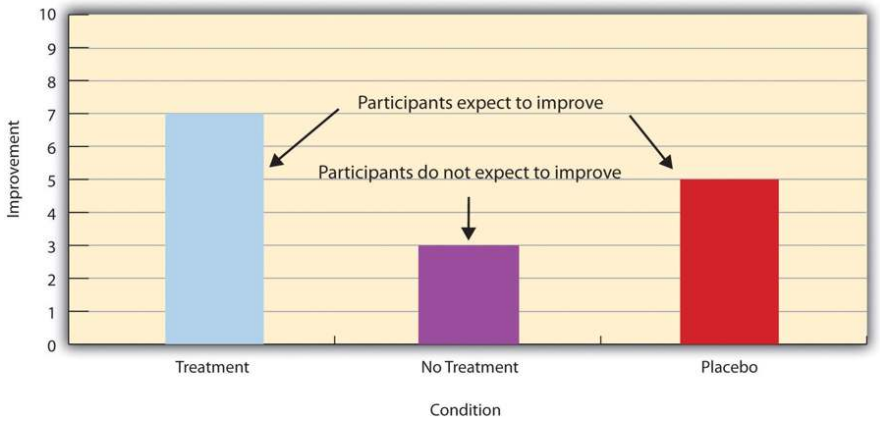


Figure 8.2.1: Hypothetical Results From a Study Including Treatment, No-Treatment, and Placebo Conditions

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 "The Powerful Placebo" box below or Psychology Today's page on placebos), but they also pose a serious problem for researchers who want to determine whether a treatment works. Figure 8.2.1 shows some hypothetical results in which participants in a treatment condition improved more on average than participants in a control condition. If these conditions (the two leftmost bars in Figure 8.2.1) 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 control condition did not. 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 8.2.1).

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).

To learn more about this and other studies on the placebo effect, check out this Hidden Brain podcast episode (Vedantam, 2019) on your favorite podcast app.

Comparison Groups

An additional option to compare the target IV condition with is a comparison group. These IV conditions are usually created by randomly assigning the participants into the currently used treatment. For a psychiatrist, this could be continuing to provide the same drug that you've been prescribing to the comparison group. Some call these control groups, which is correct in the sense that the IV condition is not receiving the target intervention. However, it is more clear to identify whether the group is not receiving anything (control group) or is receiving something but not what we're testing (comparison group).

The use of comparison groups is helpful to test whether the cost and effort of any intervention is worth it. Let's go back to our psychiatrist. Imagine that the cost of the current drug is pretty low, but the cost of the new, comparison treatment is pretty high. We want clients to receive the most effective treatment, but a comparison of the effectiveness of the new drug with the old drug may show that the new drug is equally effective. In this case, there's no benefit to switching to the new, higher cost treatment. If the psychiatrist compared the new drug to a wait-list control group, they would not have this important information about the comparative effectiveness of the new treatment.

Final Considerations of Comparisons

You may have figured it out, but it is good to make it clear: You can combine all of these control and comparison groups in one study! A psychiatrist could compare participants who:

- Receive a new drug treatment (intervention)
- Receive no drug treatment at all, but will receive medication once the study is completed (wait-list control)
- Receive a sugar pill (placebo)
- Receive the commonly-used drug (comparison group)

While that would require more participants, you may decide that the information gained in the comparisons is worth it.

A last consideration about all of these comparison group options is the principle of informed consent. Participants must be informed of what condition they may be assigned to, so they should be told that they will be assigned to *either* a treatment or control group *or* a wait-list control group *or* a placebo condition *or* a comparison group. Researchers have decided that it is both necessary and ethical to not tell the participants which of these groups each individual participant is in until the experiment ends.

Ongoing Equivalence

The second fundamental feature of an experiment is ongoing equivalence. As you might guess, ongoing equivalence means that the IV groups continue to be similar, other than what they experience with the IV. Note that you can't have ongoing equivalence if you didn't start out with initial equivalence; two groups can't stay similar if you didn't start out similar. For ongoing equivalence, the researcher exerts control over, or minimizes the variability in, variables other than the independent and dependent variable, so that the only difference between the groups is the IV. These other variables are called extraneous variables. Darley and Latané (1968) tested all their participants in the same room, exposed them to the same emergency situation, and so on. 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.

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. angry) 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 video clip designed to elicit happiness or anger) and then asked to recall as many happy childhood events as they can. The two leftmost columns of Table **8.2.1** 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 angry 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 **8.2.1**. 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 angry 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 **8.2.1**, which makes the effect of the independent variable easier to detect (although real data never look quite *that* good).

Table **8.2.1**: Hypothetical Number of Happiness Memories Recalled (Noiseless Data and Realistic Noisy Data)

Idealized "noiseless" data		Realistic "noisy" data	
Happy Video	**Angry Video**	**Happy Video**	**Angry Video**
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
$M = 4$	$M = 3$	$M = 4$	$M = 3$

One way to avoid 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 (Knecht et al., 2000). 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 control extraneous variables by limiting participants to one very specific category of person, such as 20-year-old, white, bisexual, right-handed psychology majors who are cis-gender women. The obvious downside to this approach is that it would lower the external validity of the study—in particular, the extent to which the results can be generalized beyond the people actually studied. For example, it might be unclear whether results obtained with a sample of younger Black women would apply to older Latino men. In many situations, the advantages of a diverse sample (increased external validity) outweigh the reduction in noise achieved by a homogeneous one.

Confounding Variables Can Affect Initial or Ongoing Equivalence

As discussed in the Empirical Study section of the introduction to the scientific method, 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). A confound can affect initial equivalence if there's no random assignment, or it can affect ongoing equivalence if the groups started out similar but something affected only one IV group during the study. The example of mood and memories could show confounds related to initial equivalence for any differences that random assignment didn't effectively distribute across the two IV conditions. There could also be a confound related to ongoing equivalence if all of the happy video clips were shown in Classroom A and all of the angry video clips were shown in Classroom B, Classroom B had a really noisy heater. While it might seem like the happy video clips *caused* higher recall of happy memories, it might be that the noisy classroom reduced the total number of total memories that could be recalled. The classroom noise was a confound, affecting only one IV group. As you can see, confounds make it so that you can't state that the IV caused changes in the DV (even if you used random assignment to reach initial equivalence) because you did not have ongoing equivalence. 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.

References

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 4*, 377–383.

Knecht, S., Dräger, B., Deppe, M., Bobe, L., Lohmann, H., Flöel, A., . . . Henningsen, H. (2000). Handedness and hemispheric language dominance in healthy humans. *Brain: A Journal of Neurology, 123*(12), 2512-2518. http://dx.doi.org/10.1093/brain/123.12.2512

Moseley, J. B., O'Malley, K., Petersen, N. J., Menke, T. J., Brody, B. A., Kuykendall, D. H., … Wray, N. P. (2002). A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *The New England Journal of Medicine, 347*, 81–88.

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.

Shapiro, A. K., & Shapiro, E. (1999). *The powerful placebo: From ancient priest to modern physician*. Baltimore, MD: Johns Hopkins University Press.

Footnote

1. This section was originally in Jhangiani et al.'s (2019) section 5.3, while the rest of the information from Jhangiani et al. (2019) in this page is from section 5.2. ↵

This page titled 8.2: Initial Equivalence is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **5.2: Experiment Basics** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.
- **5.3: Experimental Design** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.
- **5.5: Practical Considerations** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

8.2.1: Natural Groups Research

Learning Objectives

1. Explain why a researcher might choose to conduct natural groups design rather than an experimental design that compares randomly assigned IV groups.
2. Explain why correlation does not imply causation.

What if there's no random assignment?

Experimental designs with random assignment of participants into the IV groups are the best way to show that an IV group *caused* changes in the DV. This is because everything else that could have caused changed in the DV should have been distributed randomly across the different IV groups. There are some ideas to unpack here. First, the other potential causes should have be randomly distributed, but unless you measure these potential causes then you can't be sure; random assignment is based on chance. Second, this type of research methodology's purpose is to try to support the hypothesis that one variable *caused* changes in another variable, but sometimes it's less important to know what is the cause if we know that the variables are somehow related.

Sometimes we want to show that one variable *caused* changes in the other variable, but what we think is the cause can't be randomly assigned. When the researcher doesn't have control over the IV and can't randomly assign participants, then the research is a natural groups design; these types of designs are sometimes called posttest-only nonequivalent groups design, but that name will be discussed later in the chapter on quasi-experimental designs. This is most common when what the researcher thinks is the cause is either part of the participant (like their social group or their opinion) or it is something that the participant chose. For example, as a researcher, we cannot randomly assign participants into a demographic group (like race, gender, sexual orientation, etc.) or whether they like pineapples on pizza or not. When choice is involved, sometimes the variable *could* be randomly assigned, but it is not. For example, if an instructor wanted to know if eating before an exam improved performance, they could either randomly assign students into two groups (eat before the exam or not) and expect the students to follow the directions of their assigned group, or the instructor could ask students if they ate before the exam (no random assignment; the example with random assignment is an experiment, while the example in which the students were merely asked about their eating would be a natural groups design. The distinction is important because if the study was an experiment, then it could be concluded that making eating before can exam improves grades (cause-and-effect), but if it was a natural groups design, we could only conclude that these variables are related. Instead of eating being what caused better exam performance, it could be that students who have more time to eat before an exam also have more time to study before an exam; thus, while there could be a relationship between eating and exam performance, it might be that the real cause of exam performance is study time.

> There are no IVs in natural groups designs (technically), but we still use the term IV.

An important note here about independent variables: the definition of an IV is "the variable that the experimenter manipulates, and believes is the cause of changes in the outcome variable," which means that there are technically no IVs in natural groups designs. However, researchers often use the term IV because they think that the one variable is the cause of changed in the other, even if their study does not randomly assign participants into the groups. Plus, it's just easier to use the terminology that we're used to! But by definition, in natural groups designs we have predictor variables and outcome variables, not IVs and DVs.

Natural Groups or Correlational Research?

Many social science researchers use the term correlational research to describe studies that do not have random assignment. Dr. MO chose to use the term natural groups design to avoid the common misconception that correlational research must include a correlational statistical analysis (which requires at least two quantitative variables). The defining feature of natural groups designs is that the predictor and outcome variables are measured, neither one is manipulated. 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 an experiment, it is a natural groups design because the researcher did not (and cannot) manipulate the students’ nationalities. The defining feature of natural groups designs

research is that no variables are 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 could be natural groups designs because no independent variable is manipulated. The crucial point is that what defines a study as experimental or natural groups is not the variables being studied nor the type of statistics used to analyze the data; what defines a study is *how* the study is conducted.

Correlation Does Not Imply Causation

Remember in the Analysis chapter when you were reminded of correlations, and there was a section talking about correlations versus causation? This idea becomes relevant again when discussing natural groups designs. An amusing example of this comes from Messerli's (2012) study that showed a positive correlation (Pearson's $r = 0.79$) between the per capita chocolate consumption of a nation and the number of Nobel prizes awarded to citizens of that nation. It seems clear, however, that this does not mean that eating chocolate causes people to win Nobel prizes, or that Nobel prizes cause citizens to eat more chocolate! There are two reasons that showing that two variables are related does not mean that the predictor causes the outcome.

The first is called the directionality problem. Two variables, *X* and *Y*, can be related because *X* causes *Y* or because *Y* causes *X*. This means that one variable may be causing changes in the other variable, the research design cannot tell is which one is the causes. For example, we might find that exercise is related to how happiness such that people who exercise are happier on average than people who do not. This 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. With a natural groups design, we cannot be sure which is the causal variable.

The second reason that showing that two variables are related does not mean that the predictor causes the outcome 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*. Figure **8.2.1.1** is a poster created by one of Dr. MO's students to visual this third variable problem, which the class started to refer to as "What about Z?".

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 relationship between exercise and happiness could mean that some third variable, such as social class, *causes* both of the others. Having disposabile income could allow people more time to exercise while also leading to less stress (and more happiness).

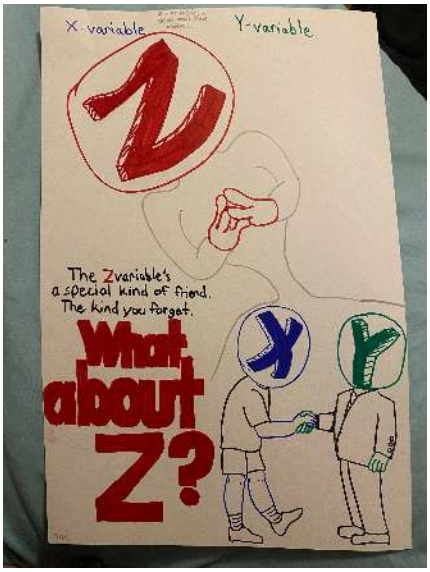


Figure **8.2.1.1**: What about Z? by Laurel Woodruff CC-BY-NC-SA

Some excellent and amusing examples of spurious correlations can be found at http://www.tylervigen.com (Figure **8.2.1.2** provides one such example).

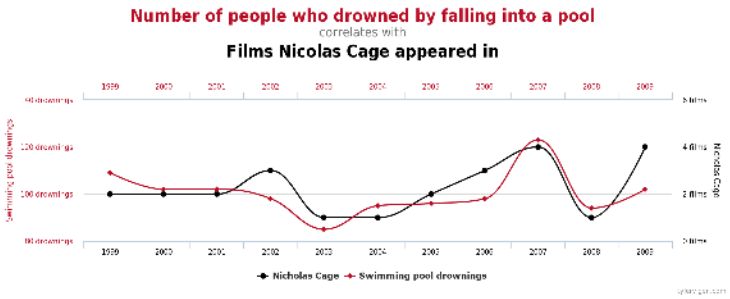


Figure **8.2.1.2**: Example of a Spurious Correlation. Source: http://tylervigen.com/spurious-correlations (CC-BY 4.0)

"Lots of Candy Could Lead to Violence"

Although researchers in psychology know that correlation does not imply causation, many journalists do not. One website about correlation and causation, http://jonathan.mueller.faculty.noctrl.edu/100/correlation_or_causation.htm, links to dozens of media reports about real biomedical and psychological research. Many of the headlines suggest that a causal relationship has been demonstrated when a careful reading of the articles shows that it has not because of the directionality and third-variable problems.

One such article is about a study showing that children who ate candy every day were more likely than other children to be arrested for a violent offense later in life. But could candy really "lead to" violence, as the headline suggests? What alternative explanations can you think of for this statistical relationship? How could the headline be rewritten so that it is not misleading?

As you have learned by reading this book, there are various ways that researchers address the directionality and third-variable problems. The most effective is to conduct an experiment. For example, instead of simply measuring how much people exercise, a researcher could bring people into a laboratory and randomly assign half of them to run on a treadmill for 15 minutes and the rest to sit on a couch for 15 minutes. Now, if the exercisers end up in more positive moods than those who did not exercise, it cannot be because their moods affected how much they exercised (because it was the *researcher* who used random assignment to determine how much they exercised). Likewise, it cannot be because some third variable (social class) affected both how much they exercised and what mood they were in. Experiments eliminate the directionality and third-variable problems and allow researchers to draw firm conclusions about causal relationships.

Why choose a natural groups design?

There are a few reasons that researchers might choose to conduct a natural groups design rather than an experiment. The first is that they are not interested in determining if there is a causal relationship between variables. Recall two goals of science are to describe and to predict; natural groups designs allow researchers to achieve both of these goals. Specifically, this strategy can be used to describe the statistical relationship between variables, and (if there is a relationship between the variables) the researchers can use scores on one variable to predict scores on the other.

Another reason that researchers would choose to use a natural groups rather than an experiment is that the they do believe that the predictor variable (sorta like an IV) does cause changed in the outcome variable (DV), but the researcher *cannot* manipulate the independent variable because it is impossible, impractical, or unethical. For example, while a researcher might be interested in the relationship between the frequency people use cannabis and their memory abilities they cannot ethically manipulate the frequency that people use cannabis. As such, they must rely on natural groups by asking about the frequency that people use cannabis and measure their memory abilities using a standardized test of memory to determine if there is a predictive relationship between the frequency people use cannabis and memory test performance.

Another strength of natural groups designs is that it is often higher in external validity than experimental research. Recall there is typically a trade-off between internal validity and external validity. As greater controls are added to experiments, internal validity is increased but often at the expense of external validity as artificial conditions are introduced that do not exist in reality. Natural group design may have low internal validity because nothing is manipulated or controlled but 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 natural groups design that is high in external validity, then the researchers can have more confidence in the validity of their theory. As a concrete example, correlational studies establishing that there is a relationship between watching violent television and aggressive behavior have been complemented by experimental studies confirming that the relationship is a causal one (Bushman & Huesmann, 2001).

References

Bushman, B. J., & Huesmann, L. R. (2001). Effects of televised violence on aggression. In D. Singer & J. Singer (Eds.), *Handbook of children and the media* (pp. 223–254). Thousand Oaks, CA: Sage.

Messerli, F. H. (2012). Chocolate consumption, cognitive function, and Nobel laureates. *New England Journal of Medicine*, *367*, 1562-1564.

This page titled 8.2.1: Natural Groups Research is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **6.3: Correlational Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

8.3: Between Groups Designs and Matched Group Designs

Learning Objectives

1. Explain the difference between group designs and matched group designs.
2. Develop ideas about the benefits and disadvantages of between groups and matched group designs.

In this section, we look at two different ways to design an experiment by randomly assigning individual participants into IV conditions so that the IV conditions are expected to have initial equivalence: between groups designs and matched group designs. The following section will focus on within groups designs, with repeated measures designs being the easiest to understand.

Between Group Design

In a between groups design, each participant in the experiment is randomly assigned into only one condition. For example, a researcher with a sample of 100 participants might assign half of them to limit their social media use and the other half of the participants are not asked to change their social media usage by having each participant flip a coin; heads means that they should limit their social media use while tails means that they do not. Or a researcher with a sample of 90 people with severe agoraphobia (fear of open spaces) might randomly assign 30 of them to receive each of three different treatments for that disorder. The randomness of the assignment procedure into the IV group to ensure initial equivalence between the groups is an important characteristic of this process; it is essential in a between groups design that the researcher assigns participants to conditions so that the different groups are, on average, highly similar to each other. Those in the limited social media use condition and those in the unlimited social media condition, for example, should include a similar proportion of each gender, and they should have similar average levels of anxiety, similar average age, and so on. The random assignment used in hope of controlling these extraneous participant variables across conditions so that they do not become confounding variables.

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 a variable that is not the IV or DV, but might be expected to affect the DV, prior the assignment into an IV condition and manipulation of the independent variable. Then, the matched groups are randomly assigned into the various IV conditions, ensuring that there is initial equivalence between the IV conditions on at least that one variable.



Figure **8.3.1**: Basketball Through Hoop by Airman 1st Class Kerelin Molina via https://commons.wikimedia.org/wiki/F...rough_hoop.jpg . *This image or file is a work of a U.S. Air Force Airman or employee, taken or made as part of that person's official duties. As a work of the U.S. federal government, the image or file is in the **public domain** in the United States*

This is a confusing concept for students to grasp, here's an example: Imagine that your Physical Education (PE) instructor wanted to know what kind of practice most helps to improve making lay-ups in basketball. The PE instructor thinks that practicing shooting from the key would help, but also notices that many players struggle with lay-ups because they dribble the basketball poorly. So, the instructor's two IV conditions will focus on shooting from the key or practicing dribbling. But one other factor might affect lay-up skill, which is height. Imagine your most recent PE class lined up from tallest to shortest. Then, pairs of students were made based on their similar heights; the two tallest students would be paired, then the next two tallest students, and so on until the two shortest students are paired. Then, each pair (who have similar heights) would use a random procedure, such as playing rock/paper/scissors (sometimes called Rochambeau, Roshambo, Ro-sham-bo, Bato Bato Pik, or Jak-en-poy, Wikipedia, 2025) to determine IV group assignment. The winner could be assigned to the group to practice shooting from the key, and the

rock/paper/scissors loser could be assigned to the group that practiced dribbling. After all pairs are assigned, there would be two groups: a group practicing shooting from the key and a group practicing dribbling. The key is that the average height of each group would be similar; there would be tall students in the group practicing shooting from the key and there would be tall students in the group who are practicing dribbling. There would also be shorter students in each group. Because the students were matched on height and then randomly assigned into the IV condition from that matched pair, we are very confident that the groups will be initially equivalent on height; height will not be an extraneous variable that could affect scores on the DV of making lay-ups. And because we still used random assignment, we should have initial equivalence on all other variables. At the end of the study, if the PE instructor found that number of lay-ups made was statistically higher for students from one type of practice, we would know that the results were not influenced by height; the matching guarantees that the matched variables will not be confounded across the experimental conditions.. However, we can't be sure of this unless we measure each variable that we are concerned about.

References

Wikipedia (2025, July 2). Rock paper scissors. https://en.wikipedia.org/wiki/Rock_paper_scissors

This page titled 8.3: Between Groups Designs and Matched Group Designs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **5.3: Experimental Design** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

8.4: Repeated Measures Design

Learning Objectives

1. Explain the difference between between groups designs and repeated measures design, list some of the pros and cons of each approach, and decide which approach to use to answer a particular research question.
2. Explain how counterbalancing helps to deal with carryover effects.

In this section, we look at another way to design an experiment. The primary distinction from the between groups and matched groups design that we previously discussed is whether each participant experiences one level of the independent variable or each participant experiences *all* levels of the independent variable. The latter are called within subjects experiments, but the easiest to understand is the repeated measures design.

Repeated Measures Design

In repeated measures designs, each participant is tested under all conditions, but the order that each participant experiences each IV condition is randomized. Remember our between groups design comparing digital stress of participants who were randomly assigned into two groups: half of the participants were randomly assigned into the IV condition of limiting their social media use, while the other half were randomly assigned to not limit their social media use. In a repeated measures design, all participants would experience each of the IV conditions but the researcher would instead randomly assign the *order* of the IV conditions. Half of the participants would first be limited in their social media use for two weeks, and then those same participants could use as much social media as they'd like for two weeks. The other half of the participants would first use social media as much as they'd like for two weeks, and then be limited in their social media use for two weeks. At the end of *each* two week period, *all* participants would have their digital stress assessed.

The primary advantage of this approach is that it provides maximum control of extraneous participant variables. Participants in each condition have the same anxiety levels, same socioeconomic status, same number of siblings, and so on, because *they are the very same people*. Repeated measures experiments, and matched groups designs, 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.

Your Research Ideas: Between Groups, Matched, or Repeated Measures?

Not all experiments can use a repeated measures design, nor would it be desirable to do so. Think about an IV that a professional in your future career might be interested in investigating its effects on an outcome to help their clients. Based on that IV, what kind of design might work best?

- Between Groups design?
- Matched Groups design? What would be the matching variable? Note that the matching variable should not be the IV nor the DV, but might be expected to influence the DV.
- Repeated Measures design? Are you sure that the IV can be experienced in any order?

Counterbalancing

Counterbalancing is testing different participants in different orders. This was included in the definition of repeated measures designs, that the order of the IV conditions is randomized for each participants. The digital stress example above also included counterbalancing; half of the participants were asked to limit their social media use at the beginning of the study, and the other half of the participants were asked to limit their social media use at the end of the study.

Based on the need for counterbalancing, you can see that the primary disadvantage of repeated measures 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, digital stress may be higher for participants who first limited their social media use and then had unlimited access. In contrast to the social media break, the unlimited use might feel overwhelming. Note that this would not be the case for participants whose first IV condition was the unlimited condition. In other words, the order of the conditions may be a confounding variable. 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.

The best method of counterbalancing is *complete counterbalancing* in which an equal number of participants complete each possible order of conditions. However, for practical issues or due to the competency of the researcher, sometimes that does not happen. We've already covered an example when there are two IV conditions when discussing the the IV conditions of limited or unlimited social media use. With three conditions, there would be six different orders (ABC, ACB, BAC, BCA, CAB, and CBA), so some participants would be tested in each of the six orders. With four conditions, there would be 24 different orders; with five conditions there would be 120 possible orders. With counterbalancing, participants are assigned to orders randomly. Thus, random assignment plays an important role in repeated measures designs just as in between groups designs. Here, instead of randomly assigning to conditions, participants are randomly assigned to different orders of conditions. In fact, it can safely be said that if a study does not involve random assignment in one form or another, it is not an experiment.

A more efficient way of counterbalancing is *partial counterbalancing through a Latin square design* which randomizes through having equal rows and columns. For example, if you have four treatments, you must have four versions. Like a Sudoku puzzle, no treatment can repeat in a row or column. For four versions of four treatments, the Latin square design would look like Table **8.4.1**:

Table **8.4.1**: Example of Latin Square for counterbalancing

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

In Table **8.4.1**, each condition appears at each ordinal position (A appears first once, second once, third once, and fourth once) and each condition precedes and follows each other condition one time. A Latin square for an experiment with 6 conditions would by 6 x 6 in dimension, one for an experiment with 8 conditions would be 8 x 8 in dimension, and so on. So while complete counterbalancing of 6 conditions would require 720 orders, a Latin square would only require 6 orders.

When the number of conditions is large, experiments can use *random counterbalancing* in which the order of the conditions is randomly determined for each participant. Using this technique, every possible order of conditions is determined and then one of these orders is randomly selected for each participant. This is not as powerful a technique as complete counterbalancing or partial counterbalancing using a Latin squares design. Use of random counterbalancing will result in more random error, but if order effects are likely to be small and the number of conditions is large, this is an option available to researchers.

There are two ways to think about what counterbalancing accomplishes. One is that it controls the order of conditions so that it is no longer a confounding variable. Instead of the limited social media condition always being first and the unlimited condition always being second, the limited social media condition comes first for some participants and second for others. Likewise, the unlimited social media condition comes first for some participants and second for others. Thus, any overall difference in the dependent variable of digital stress between the two conditions cannot have been caused by the order of conditions. A second way that counterbalancing is helpful is that if there are carryover effects, counterbalancing makes it possible to detect them because the

different orders can be included in statistical analyses. One can analyze the data separately for each order to see whether it had an effect.

Disadvantages of Repeated Measures Designs

In addition to order effects, another disadvantage of repeated measures designs could be that it might be easier for participants to guess the hypothesis. For example, a participant who is asked about their digital stress after two limiting and two weeks not limiting their social media use (in whatever order) may recognize that the research is about social media's effect on digital stress. This knowledge could lead participants to rate their digital stress in the direction that they think is expected.

Type of Design

Many research topics can be studied using a between groups design, matched group design *or* a repeated measures design, so you must decide which of the approaches best fits the research question based on the design's relative merits for the particular situation.

Between groups designs have the advantage of being conceptually simpler and requiring less testing time per participant. They also avoid carryover effects without the need for counterbalancing. Matched group designs also avoid carryover effects, but take a little more time and effort because the matching variable must be measured and then the participants must be matched then randomly assigned into an IV condition. Repeated measures designs have the advantage of controlling more extraneous participant variables than either matched groups or between groups, which generally reduces noise in the data and makes it easier to detect any effect of the independent variable upon the dependent variable. Repeated measures designs also require fewer participants than the other designs to detect an effect of the same size.

There's no real right answer as the best design will differ based on your participants, IV, and DV. Remember also you are choosing one type of design for one experiment. This does not preclude using another type in a different study. Replication is necessary in research, and using different types of designs that tend to find the same results improves the internal validity of your findings.

This page titled 8.4: Repeated Measures Design is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **5.3: Experimental Design** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

8.5: Practical Considerations

Learning Objectives

1. Describe several strategies for recruiting participants for an experiment.
2. Explain why it is important to standardize the procedure of an experiment and several ways to do this.
3. Explain what pilot testing is and why it is important.

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.

Recruiting Participants

Of course, at the start of any research project, you should be thinking about how you will obtain your participants. Unless you have access to people with schizophrenia or incarcerated juvenile offenders, for example, then there is no point designing a study that focuses on these populations. But even if you plan to use a convenience sample, you will have to recruit participants for your study.

There are several approaches to recruiting participants. One is to use participants from a formal **subject pool** —an established group of people who have agreed to be contacted about participating in research studies. For example, at many colleges and universities, there is a subject pool consisting of students enrolled in introductory psychology courses who must participate in a certain number of studies to meet a course requirement. Researchers post descriptions of their studies and students sign up to participate, usually via an online system. Participants who are not in subject pools can also be recruited by posting or publishing advertisements or making personal appeals to groups that represent the population of interest. For example, a researcher interested in studying older adults could arrange to speak at a meeting of the residents at a retirement community to explain the study and ask for volunteers.



"Study" Retrieved from http://imgs.xkcd.com/comics/study.png (CC-BY-NC 2.5)

The Volunteer Subject

Even if the participants in a study receive compensation in the form of course credit, a small amount of money, or a chance at being treated for a psychological problem, they are still essentially volunteers. This is worth considering because people who volunteer to participate in psychological research have been shown to differ in predictable ways from those who do not volunteer. Specifically, there is good evidence that on average, volunteers have the following characteristics compared with non-volunteers (Rosenthal & Rosnow, 1976)[1]:

- They are more interested in the topic of the research.
- They are more educated.

- They have a greater need for approval.
- They have higher IQ.
- They are more sociable.
- They are higher in social class.

This difference can be an issue of external validity if there is a reason to believe that participants with these characteristics are likely to behave differently than the general population. For example, in testing different methods of persuading people, a rational argument might work better on volunteers than it does on the general population because of their generally higher educational level and IQ.

In many field experiments, the task is not recruiting participants but selecting them. For example, researchers Nicolas Guéguen and Marie-Agnès de Gail conducted a field experiment on the effect of being smiled at on helping, in which the participants were shoppers at a supermarket. A confederate walking down a stairway gazed directly at a shopper walking up the stairway and either smiled or did not smile. Shortly afterward, the shopper encountered another confederate, who dropped some computer diskettes on the ground. The dependent variable was whether or not the shopper stopped to help pick up the diskettes (Guéguen & de Gail, 2003) [2]. There are two aspects of this study that are worth addressing here. First, n otice that these participants were not "recruited," which means that the IRB would have taken care to ensure that dispensing with informed consent in this case was acceptable (e.g., the situation would not have been expected to cause any harm and the study was conducted in the context of people's ordinary activities). Second, even though informed consent was not necessary, the researchers still had to select participants from among all the shoppers taking the stairs that day. I t is extremely important that this kind of selection be done according to a well-defined set of rules that are established before the data collection begins and can be explained clearly afterward. In this case, with each trip down the stairs, the confederate was instructed to gaze at the first person he encountered who appeared to be between the ages of 20 and 50. Only if the person gazed back did they become a participant in the study. The point of having a well-defined selection rule is to avoid bias in the selection of participants. For example, if the confederate was free to choose which shoppers he would gaze at, he might choose friendly-looking shoppers when he was set to smile and unfriendly-looking ones when he was not set to smile. As we will see shortly, such biases can be entirely unintentional.

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.

Experimenter's Sex as an Extraneous Variable

It is well known that whether research participants are male or female can affect the results of a study. But what about whether the experimenter is male or female? There is plenty of evidence that this matters too. Male and female experimenters have slightly different ways of interacting with their participants, and of course, participants also respond differently to male and female experimenters (Rosenthal, 1976)[3].

For example, in a recent study on pain perception, participants immersed their hands in icy water for as long as they could (Ibolya, Brake, & Voss, 2004) [4]. Male participants tolerated the pain longer when the experimenter was a woman, and female participants tolerated it longer when the experimenter was a man.

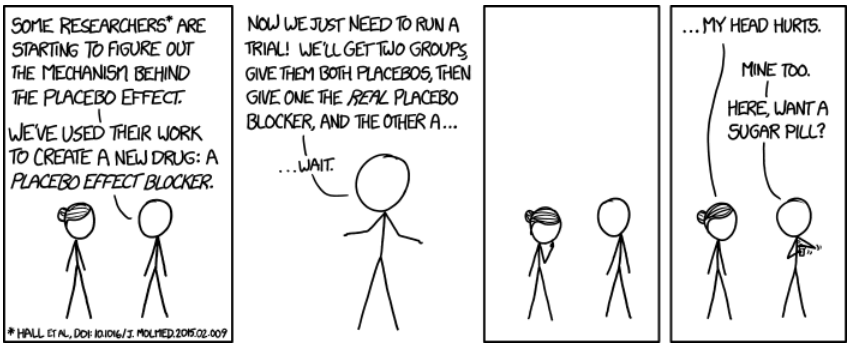
Researcher Robert Rosenthal has spent much of his career showing that this kind of unintended variation in the procedure does, in fact, affect participants' behavior. Furthermore, one important source of such variation is the experimenter's expectations about how participants "should" behave in the experiment. This outcome is referred to as an **experimenter expectancy effect** (Rosenthal, 1976)[5]. 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) [6]. 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.



“Placebo Blocker” retrieved from http://imgs.xkcd.com/comics/placebo_blocker.png (CC-BY-NC 2.5)

Record Keeping

It is essential to keep good records when you conduct an experiment. As discussed earlier, 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 analy z e 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.

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.

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.

References

1. Rosenthal, R., & Rosnow, R. L. (1976). *The volunteer subject*. New York, NY: Wiley. ↵
2. Guéguen, N., & de Gail, Marie-Agnès. (2003). The effect of smiling on helping behavior: Smiling and good Samaritan behavior. *Communication Reports*, *16*, 133–140. ↵
3. Rosenthal, R. (1976). *Experimenter effects in behavioral research* (enlarged ed.). New York, NY: Wiley. ↵
4. Ibolya, K., Brake, A., & Voss, U. (2004). The effect of experimenter characteristics on pain reports in women and men. *Pain*, *112*, 142–147. ↵
5. Rosenthal, R. (1976). *Experimenter effects in behavioral research* (enlarged ed.). New York, NY: Wiley. ↵
6. Rosenthal, R., & Fode, K. (1963). The effect of experimenter bias on performance of the albino rat. *Behavioral Science*, *8*, 183-189. ↵

This page titled 8.5: Practical Considerations is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **5.5: Practical Considerations** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

8.6: Summary of Experimental Research

Key Takeaways

- An experiment is a type of empirical study that features the manipulation of an independent variable, the measurement of a dependent variable, and control of extraneous variables.
- An extraneous variable is any variable other than the independent and dependent variables. A confound is an extraneous variable that varies systematically with the independent variable.
- Experimental research on the effectiveness of a treatment requires both a treatment condition and a control condition, which can be a no-treatment control condition, a placebo control condition, or a wait-list control condition. Experimental treatments can also be compared with the best available alternative.
- Experiments can be conducted using either between groups, matched groups, or repeated measures design. Deciding which to use in a particular situation requires careful consideration of the pros and cons of each approach.
- Random assignment is a fundamental element of experimental research. The purpose of these techniques is to control extraneous variables so that they do not become confounding variables.
- Studies are high in internal validity to the extent that the way they are conducted supports the conclusion that the independent variable caused any observed differences in the dependent variable. Experiments are generally high in internal validity because there is initial equivalence of the participants in each IV group, and there's ongoing equivalence of the IV groups.
- Initial equivalence and ongoing equivalence provides internal validity because they show that nothing other than the IV could have caused changes in the DV. Initial equivalence is usually created through random assignment. Ongoing equivalence is usually created through having initial equivalence, and then attempting to control all other extraneous variables.
- It is important to standardize experimental procedures to minimize extraneous variables, including experimenter expectancy effects.
- It is important to conduct one or more small-scale pilot tests of an experiment to be sure that the procedure works as planned.

What's Next?

Now that we've covered the components of experiments, we're going to talk about other types of studies.

Exercises

- Practice: List five potential IVs that can be randomly assigned and manipulated by the researcher in an experiment. List five potential IVs that cannot be randomly assigned and manipulated by the researcher in an experiment.
- Practice: For each of the following topics, decide whether that topic could be studied using an experimental research design that included randomm assignment to reach initial equivalence between the IV conditions, and explain why or why not.
  - Effect of parietal lobe damage on people’s ability to do basic arithmetic.
  - Effect of being clinically depressed on the number of close friendships people have.
  - Effect of group training on the social skills of widowed men.
  - Effect of paying people to lose weight.
- Discussion: Imagine that a study found that writing a story to learn 10 vocabulary words improved memory more than re-writing the same 10 words and their definitions multiple times by the same participants. If the story condition was always tested after the re-writing condition, explain a fundamental problem with this research design and at least two ways that it might be corrected.
- Discussion: For each of the following topics, decide which would be better.
  - You want to test the relative effectiveness of two training programs for running a marathon.
  - Using photographs of people as stimuli, you want to see if smiling people are perceived as more intelligent than people who are not smiling.
  - In a field experiment, you want to see if the way a donation letter is written affects whether or not passersby give him any money.
  - You want to see if concrete nouns (e.g., dog) are recalled better than abstract nouns (e.g., truth).

- Discussion: Imagine a study in which you will visually present participants with a list of 20 words, one at a time, wait for a short time, and then ask them to recall as many of the words as they can. In the stressed condition, they are told that they might also be chosen to give a short speech in front of a small audience. In the unstressed condition, the participants are not told that they will not have to interact with anyone else. What are several specific things that you could do to standardize the procedure?

This page titled 8.6: Summary of Experimental Research is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **5.6: Experimental Research (Summary)** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

9: Quasi-Experimental Research

9.1: Prelude to Quasi-Experimental Research
9.2: Quasi-Experimental Designs with One Group
9.3: Quasi-Experimental Designs with Two or More Groups
9.4: Threats to Internal Validity
9.5: Summary of Quasi-Experimental Research

9: Quasi-Experimental Research is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

9.1: Prelude to Quasi-Experimental Research

The prefix *quasi* means “resembling.” Thus quasi-experimental research is research that resembles experimental research but is not true experimental research.

Hatzenbuehler et al. (2012) is an example of a type of quasi-experimental design. Hatzenbuehler et al. (2012) noted previous research showing that bans on same-gender marriage was related to increased mental health issues but lower rates for medical and mental health care use. If the banning of rights seem to be related to mental and medical issues, could providing legal recognition be related to improved mental health? To test this, medical care visits, medical care costs, mental health care visits, and mental health care costs were compared for gay and bisexual men in Massachusets between a year before Massachusets became the first state to legally recognize same-gender marriage (Hatzenbuehler et al., 2012) to a year after same-gender marriage was legally recognized in the state. In support, Hatzenbuehler et al. (2012) found that both the number of visits and the costs of both medical and mental health care decreased after legal recognition of same-gender marriage for this sample of gay and bisexual men. So how is this not a true experiment? The next section will help you understand, and describe other quas-experimental designs.

References

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues in field settings*. Boston, MA: Houghton Mifflin.

Hatzenbuehler, M. L, O'Cleirigh, C., Grasso, C., Mayer, K., Safren, S., & Bradford, J. (2012). Effect of same-sex marriage laws on health care use and expenditures in sexual minority men: A quasi-natural experiment. *American Journal of Public Health*, *102*(2), 285-291.

9.1: Prelude to Quasi-Experimental Research is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **8.1: Prelude to Quasi-Experimental Research** has no license indicated.

9.2: Quasi-Experimental Designs with One Group

Learning Objectives

1. Describe at least one way that quasi-experimental designs different from experimental designs.
2. Describe how interrupted time series designs are similar and different from pretest-posttest designs.

Recall that with a true experiment, random assignment to conditions is used to ensure the groups are equivalent or counterbalancing is used to guard against order effects. Quasi-experiments are missing one of these safeguards. In quasi-experimental designs, an independent variable is manipulated, but either a control group is missing or participants are not randomly assigned to conditions (Cook & Campbell, 1979). In this section, we will discuss quasi-experimental designs that compare one group on scores before and after something happens. The following section will discuss similar quasi-experimental designs but that include two or more groups.

One-Group Pretest-Posttest Design

In a one-group pretest-posttest design, the dependent variable is measured once before the treatment is implemented (called a pretest), and once after it is implemented (called a posttest). As described in the prelude, Hatzenbuehler et al. (2012) could be considered a one-group pretest-posttest design. Why? Because there was one group (sexual minority men) in which rates of the DV were measured before implementation of a treatment or event (the law was passed) and again after the event. This type of pretest-posttest design sounds a lot like a repeated measures design experiment because the participants are measured more than once; each participant is tested first under the control condition and then under the treatment condition. However, it is not a repeated measures design because the order of conditions is not counterbalanced. In Hatzenbuehler et al. (2012), the event was legal recognition of same-gender marriage, but we cannot randomly assign the order of time (before/after).

The implementation of the treatment or event is often referred to as an IV because we think that it *caused* changes in the DV, but technically these are not IVs because the definition of independent variables is "the variable that the experimenter manipulates." In quasi-experimental designs, the experimenter was not in control of the IV. So even if if the average posttest score is statistically different than the average pretest score, we cannot conclude that the treatment is responsible for the change because there may be other explanations for why the posttest scores may have changed. These alternative explanations pose threats to internal validity, and are discussed later in this chapter.

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. This is like the one-group pretest-posttest design with the addition of measuring the outcome multiple times. 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 9.2.1 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 9.2.1 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 9.2.1 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.

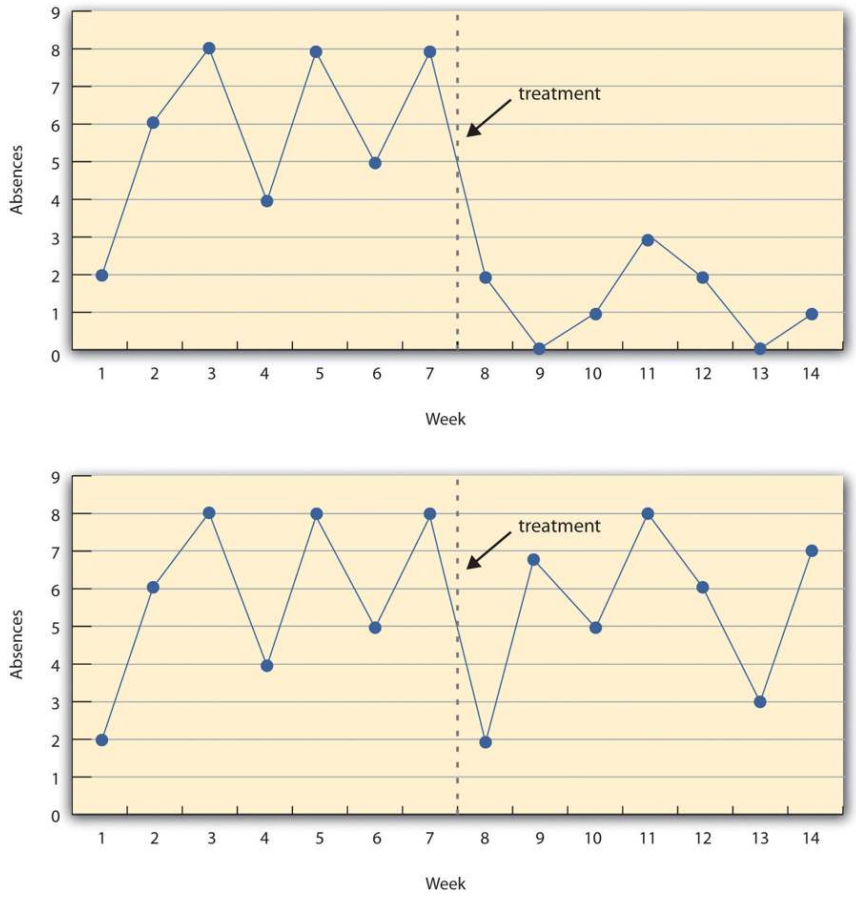


Figure **9.2.1**: A Hypothetical Interrupted Time-Series Design. The top panel shows data that suggest that the treatment caused a reduction in absences. The bottom panel shows data that suggest that it did not.

The prelude didn't get into details, but because Hatzenbuehler et al. (2012) measured medical care visits, medical care costs, mental health care visits, and mental health care costs over a two year period, we can assume that the results could have been analyzed as an interrupted time series design. Instead, Hatzenbuehler et al. (2012) averaged the costs and visits for each time period for their comparisons.

References

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues in field settings*. Boston, MA: Houghton Mifflin.

Hatzenbuehler, M. L, O'Cleirigh, C., Grasso, C., Mayer, K., Safren, S., & Bradford, J. (2012). Effect of same-sex marriage laws on health care use and expenditures in sexual minority men: A quasi-natural experiment. *American Journal of Public Health, 102*(2), 285-291.

This page titled 9.2: Quasi-Experimental Designs with One Group is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **8.2: One-Group Designs** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

9.3: Quasi-Experimental Designs with Two or More Groups

Learning Objectives

1. Describe the different types of nonequivalent groups quasi-experimental designs.
2. Describe the strengths of the different types of nonequivalent groups quasi-experimental designs.

Recall that when participants in a between groups design experiment are randomly assigned to conditions, the resulting groups are likely to be quite similar (initially 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 nonequivalent. A nonequivalent groups design, then, is a design that compares groups of participants who have not been randomly assigned to conditions. There are several types of nonequivalent groups designs we will consider here.

Nonequivalent Group Designs

Natural Groups Designs (AKA Posttest-Only Nonequivalent Groups Design)

In the chapter on experimental designs, there was a quick sidetrack about natural groups designs to help you see the strengths.

Exercise 9.3.1

Without looking it up, try to describe what you can remember about natural groups designs.

Answer

The definition of natural groups design is "Research study that compares groups that were not randomly assigned into the IV conditions."

Natural groups designs is the posttest-only nonequivalent groups design. Just like in natural groups designs, there is a group who is exposed to a treatment or experiences an event and another (nonequivalent) group is not exposed to the treatment or event, and then the two groups are compared. Some researchers may argue that natural groups designs are only when the research is comparing participant variables, and posttest-only nonequivalent groups design is only for research in which a treatment was implemented, but this distinction is not needed for our purposes.

Let's discuss an example to reinforce the issues when there's neither equivelent groups (created through random assignment) nor is there a comparison of a pretest to a posttest. 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 posttest-only nonequivalent groups design because the students are not randomly assigned to classes by the researcher, and there was no pretest to test for equivalency between the classes. This means that 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. In addition to the teachers' styles and the students in the class, the classroom buildings might be very different and might cause different levels of achievement or motivation among the students. Maybe Ms. Williams has a smart board but Mr. Jones's classroom has a regular white board. 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 (or other!) confounding variables.

Thus, posttest-only nonequivalent groups designs are quasi-experimental designs because there is no initial equivalence between the groups; without true random assignment of the students to conditions, there remains the possibility that the DV was influenced by confounding variables that the researcher was not able to control. And without a pretest, its both impossible to know if the groups start out equivalent on the outcome, but it's also impossible to know if the scores on the outcome changed at all throughout the study.

Pretest-Posttest Nonequivalent Groups Design

Another way to improve upon the posttest-only nonequivalent groups design is to add a pretest. This is like the one-group pretest-posttest design, except now we do have a comparison group. In the pretest-posttest nonequivalent groups design, there is a treatment group that is given a pretest, receives a treatment, and then is given a posttest, and at the same time there is a nonequivalent 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.

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.

You might be thinking that there's no way that Hatzenbuehler et al. (2012) could also be an example here, but there was some additional groups in the study that haven't been mentioned yet. Not only did Hatzenbuehler et al. (2012) compare the same sexual minority men before and after same-gender marriage was legal in Massachusets, but Hatzenbuehler et al. (2012) also knew if these men were were long-term romantic relationship or not. This enabled a posttest-only nonequivalent groups design because partnered participants (one group) completed their health visits before (pretest) the legal recognition of same-gender marriage (treatment) and after (posttest), while nonpartnered men (nonequivalent group) also completed their health visits before (pretest) the legal recognition of same-gender marriage (treatment) and after (posttest). Hatzenbuehler et al. (2012) may not be a true example of a pretest-posttest nonequivalent groups design because the outcomes (dependent variables) are not directly related to the comparison groups or treatment like the example about the anti-drug program above, but it is interesting that one study can be similar to many different types of design, depending on how much detail you can see.

Interrupted Time-Series Design with Nonequivalent Groups

Just like the pretest-posttest nonequivalent groups design is an improvement on the one-group pretest-posttest design, the interrupted time series design with nonequivalent groups is an improvement on the the interrupted time-series design by to adding a comparison group. The interrupted time-series design with nonequivalent 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 nonequivalent 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 nonequivalent 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.

Pretest-Posttest Design With Switching Replication

Some of these nonequivalent control group designs can be further improved by adding a switching replication. Using a pretest-posttest design with switching replication design, nonequivalent groups are administered a pretest of the dependent variable, then one group receives a treatment while a nonequivalent 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.

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 nonequivalent 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.

A strength 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). Because all participants are not in all conditions, there is also not the concern over order effects like there is in a repeated measures designe.

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. 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. However, not all treatments can be easily removed, or the results of the treatments will not fade. This is an issue in repeated measures designs, as well in the switching replication with treatment removal design.

Good enough?

You may have noticed that each type of quasi-experimental design seemed a little better. Nonequivalent designs are better than one-group designs because there's a comparison group. Pretest-posttest designs are better than posttest-only designs because there's a comparison between before and after the treatment. Interrupted time series designs may be better with more measurements of the outcome. And replication, and then replication with removal, are better because there is strong evidence that nothing else other than the treatment could affect the the outcome. But, researchers still have this nagging feeling that maybe there is something else; is there another variable that affects both the treatment group and outcome, but not the comparison group (What about Z?). If you really want to show that the treatment *caused* changes in the outcome, then you really need initial equivalence and ongoing equivalence of the groups. We need to know that nothing could have affected the outcome other than our intervention. The last quasi-experimental designs do get at this, but the other quasi-experimental designs fall short of being able to show cause and effect.

References

Hatzenbuehler, M. L, O'Cleirigh, C., Grasso, C., Mayer, K., Safren, S., & Bradford, J. (2012). Effect of same-sex marriage laws on health care use and expenditures in sexual minority men: A quasi-natural experiment. *American Journal of Public Health, 102*(2), 285-291.

This page titled 9.3: Quasi-Experimental Designs with Two or More Groups is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **8.3: Non-Equivalent Groups Designs** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

9.4: Threats to Internal Validity

Learning Objectives

1. Identify the threats to internal validity associated with quasi-experimental designs.
2. Identify the threats to internal validity associated with experimental designs.

We've touched on the fact that quasi-experimental designs have issues with showing that the treatment, and nothing other than the treatment, could have caused changes in the outcome. This causal interpretation is about internal validity. As a reminder, internal validity of a study is the degree to which we can confidently infer a causal relationship between variables.

Researchers have come up with some of the most common alternative explanations that make it hard to make causal interpretations. These are called threats to validity. We will start by discussing threats to internal validity that are common in quasi-experimental designs.

Threats to Validity in Quasi-Experimental Designs

In Table 9.4.1, you can see descriptions of some common threats to internal validity in quasi-experimental designs. The right-most column will also describe what this threat to internal validity could look like in a quasi-experimental study comparing classes that use OER (openly-licensed educational resources, like the textbook you are reading) to classes that do not use OER, with the outcome being student learning. For these examples, the study is expected to be a pretest-posttest nonequivalent groups design, with the measure of student learning be chapter quizzes.

Table 9.4.1: Common Threats to Internal Validity for Quasi-Experimental Designs

Name	Description	Example
History	The treatment group experiences events (other than the IV) that the control group does not experience.	Many instructors started using OER during the pandemic because it was easier for students to access than commercial textbooks. However, if we compared this group of students to students in previous semester, then we would have the threat of history since the previous students (comparison group) did not have to attend school during a pandemic.
(Re-)Testing	Repeated exposure to DV. [This threat to internal validity is more of a concern when there is no comparison group. When there is a comparison group, we could see that both groups improve, but test to see if the intervention group improved more.]	Students may improve their test-taking through the practice of weekly chapter quizzes.
Instrumentation	When the instrument changes through use. This includes observer bias.	Many instructors use quizzes developed by the publisher, so the DV would automatically change for those instructors any time a different textbook is use. This still could be a threat for instructors who develop their own quizzes if they change some questions in their quiz between the pretest and posttest.
Regression to the Mean	Just due to chance, extreme scores tend to regress towards the mean in later measurements.	Some students, some sections, some semesters just go better than others. If, by chance, a group of students, a section or a commercial textbook, or a good semester was in the comparison group on the pretest, then their inordinately high scores could decrease (regress) to more typical levels (the mean) on the posttest. This would make it look like the commercial textbook hurt student learning, even if it was an artifact of statistics and probability.
Attrition/ Mortality	Participants drop-out (further reducing initial equivalence).	Because of the high cost of textbooks, it's more likely that students who have to buy a textbook, compared to students in an OER section, would drop out of the class because they need to work more.
Selection	There are relevant differences between the groups before the experiment begins.	Alternatively, an instructor may choose to use the OER in thei online section and a traditional textbook in their face-to-face section. However, there are lots of reasons why students would take an online course (or take a face-to-face course) that could also affect their scores on the chapter quizzes. Online students tend to work more

Name	Description	Example
		hours or have more family responsibilities than students in on-campus sections. These work and family responsibilities may limit how much time the student can study for the course, so poor scores on the chapter quizzes would be due to time spent on studying rather than the type of textbook. Also, the example above comparing students from before the pandemic to students after the pandemic is a good example of this.

One less common threat is maturation. During longer studies, like research that lasts a years rather than months, participants may naturally change between the pretest and the posttest in ways that they were going to anyway because they are growing and learning. If it were a year long course program, participants might learn how to study better or inefficient it is to try to multitasking. It could be that natural development that leads to better scores, not the intervention or comparison. Like testing (sometimes called re-testing), this threat to internal validity applies more to posttest only nonequivalent groups designs (also called natural groups designs).

A threat to internal validity related to regression to the mean 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.

What do you think?

Do you think that spontaneous remission could happen in our study on student learning comparing different types of textbooks?

Threats to Internal Validity in Between Groups Designs

With quasi-experimental designs, we know that something else may have caused the effect because we didn't have initial equivalence between the groups (or may have only had one group). However, between groups experimental designs also have threats to validity. Quasi-Experimental designs share the threats to internal validity shown in Table **9.4.2** with between groups experimental designs. Again, the study in these examples a pretest-posttest nonequivalent groups design between sections that use an OER to sections that use a commercial textbook. The measure of student learning is still chapter quizzes.

Table **9.4.2**: Common Threats to Internal Validity for Experimental and Quasi-Experimental Designs

Name	Description	Example
Contamination (3 similar kinds)	• Rivalry: Control group is creative or motivated to out-perform treatment group. • Diffusion of Treatment: Control group affected by the treatment. • Equalization: Control group receives similar treatment through alternate means. Note that in all three types, the control group is doing something to overcome their disadvantage (compared to the treatment group).	• Rivalry: The group with the traditional textbook might learn about the group with a free, online textbook. Thinking that it is easier to access the course materials, the traditional textbook group may put more reminders to bring their textbook with them to class and study session. • Diffusion of Treatment: The group with the free, online acces to their textbook may share the link with their friends in a section of the course that only had the traditional textbook. • Equalization: The group with the traditional textbook might search online for learning materials for the course, or they may spend more time on the learning materials posted in their course shell.
Experimenter Expectancy Effects	Experimenter's beliefs may affect their behavior towards the participants, or their measurements. This can also include observer bias.	In this case, the experimenter is probably the instructor so they would expect higher scores for the OER group. They may unconsciously treat the students in the OER section better, or grade their quizzes with more generosity.
Novelty Effects	New activities or additional attention (Hawthorne effect) can change participants' behaviors.	Students who haven't enrolled in a class with an OER might find having a free, online textbook very novel and interesting, which may lead to them spending more time reviewing the textbook. This could affect quiz scores, even if nothing else about having OER affect the students' learning.

Thinking back to Hatzenbuehler et al. (2012), what threats to internal validity might be relevant?

Why do you think?

What threats to internal validity for quasi-experimental designs seem to apply to the comparison of medical care visits, medical care costs, mental health care visits, and mental health care costs for gay and bisexual men in Massachusets between a year before Massachusets became the first state to legally recognize same-gender marriage (Hatzenbuehler et al., 2012) to a year after same-gender marriage was legally recognized in the state? Why those?

- History
- (Re-)Testing
- Instrumentation
- Regression to the Mean
- Attrition/ Mortality
- Selection
- Maturation

- Spontaneous Remission

What threats to internal validity for any experimental design seem to apply to Hatzenbuehler et al. (2012)? Why those?

- Contamination: Rivalry
- Contamination: Diffusion of Treatment
- Contamination: Equalization
- Experimenter Expectancy Effects
- Novelty Effects

References

Hatzenbuehler, M. L, O'Cleirigh, C., Grasso, C., Mayer, K., Safren, S., & Bradford, J. (2012). Effect of same-sex marriage laws on health care use and expenditures in sexual minority men: A quasi-natural experiment. *American Journal of Public Health*, *102*(2), 285-291.

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.

This page titled 9.4: Threats to Internal Validity is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **8.2: One-Group Designs** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

9.5: Summary of Quasi-Experimental Research

Key Takeaways

- Quasi-experimental research involves the manipulation of an independent variable without the random assignment of participants to conditions or counterbalancing of orders of conditions.
- There are quasi-experimental designs that compare the scores of participants before and after the treatment without a comparison group; the comparisons are between the timeframes.
- There are quasi-experimental designs that compare the scores of participants before and after the treatment and a (nonequivalent) comparison group.
- Quasi-experimental research does not provide initial equivalence between the groups because it does not involve random assignment to conditions or counterbalancing. For this reason, quasi-experimental research has lower internal validity than experimental designs. This means that experiments are stronger evidence that the IV *caused* changes in the DV than quasi-experimental designs. However, the more that alternative explanations of changes to the outcome variable (DV) other than the intervention group, the more higher the internal validity. Of all of the quasi-experimental designs, those that include a switching replication are highest in internal validity.
- There are typical issues in quasi-experimental designs that could affect internal validity. These are called threats to internal validity. Additional threats to internal validity can affect experimental designs, as well.

What's Next?

Now that you're a little more comfortable with experimental designs and quasi-experimental designs, we're going to learn about slightly more complex designs as they include more than one IV (not just more than one group).

Exercises

- Practice: Imagine that two professors decide to test the effect of giving daily quizzes on student performance in a statistics course. They decide that Professor A will give quizzes but Professor B will not. They will then compare the performance of students in their two sections on a common final exam. List five other variables that might differ between the two sections that could affect the results.
- Discussion: Imagine that a group of obese children is recruited for a study in which their weight is measured, then they participate for 3 months in a program that encourages them to be more active, and finally their weight is measured again.
  - What kind of quasi-experimental design would this be?
  - For each of the following threats to internal validity for quasi-experimental designs, explain how each of the following might affect the results (or not):
    - History
    - (Re-)Testing
    - Instrumentation
    - Regression to the Mean
    - Attrition/ Mortality
    - Selection
    - Maturation
    - Spontaneous Remission
  - For each of the following threats to internal validity for all designs, explain how each of the following might affect the results (or not):
    - Contamination: Rivalry
    - Contamination: Diffusion of Treatment
    - Contamination: Equalization
    - Experimenter Expectancy Effects
    - Novelty Effects

- Practice: Some schools have a lab section required for their research methods class. To see if the lab portion of this class improves scores on the class research project's manuscript, an instructor could randomly assign half of the class to stay for the lab section, and the other half of the class could be banned from the lab section of this course. In this scenario, which threats to internal validity might be present? Why?
  - Contamination: Rivalry
  - Contamination: Diffusion of Treatment
  - Contamination: Equalization
  - Experimenter Expectancy Effects
  - Novelty Effects

This page titled 9.5: Summary of Quasi-Experimental Research is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **8.4: Quasi-Experimental Research (Summary)** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

10: Factorial Designs

10.1: Prelude to Factorial Designs

10.2: Setting Up a Factorial Experiment

10.3: Interpreting the Results of a Factorial Experiment

10.3.1: Practice Interpreting Main Effects and Interactions

10.3.2: Interpreting Main Effects and Interactions in Graphs

10.3.3: Interpreting Beyond 2x2 in Graphs

10.4: Interpreting Interactions- Do Main Effects Matter?

10.5: Summary of Factorial Designs

10: Factorial Designs is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

10.1: Prelude to Factorial Designs

Have you heard the the phrase "cleanliness is next to godliness"? Schnall et al. (2008) wanted to test if judgments of morality differed by the cleanliness of the room in which the judgments were made. In an experiment, Schnall et al. (2008) manipulated participants' feelings of disgust by testing them in either a clean room or a messy room that contained dirty dishes, an overflowing wastebasket, and a chewed-up pen. The cleanliness of the room was their IV. Schnall et al. (2008) also used a self-report questionnaire to measure the amount of attention that people pay to their own bodily sensations, which Schnall et al. (2008) called this "private body consciousness." Because Schnall et al. (2008) thought that private body consciousness would affect their DV, private body consciousness was another IV. Schnall et al. (2008) measured their primary dependent variable, the harshness of people's moral judgments, by describing different behaviors (e.g., eating one's dead dog, failing to return a found wallet) and having participants rate the moral acceptability of each one on a scale of 1 to 7. As a manipulation check, Schnall et al. (2008) also asked participants to rate their current level of disgust and other emotions. The primary results of Schnall et al. (2008) was that participants in the messy room were, in fact, more disgusted and made harsher moral judgments than participants in the clean room —but only if they scored relatively high in private body consciousness.

Schnall et al. (2008) is an example of a factorial design, a study with multiple IVs that are systematically combined to see how each combination affects the DV.

References

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin, 34*, 1096–1109.

10.1: Prelude to Factorial Designs is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **9.1: Factorial Designs** has no license indicated.

10.2: Setting Up a Factorial Experiment

Learning Objectives

1. Explain why researchers often include multiple independent variables in their studies.
2. Define factorial design, and use a factorial design table to represent and interpret simple 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. Schnall and her colleagues studied the effect of both disgust and private body consciousness in the same study. 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. For example, instead of conducting one study on the effect of disgust on moral judgment and another on the effect of private body consciousness on moral judgment, Schnall and colleagues were able to conduct one study that addressed both questions. 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. Schnall and her colleagues, for example, observed an interaction between disgust and private body consciousness because the effect of disgust depended on whether participants were high or low in private body consciousness. As we will see, interactions are often among the most interesting results in psychological research.

Factorial Designs

Overview

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 **10.2.1**. 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.

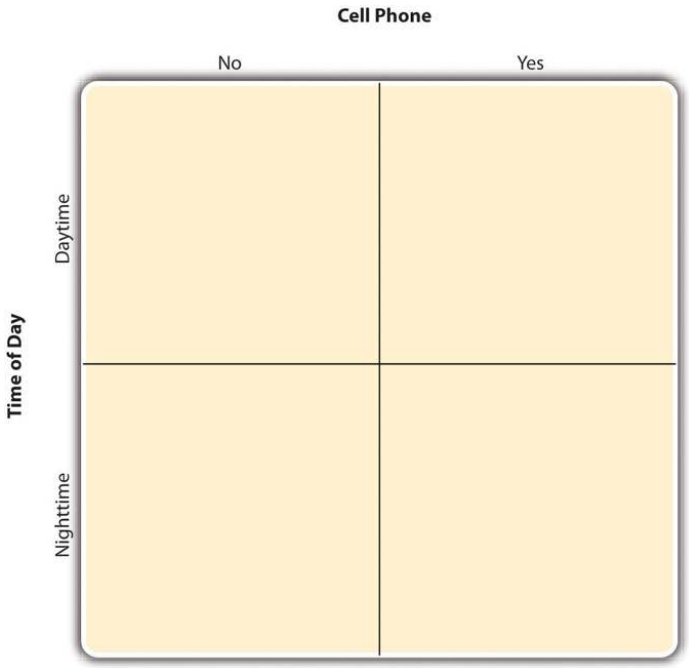


Figure **10.2.1**: Factorial Design Table Representing a 2 × 2 Factorial Design

In principle, factorial designs can include any number of independent variables with any number of levels. For example, an experiment could include the type of psychotherapy (cognitive vs. behavioral), the length of the psychotherapy (2 weeks vs. 2 months), and the sex of the psychotherapist (female vs. male). This would be a 2 × 2 × 2 factorial design and would have eight conditions. Figure **10.2.2** 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 × 2 × 2 × 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 more information, see the discussion about the importance of adequate statistical power in Chapter 13). As a result, in the remainder of this section, we will focus on designs with two independent variables. The general principles discussed here extend in a straightforward way to more complex factorial designs.

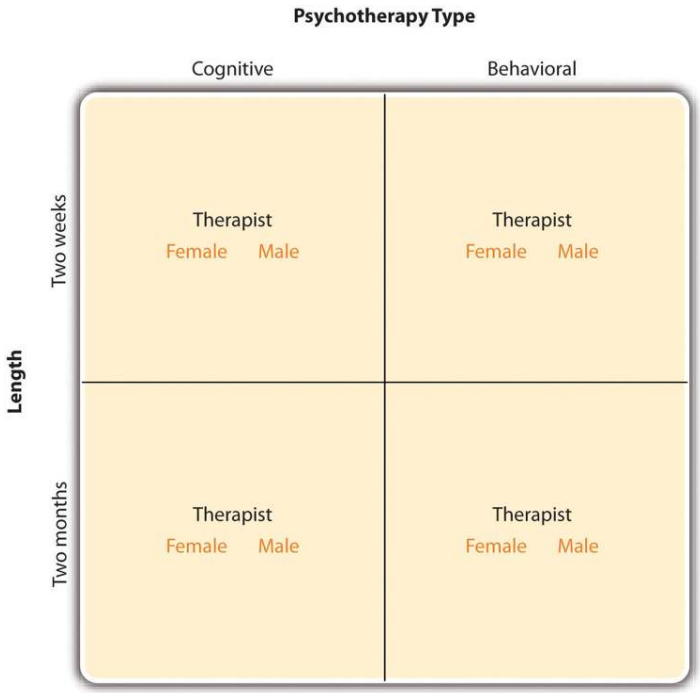


Figure **10.2.2**: Factorial Design Table Representing a 2 × 2 × 2 Factorial Design

Assigning Participants to Conditions

Recall that 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 advantages and disadvantages of these two approaches are the same as those discussed in Chapter 5. 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 and controls extraneous participant variables.

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.

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. The study by Schnall and colleagues is a good example. One independent variable was disgust, which the researchers manipulated by testing participants in a clean room or a messy room. The other was private body consciousness, a participant variable which the researchers simply measured. Another 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, Kosslyn, Delamater, Fama, & Barsky, 1999) [1]. The manipulated independent variable was the type of word. Some were negative health-related words (e.g., *tumor, coronary*),

and others were not health related (e.g., *election*, *geometry*). The non-manipulated independent variable was whether participants were high or low in hypochondriasis (excessive concern with ordinary bodily symptoms). The result of this study was that the participants high in hypochondriasis were better than those low in hypochondriasis at recalling the health-related words, but they were no better at recalling the non-health-related words.

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, hypochondriasis, self-esteem, gender, and so on), and as such, they are by definition between-subjects factors. For example, people are either low in hypochondriasis or high in hypochondriasis; 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. For example, Schnall and her colleagues were justified in concluding that disgust affected the harshness of their participants' moral judgments because they manipulated that variable and randomly assigned participants to the clean or messy room. But they would not have been justified in concluding that participants' private body consciousness affected the harshness of their participants' moral judgments because they did not manipulate that variable. It could be, for example, that having a strict moral code and a heightened awareness of one's body are both caused by some third variable (e.g., neuroticism). Thus it is important to be aware of which variables in a study are manipulated and which are not.

Non-Experimental Studies With Factorial Designs

Thus far we have seen that factorial experiments can include manipulated independent variables or a combination of manipulated and non-manipulated independent variables. But factorial designs can also include *only* non-manipulated independent variables, in which case they are no longer experiments but are instead non-experimental in nature. Consider a hypothetical study in which a researcher simply measures both the moods and the self-esteem of several participants—categorizing them as having either a positive or negative mood and as being either high or low in self-esteem—along with their willingness to have unprotected sexual intercourse. This can be conceptualized as a 2 × 2 factorial design with mood (positive vs. negative) and self-esteem (high vs. low) as non-manipulated between-subjects factors. Willingness to have unprotected sex is the dependent variable.

Again, because neither independent variable in this example was manipulated, it is a non-experimental study rather than an experiment. (The similar study by MacDonald and Martineau [2002][2] was an experiment because they manipulated their participants' moods.) 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.

References

1. 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*, 67–78. ↵
2. 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. ↵

This page titled 10.2: Setting Up a Factorial Experiment is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **9.2: Setting Up a Factorial Experiment** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

10.3: Interpreting the Results of a Factorial Experiment

Learning Objectives

1. Distinguish between main effects and interactions, and recognize and give examples of each.
2. Sketch and interpret bar graphs and line graphs showing the results of studies with simple factorial designs.
3. Distinguish between main effects and simple effects, and recognize when an analysis of simple effects is required.

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 **10.3.1** 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 **10.3.1** 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.

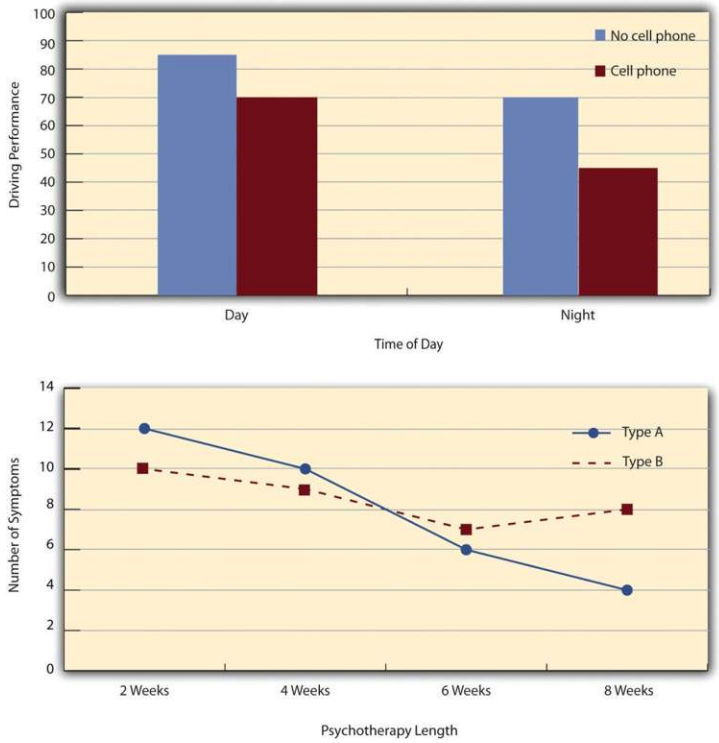


Figure **10.3.1**: Two Ways to Plot the Results of a Factorial Experiment With Two Independent Variables

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 **10.3.1** 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 **10.3.1**, for example, shows a clear main effect of psychotherapy length. The longer the psychotherapy, the better it worked.

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 her 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 she is bringing with her then there is an interaction. For instance, if you will go to see the cheesy romantic comedy if she brings her hot friend you want to get to know better, but you will not go to this movie if she brings anyone else, then there is an interaction. Drug interactions are another good illustration of everyday interactions. Many older men take Viagara to assist them in the bedroom, and many men take nitrates to treat angina or chest pain. So each of these drugs is beneficial on its own (there are main effects of each on older men's well-being). But the combination of these two drugs can be lethal. In other words, there is a very important interaction between Viagara and heart medication that older men need to be aware of to prevent their untimely demise.

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). Schnall and her colleagues also demonstrated an interaction because the effect of whether the room was clean or messy on participants' moral judgments depended on whether the participants were low or high in private body consciousness. If they were high in private body consciousness, then those in the messy room made harsher judgments. If they were low in private body consciousness, then whether the room was clean or messy did not matter.

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 with hypochondriasis are especially attentive to any negative health-related information. This led to the hypothesis that people high in hypochondriasis would recall negative health-related words more accurately than people low in hypochondriasis but recall non-health-related words about the same as people low in hypochondriasis. And of course, this is exactly what happened in this study.

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 **10.3.2**. 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). This is much like the study of Schnall and her colleagues where there was an effect of disgust for those high in private body consciousness but not for those low in private body consciousness. 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 **10.3.2**, 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)[1]. 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.

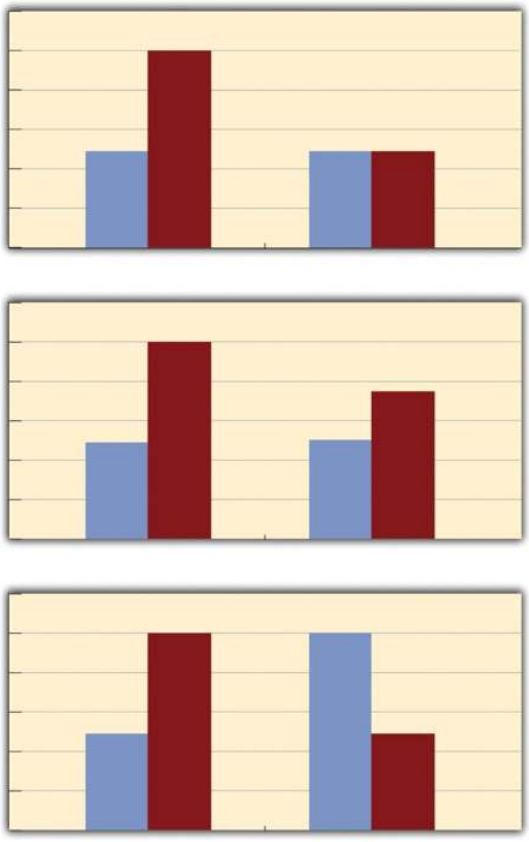


Figure **10.3.2**: Bar Graphs Showing Three Types of Interactions. In the top panel, one independent variable has an effect at one level of the second independent variable but not at the other. In the middle panel, one independent variable has a stronger effect at one level of the second independent variable than at the other. In the bottom panel, one independent variable has the opposite effect at one level of the second independent variable than at the other.

Figure **10.3.3** 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).

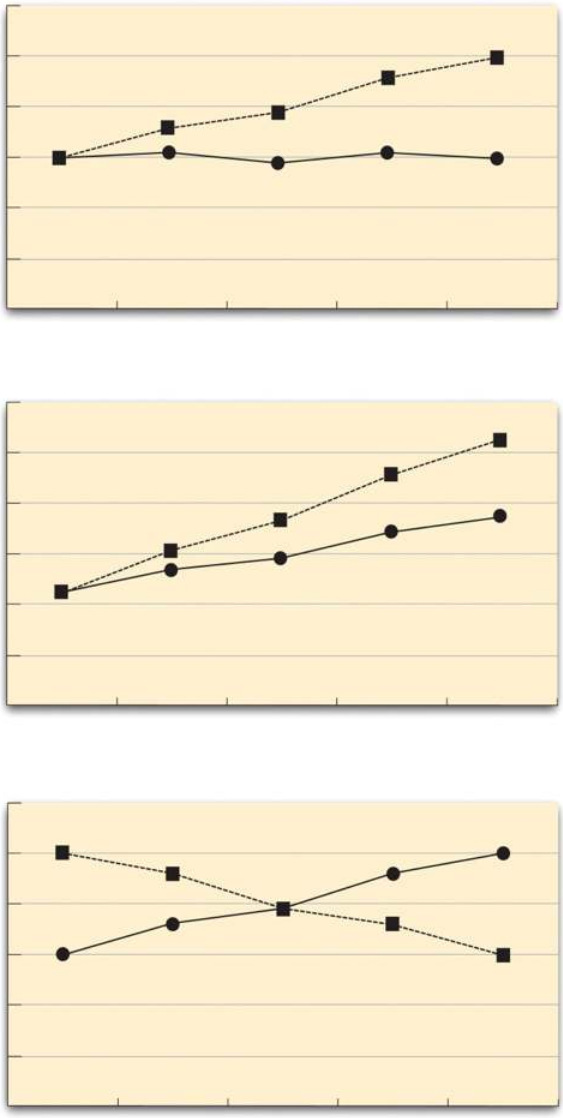


Figure **10.3.3**: Line Graphs Showing Different Types of Interactions. In the top panel, one independent variable has an effect at one level of the second independent variable but not at the other. In the middle panel, one independent variable has a stronger effect at one level of the second independent variable than at the other. In the bottom panel, one independent variable has the opposite effect at one level of the second independent variable than at the other.

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 test performance 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 memory test performance. 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 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 test performance 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. As we saw previously, 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 and four simple effects.

Schnall and colleagues found a main effect of disgust on moral judgments (those in a messy room made harsher moral judgments). However, they also discovered an interaction between private body consciousness and disgust. In other words, the effect of disgust depended on private body consciousness. The presence of this interaction suggests the main effect may be a bit misleading. That is, it is not entirely accurate to say that those in a messy room made harsher moral judgments because this was only true for half of the participants. Using simple effects analyses, they were able to further demonstrate that for people high in private body consciousness, there was an effect of disgust on moral judgments. Further, they found that for those low in private body consciousness there was no effect of disgust on moral judgments. By examining the effect of disgust at each level of body consciousness using simple effects analyses, Schnall and colleagues were able to better understand the nature of the interaction.

As described previously, Brown and colleagues found an interaction between type of words (health related or not health related) and hypochondriasis (high or low) on word recall. To break down this interaction using simple effects analyses they examined the effect of hypochondriasis at each level of word type. Specifically, they examined the effect of hypochondriasis on recall of health-related words and then they subsequently examined the effect of hypochondriasis on recall of non-health related words. They found that people high in hypochondriasis were able to recall more health-related words than people low in hypochondriasis. In contrast, there was no effect of hypochondriasis on the recall of non-health related words.

Once again examining simple effects provides a means of breaking down the interaction and therefore it is only necessary to conduct these analyses when an interaction is present. When there is no interaction then the main effects will tell the complete and accurate story. 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 and 4 simple effects. A researcher using a 2×3 design with six conditions would need to look at 2 main effects and 5 simple effects, while a researcher using a 3×3 design with nine conditions would need to look at 2 main effects and 6 simple effects. As you can see, while the number of main effects depends simply on the number of independent variables included (one main effect can be explored for each independent variable), 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).

References

1. Gilliland, K. (1980). The interactive effect of introversion-extraversion with caffeine induced arousal on verbal performance. *Journal of Research in Personality, 14*, 482–492. ↵

This page titled 10.3: Interpreting the Results of a Factorial Experiment is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **9.3: Interpreting the Results of a Factorial Experiment** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

10.3.1: Practice Interpreting Main Effects and Interactions

Dr. MO's student researchers did a couple studies on growth mindset, which is the idea that mistakes help you learn and that more time and effort leads to brain development. One group of student researchers collected data on improving growth mindset that should be analyzed with a factorial ANOVA. Here it is:

Scenario

The student researchers looked at the Difference scores of 106 students at Dr. MO's community college. The Difference scores were calculated by subtracting each students' Post-Test Mindset Quiz score (measured at the end of the semester) from their own Pre-Test Mindset Quiz score (measured at the beginning of the same semester); thus, positive scores mean that the student improved their mindset. In this study, we had two variables. One was the intervention, which had two levels: No Intervention (comparison control group) and Intervention (in which faculty tried different activities to improve mindset). We also collected information on what department these activities were happening in, and had enough data to analyze students in English classes and students in Psychology classes.

Can you identify the groups and variables in this scenario?

Example 10.3.1.1

Answer the following questions to understand the variables and groups that we are working with.

1. Who is the sample?
2. Who do might be the population?
3. What are the IVs and levels for each IV?
4. What is the DV (quantitative variable being measured)?
5. Is this a 2x2 factorial design? If not, what kind of design is it?
6. List out each of the *combinations* of the levels of the IVs:

Solution

1. The sample was 106 community college students taking an English class or a Psychology class.
2. The population could be all community colleges students, or maybe all community college students taking English or Psychology. Or maybe all community college students in a general education course?
3. One IV is Intervention, with the levels being Yes or No. The other IV is Department, with the levels being English or Psychology.
4. The DV is the Difference score from the Mindset Quiz pre-test to the Mindset Quiz post-test.
5. Yes, this is a 2x2 factorial design because there are two IVs (so there are two numbers __x__); the first IV (Intervention) has two levels and the second IV (Department) has two levels.
6. List out each of the *combinations* of the levels of the IVs:
    1. Intervention in English
    2. Intervention in Psychology
    3. No Intervention in English
    4. No Intervention in Psychology

Can you plug these IVs into a factorial design table?

Example 10.3.1.2

Complete a table showing the factorial design IVs and DVS.

Solution

Table **10.3.1.1**- 2x2 Factorial Grid of Intervention by Department

IV Levels	**IV1- Yes Intervention**	**IV1- No Intervention**
IV2: English Department	2x2: Students who experienced an intervention in an English class.	2x2: Students who did not experience an intervention but who were in an English class.
IV2: Psychology Department	2x2: Students who experienced an intervention in a Psychology class.	2x2: Students who did not experience an intervention but who were in a Psychology class.

You could have put Department as IV1 (columns) and Intervention as IV2 (rows). It's really about what makes the most sense to you; Dr. MO wanted what she thinks is the main influence of mindset on the top.

Participants in each "cell" of this design have a unique combination of IV conditions.

Three Effects

With a 2x2 factorial design, you have three effects to look at. Remember, "effects" are the results of the DV, what was measured. Here are the three effects that you need to look at:

1. The main effect of the one IV: How does one IV affect the DV (independent of the other IV)
2. The main effect of the other IV: How does the other IV affect the DV (independent of the first IV)
3. The interaction of the two IVs -- how they **jointly** affect the DV

Example 10.3.1.3

In our mindset scenario, what are these three effects?

Solution

1. The main effect of the intervention; did the intervention improve Mindset Quiz scores?
2. The main effect of the department; did which class you were in affect Mindset Quiz scores?
3. The interaction of the intervention by department; is one department more likely to improve Mindset Quiz scores than another department when there's an intervention? Another way to think about these variables is to ask whether the departments started out with different average Mindset Quiz scores, so that even in the No Intervention condition the Difference score would be statistically significantly different between the two departments.

Let's look at these main effects in Table **10.3.1.2**, in which the marginal means were included. Marginal means are, you guessed, it the means on the margins of the table. These means on the margin show the means for each level of each IV, which are the *main effects*. The marginal means do not show the combination of the IVs' levels, so they do not show an interaction.

Table **10.3.1.2**- 2x2 Factorial Grid of Intervention by Department with Marginal Means

IV Levels	**IV1- Yes Intervention**	**IV1- No Intervention**	Marginal Means of IV2
IV2: English Department	2x2: Students who experienced an intervention in an English class.	2x2: Students who did not experience an intervention but who were in an English class.	**DV = 2.49**
IV2: Psychology Department	2x2: Students who experienced an intervention in a Psychology class.	2x2: Students who did not experience an intervention but who were in a Psychology class.	**DV = 1.34**
Marginal Means of IV1	**DV = 2.08**	**DV = 1.76**	

Again, you could have put Department as IV1 (columns) and Intervention as IV2 (rows). And again, participants in each "cell" of this design have a unique combination of IV conditions.

Main Effects

To identify main effects of any IV, you compare the means for just that variable. These are called marginal means because they are always located at the outside (margins) of the factorial design talbe. As shown in Table **10.3.1.2**, the means for the intervention IV is 2.08 for those who received the intervention and 1.76 for the group who did not receive the intervention. We won't look at statistical significant here, but it *looks* like the students who experienced an Intervention had a larger Difference score than students who did not have an intervention. This is what was expected; students who experienced an intervention had a higher Mindset Quiz score at the end of the semester (post-test), after the interventions, than at the beginning of the semester (pretest). This is your *main effect* of Intervention. Now you can practice on the other independent variable.

? Exercise 10.3.1.1

Based on Table **10.3.1.2**, what are the marginal means for department?

Answer

Marginal Means:

- English=2.49
- Psychology=1.34

Which department seemed to have a higher Difference score? There seems to be a main effect of department such that students in English had a higher Difference score than students in Psychology classes. We can't be say much more than that without looking at actual statistical results. Instead, we will look at the individual cells of our grid to see if there was an *interaction* between Department and Intervention on the Difference scores.

Interaction

Table **10.3.1.3** has the means for each combination of each IV level in the individual cells. Looking at Table **10.3.1.3**, how do the inside cells (simple effects) seem relate to each other?

Table **10.3.1.3**- 2x2 Factorial Grid of Intervention by Department with Marginal Means & Cell Means

IV Levels	**IV1- Yes Intervention**	**IV1- No Intervention**	Marginal Means of IV2
IV2: English Department	**DV=3.70**	**DV=1.29**	**DV=2.49**
IV2: Psychology Department	DV=0.46	DV=2.23	**DV=1.34**
Marginal Means of IV1	**DV=2.08**	**DV=1.76**	

Again, you could have put Department as IV1 (columns) and Intervention as IV2 (rows). And again, participants in each "cell" of this design have a unique combination of IV conditions.

✔ Example 10.3.1.5

Answer the following questions related the cell means in Table **10.3.1.3**,

1. Was one cell substantially higher than the others?
2. Was one cell substantially lower than the others?

Solution

1. It looks like students in English class who experienced an intervention had higher Difference scores.
2. It looks like students in a Psychology class who experienced an intervention had a really low difference score.

This pattern of cell means is your *interaction* of Intervention and Department: Experiencing the intervention *interacts* with the department to affect Differences in Mindset Quiz scores such that the intervention seems to improve mindset for those in English but not in Psychology (the Psychology students who experienced an intervention were essentially unchanged). What's strange is

that Psychology students in the control condition (No Intervention) might have actually improved their Mindset Quiz scores more than both English students with no intervention and Psychology student with the intervention! You might be wondering why these strange effects for the Psychology students, and Dr. MO has no answer for you. This is real data, and there's no good explanation for this with the variables that we have. ::shrug::

As you can see, the main effects of each IV can relate to the interaction in several different ways. Let's look at that next.

This page titled 10.3.1: Practice Interpreting Main Effects and Interactions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.3.2: Interpreting Main Effects and Interactions in Graphs

Designs with multiple factors are very common. When you read a research article (and you will!), you will often see tables and graphs that show the results from designs with multiple factors. It would be good for you if you were comfortable interpreting the meaning of those results. You should look at the main effects, but understand that any statistically significant interaction is the real result because the interaction tells you more about the levels of each IV than any main effect can.

2x2 designs

Let's take the case of 2x2 designs. There will always be the possibility of two main effects and one interaction. You will always be able to compare the means for each main effect and interaction. If the two means from one variable are different, then there is a main effect. If the two means from the other variable are different, then there is a main effect. And then you look at the cell means (the means in the four combinations of the two IV's two levels), and those could have all sorts of relationships. There a bunch of ways all of this can turn out. Check out the ways, there are eight of them:

1. no IV1 main effect, no IV2 main effect, no interaction (This is the null hypothesis!)
2. IV1 main effect, no IV2 main effect, no interaction
3. IV1 main effect, no IV2 main effect, yes interaction
4. IV1 main effect, IV2 main effect, no interaction
5. IV1 main effect, IV2 main effect, yes interaction
6. no IV1 main effect, IV2 main effect, no interaction
7. no IV1 main effect, IV2 main effect, yes interaction
8. no IV1 main effect, no IV2 main effect, yes interaction

OK, so if you run a 2x2, any of these eight general patterns could occur in your data. That's a lot to keep track of. As you develop your skills in examining graphs that plot means, you should be able to look at the graph and visually guesstimate if there is, or is not, a main effect or interaction. You will need you inferential statistics to tell you for sure, but it is worth knowing how to know see the patterns. Look at the eight patterns in Figure **10.3.2.1**.

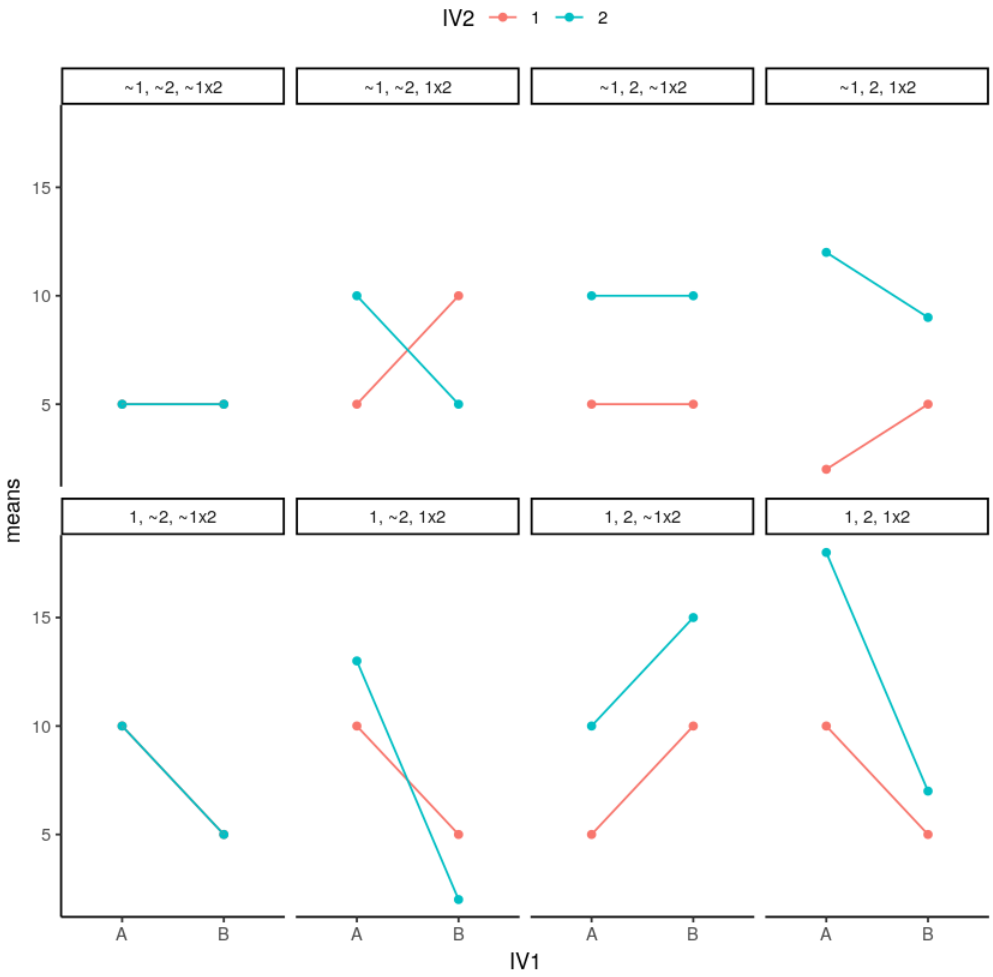


Figure **10.3.2.1**: Line graphs showing 8 possible general outcomes for a 2x2 design. (CC-BY-SA Matthew J. C. Crump from 10.1 of Answering Questions with Data)

Whenever the lines cross, or would cross if they kept going, you have a possibility of an interaction. Whenever the lines are parallel, there *can't* be an interaction. When *both* of the points on the A side are higher *or* lower than *both* of the points on the B side, then you have a main effect for IV1 (A vs B). Whenever the green line is above *or* below the red line, then you have a main effect for IV2 (1 vs. 2).

We know this is complicated. You should see what all the possibilities look like when we start adding more levels or more IVs! It gets nuts. Because of this nuttiness, it is often good practice to make your research designs simple (as few IVs and levels as possible to test your question). That way it will be easier to interpret your data. Whenever you see that someone ran a 4x3x7x2 design, your head should spin. It's just too complicated.

Do you remember Stephen Chew's video series on how to learn? Well, his second video (What Students Should Know about How People Learn) has a factorial design (with an extra control group).

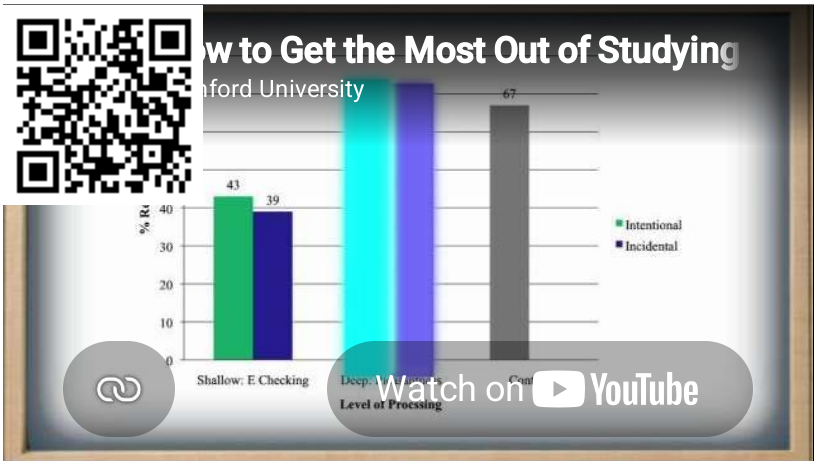


Figure 10.3.2.2: What Students Should Know about How People Learn. (Stephen Chew via YouTube)

Watching this video can help you understand factorial designs, including main effects and interactions), as well as learn about learning!

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)
- Dr. Stephen Chew's "What Students Should Know About How People Learn" (Part 2 of "How to Get the Most Out of Studying" series)

This page titled 10.3.2: Interpreting Main Effects and Interactions in Graphs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **10.1: Looking at main effects and interactions** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

10.3.3: Interpreting Beyond 2x2 in Graphs

Our graphs so far have focused on the simplest case for factorial designs, the 2x2 design, with two IVs, each with 2 levels. It is worth spending some time looking at a few more complicated designs and how to interpret them. We talked about more complicated designs in the Factorial Notations and Square Tables section, but here's a more focused approach to interpreting the graphs of these advanced designs.

2x3 design

In a 2x3 design there are two IVs. IV1 has two levels, and IV2 has three levels. Typically, there would be one DV. Let's talk about the main effects and interaction for this design.

First, let's make the design concrete. Let's imagine we are running a memory experiment. We give people some words to remember, and then test them to see how many they can correctly remember. Our DV is the proportion (percentage) that participants remembered correctly out of all tries. We know that people forget things over time. Our first IV will be time of test, immediate versus one week later. The time of test IV will produce a forgetting effect. Generally, people will have a higher proportion correct on an immediate test of their memory for things they just saw, compared to testing a week later.

We might be interested in manipulations that reduce the amount of forgetting that happens over the week. Let's make the second IV the number of time people got to study the items before the memory test, once, twice or three times. We call IV2 the repetition manipulation.

We might expect data that looks like Figure **10.3.3.1**. The top line shows the means when there is no delay (Immediate) for the three levels of repetition. Notice that the proportion correct (y-axis) increases for the Immediate group with each repetition. The bottom ling shows the One Week Delay group over the three levels of repetition. Again, more repetition seems to increase the proportion correct. Let's talk about this graph in terms of main effects and interaction.

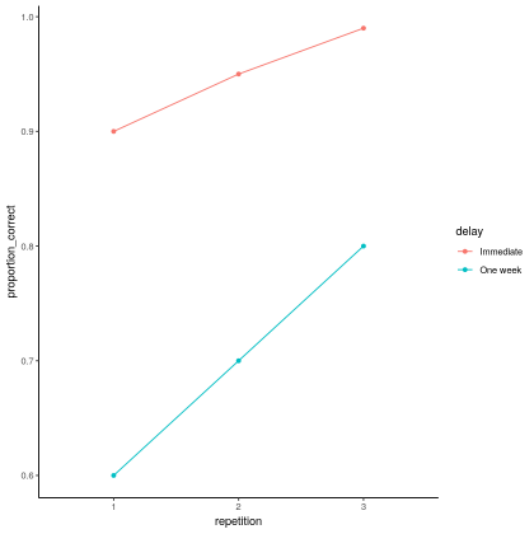


Figure **10.3.3.1**: Example means for a 2x3 factorial design. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

First, the main effect of delay (time of test) is shown by in each differently-colored line, and seems obvious; the red line is on the top, way above the aqua line. This tells us that the proportion correct on the memory test is always higher when the memory test is taken immediately compared to after one week.

Second, the main effect of repetition is presented on the x-axis, and seems to be clearly present. The more times people saw the items in the memory test (once, twice, or three times), the more they remembered, as measured by increasingly higher proportion correct as a function of number of repetitions.

Is there an interaction? Yes, there is. Remember, an interaction occurs when the effect of one IV depends on the levels of an another. The delay IV measures the forgetting effect. Does the size of the forgetting effet change across the levels of the repetition variable? Yes it does. With one repetition the forgetting effect is 0.9 - 0.6 = 0.4. With two repetitions, the forgetting effect is a little

bit smaller, and with three, the repetition is even smaller still. So, the size of the forgetting effect changes as a function of the levels of the repetition IV. There is evidence in the means for an interaction. You would have to conduct an inferential test on the interaction term to see if these differences were likely or unlikely to be due to sampling error.

If there was no interaction, and say, no main effect of repetition, we would see something like Figure **10.3.3.2**.

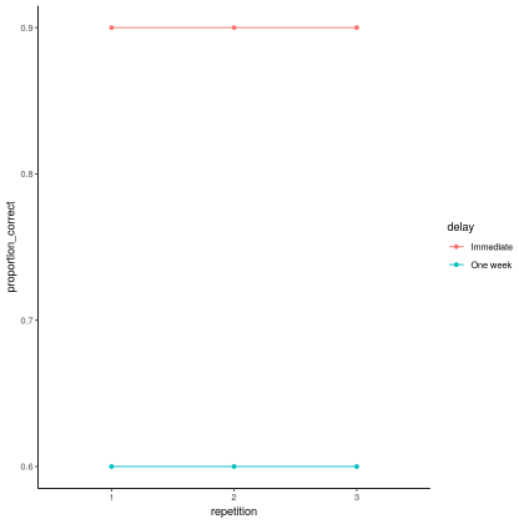


Figure **10.3.3.2**: Example means for a 2x3 design when there is only one main effect. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

What would you say about the interaction if you saw something like Figure **10.3.3.3**? The Immediate group is high, but repetition doesn't seem to matter. The One Week Delay group is flat until the third repetition, then increases the proportion correct.

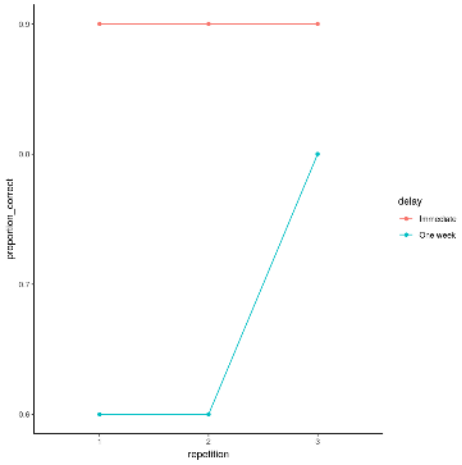


Figure **10.3.3.3**: Example means for a 2x3 design showing another pattern that produces an interaction. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

There is evidence in the means for an interaction. Remember, we are measuring the forgetting effect (effect of delay) three times. The forgetting effect is the same for repetition condition 1 and 2, but it is much smaller for repetition condition 3. The size of the forgetting effect depends on the levels of the repetition IV, so here again there is an interaction.

2x2x2 designs

Let's take it up a notch and look at a 2x2x2 design. Here, there are three IVs with 2 levels each. There are three main effects, three two-way (2x2) interactions, and one 3-way (2x2x2) interaction.

We will use the same example as before but add an additional manipualtion of the kind of material that is to be remembered. For example, we could present words during an encoding phase either visually or spoken (auditory) over headphones.

Figure **10.3.3.4** shows two pairs of lines, one side (the panel on the left) is for the auditory information to be remembered, and the panel on the right is when the information was presented visually. The lines still show the delay, and the y-axis still shows the

number of repetitions. In this version of the study, the was only two repetitions levels: once or twice. The two lines on the left show auditory IV levels and the two lines on the right show visual information. The top lines show when there's no delay, and the diagonal lines show when there is a week delay. There are only two levels of repetition, so there are only two dots representing this IV (1 repetition on the right and 2 repetitions on the left for both auditory and visual information).

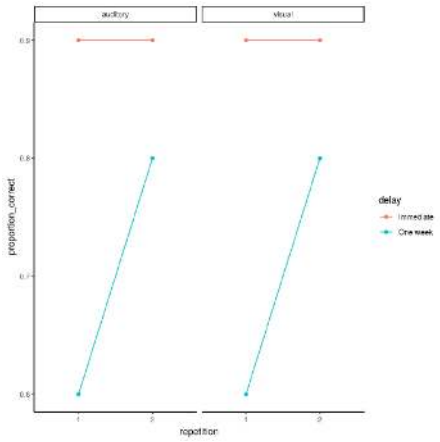


Figure **10.3.3.4**: Example means from a 2x2x2 design with no three-way interaction. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

You can think of the 2x2x2, as two 2x2s, one for auditory and one for visual. What's the take home from this example data? We can see that the graphs for auditory and visual are the same. They both show a 2x2 interaction between delay and repetition. People forgot more things across the week when they studied the material once, compared to when they studied the material twice. There is a main effect of delay, there is a main effect of repetition, but there is no main effect of modality (no difference between auditory or visual information), and there is not a three-way interaction.

What is a three-way interaction anyway? That would occur if there was a difference between the 2x2 interactions. For example, consider the next pattern of results (Figure **10.3.3.5**).

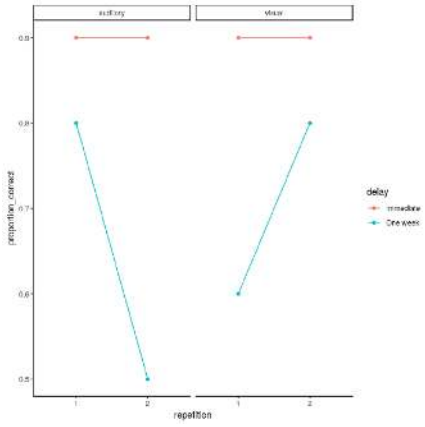


Figure **10.3.3.5**: Example means from a 2x2x2 design with a three-way interaction. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

We are looking at a 3-way interaction between modality, repetition and delay in Figure **10.3.3.5**. What is going on here? These results would be very strange, but here is an interpretation.

> For auditory stimuli, we see that there is a small forgetting effect when people studied things once, but the forgetting effect gets bigger if they studies things twice. A pattern like this would generally be very strange, usually people would do better if they got to review the material twice.
>
> The visual stimuli show a *different pattern*. This different pattern is where we get the three-way interaction. Here, the forgetting effect is large when studying visual things once, and it get's smaller when studying visual things twice.
>
> We see that there is an interaction between delay (the forgetting effect) and repetition for the auditory stimuli; BUT, this interaction effect is *different* from the interaction effect we see for the visual stimuli. The 2x2 interaction for the auditory

stimuli is *different* from the 2x2 interaction for the visual stimuli. In other words, *there is an interaction between the two interactions*, as a result there is a three-way interaction, called a 2x2x2 interaction.

We will note a general pattern here. Imagine you had a 2x2x2x2 design. That would have a 4-way interaction. What would that mean? It would mean that the pattern of the 2x2x2 interaction changes across the levels of the 4th IV. If two three-way interactions are different, then there is a four-way interaction.

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 10.3.3: Interpreting Beyond 2x2 in Graphs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **10.4: More complicated designs** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

10.4: Interpreting Interactions- Do Main Effects Matter?

The interpretation of main effects and interactions can get tricky. Consider the concept of a main effect. This is the idea that a particular IV has a consistent effect. For example, drinking five cups of coffee makes you more awake compared to not drinking five cups of coffee. The main effect of drinking five cups of coffee versus not drinking coffee will generally be true across the levels of other IVs in our life. For example, let's say you conducted an experiment testing whether the effect of drinking five cups of coffee versus not, changes depending on whether you are in your house or in a car. Perhaps the situation matters? No, probably not so much. You will probably still be more awake in your house, or your car, after having five cups of coffee, compared to if you hadn't.

The coffee example is a reasonably good example of a consistent main effect. Another silly kind of example might be the main effect of shoes on your height. For example, if your IV was wearing shoes or not, and your DV was height, then we could expect to find a main effect of wearing shoes on your measurement of height. When you wear shoes, you will become taller compared to when you don't wear shoes. Wearing shoes adds to your total height. In fact, it's hard to imagine how the effect of wearing shoes on your total height would ever interact with other kinds of variables. You will be always be that extra bit taller wearing shoes. Indeed, if there was another manipulation that could cause an interaction that would truly be strange. For example, imagine if the effect of being inside a bodega or outside a bodega interacted with the effect of wearing shoes on your height. That could mean that shoes make you taller when you are outside a bodega, but when you step inside, your shoes make you shorter…but, obviously this is just totally ridiculous. That's correct, it is often ridiculous to expect that one IV will have an influence on the effect of another, especially when there is no good reason.

The summary here is that it is convenient to think of main effects as a consistent influence of one manipulation. However, when an *interaction* is observed, this messes up the consistency of the main effect. That is the very definition of an interaction. It means that some main effect is *not* behaving consistently across different situations. Indeed, whenever we find an interaction, sometimes we can question whether or not there really is a general consistent effect of some manipulation, or instead whether that effect only happens in specific situations.

For this reason, you will often see that researchers report their findings this way:

"We found a main effect of X, BUT, this main effect was qualified by an interaction between X and Y".

Notice the big *BUT*. Why is it there? The sentence points out that before they talk about the main effect, they need to first talk about the interaction, which is making the main effect behave inconsistently. In other words, the interpretation of the main effect *depends* on the interaction, the two things have to be thought of together to make sense of them.

Here are some examples to help you make sense of these issues:

Consistent

A consistent main effect and an interaction means that the results of the interaction follows the results of the main effect. What's going on here? As you can see, the top line (Level 2 of IV2) goes down. The bottom line (Level 1 of IV2) goes up. IV1 is represented by the endpoints of the lines, with Level A of IV1 on the left and Level B of IV1 on the right.

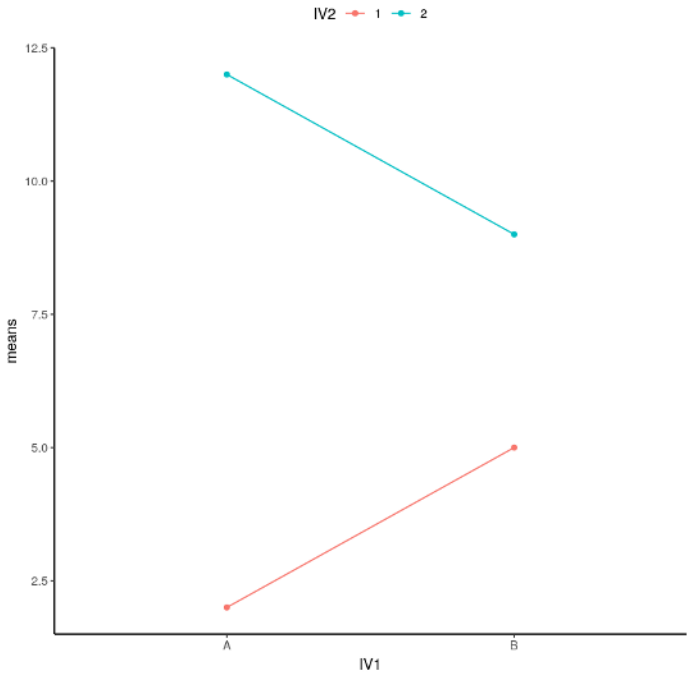


Figure 10.4.1: Example means showing a generally consistent main effect along with an interaction. (CC-BY-SA Matthew J. C. Crump from 10.2 of Answering Questions with Data)

There is a main effect of IV2: the level 1 means (red line on the bottom) are both lower than the level 2 means (aqua line on the top). There is also an interaction. The size of the difference between the red and aqua points in the A condition (left) is bigger than the size of the difference in the B condition. This means that you couldn't just look at the main effect of IV2, you have to also take into account how IV1 affects IV2.

How would we interpret this?

We could say there WAS a main effect of IV2, BUT it was qualified by an IV1 x IV2 interaction.

What's the qualification?

The size of the IV2 effect changed as a function of the levels of IV1. It was big for level A, and small for level B of IV1.

What does the qualification mean for the main effect?

Well, first it means the main effect can be changed by the other IV. That's important to know. Does it also mean that the main effect is not a real main effect because there was an interaction? Not really, there is a generally consistent effect of IV2. The green points are above the red points in all cases. Whatever IV2 is doing, it seems to work in at least a couple situations, even if the other IV also causes some change to the influence.

Let's try an example to show that a consistent interaction, when the interaction is similar to one or both sets of main effects.

Table 10.4.1- Consistent 2x2

IV Levels	**IV1 (Task Presentation): Paper**	**IV1 Task Presentation): Computer**
IV2 (Task Difficulty): Easy	80	90
IV2 (Task Difficulty): Hard	40	70
Marginal Means of IV1:	60	80

How would we interpret this?

As shown in Table 10.4.1, there is a main effect for Task Presentation; overall, performance was better using computer than using paper. (There is also a main effect for Task Difficulty; overall, performance was better on the easier task, but the means are shown to make these examples easier to follow.).

What's the qualification?

The interaction shows that both Task Presentation and Task Difficulty affect the DV.

What does the qualification mean for the main effect?

The pattern of the main effect of Task Presentation is consistent for *both* easy tasks *and* for hard tasks.

Consistency amongst interactions and mains effect is the simplest relationship between a main effect and an interaction. Let's look at one that makes interpreting the main effects more difficult.

Inconsistent

An inconsistent main effect and an interaction is shown in Figure **10.4.2**. What's going on here? As you can see, the top line (Level 2 of IV2) goes down sharply. The bottom line (Level 1 of IV2) stays flat. IV1 is represented by the endpoints of the lines, with Level A of IV1 on the left and Level B of IV1 on the right.

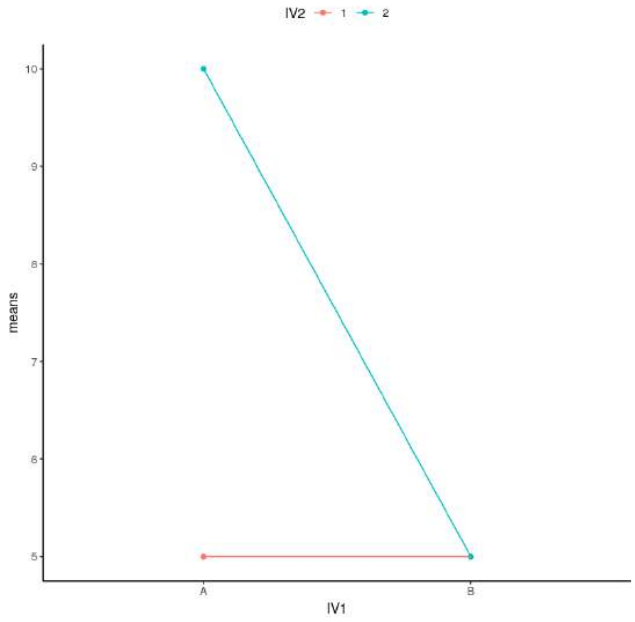


Figure **10.4.2**: Example data showing how an interaction exists, and a main effect does not, even though the means for the main effect may show a difference. (CC-BY-SA Matthew J. C. Crump from 10.2 of Answering Questions with Data)

You should see an interaction here straight away. The difference between the aqua and red points in condition A (left two dots) is huge, and there is no difference between them in condition B. Is there an interaction? Yes!

Are there any main effects here? With data like this, sometimes an ANOVA will suggest that you do have significant main effects. For example, what is the mean difference between level 1 and 2 of IV2? That is the average of the green points ((10+5)/2 = 15/2= 7.5) compared to the average of the red points (5). There will be a difference of 2.5 for the main effect (7.5 vs. 5).

Starting to see the issue here? From the perspective of the main effect (which collapses over everything and ignores the interaction), there is an overall effect of 2.5. In other words, level 2 adds 2.5 in general compared to level 1. However, we can see from the graph that IV2 does not do anything in general. It does not add 2.5s everywhere. It adds 5 in condition A, and nothing in condition B. It only does one thing in one condition.

What is happening here is that a “main effect” is produced by the process of averaging over a clear interaction.

How would we interpret this?

We might have to say there was a main effect of IV2, BUT we would definetely say it was qualified by an IV1 x IV2 interaction.

What's the qualification?

The size of the IV2 effect completely changes as a function of the levels of IV1. It was big for level A, and nonexistent for level B of IV1.

What does the qualification mean for the main effect?

In this case, we might doubt whether there is a main effect of IV2 at all. It could turn out that IV2 does not have a general influence over the DV all of the time, it may only do something in very specific circumstances, in combination with the presence of other factors.

Let's try that with some examples.

Table **10.4.2**- Inconsistent 2x2

IV Levels	**IV1 (Task Presentation): Paper**	**IV1 Task Presentation): Computer**
IV2 (Task Difficulty): Easy	90	90
IV2 (Task Difficulty): Hard	40	70
Marginal Means of IV1:	65	80

How would we interpret this?

Sometimes the interaction shows effects that the main effects miss or are opposite of the main effects. In Table **10.4.2**, there is a main effect for Task Presentation, overall performance was better using computer presentation than using paper presentation. (There is also a main effect for Task Difficulty; overall, performance was better on the easier task, but the means are shown to make these examples easier to follow.).

What's the qualification?

Example 10.4.1

In Table **10.4.2**, what did the main effect of Task Presentation miss?

Solution

Performance is better on the computer but *only for hard tasks* (not for easy tasks). Performance the same for Paper and Computer for Easy tasks.

What does the qualification mean for the main effect?

This qualification means that the interaction is inconsistent with the main effects; performance is affected by *both* Task Difficulty and Task Presentation.

Another form of inconsistency is when the main effects do not show a difference, but the IV levels do affect the DV.

Table **10.4.3**- Inconsistent 2x2 Showing Null Main Effect

IV Levels	**IV1 Task Presentation): Computer**	**IV1 (Task Presentation): Paper**
IV2 (Task Difficulty): Easy	90	70
IV2 (Task Difficulty): Hard	40	60
Marginal Means of IV1:	65	65

How would we interpret this?

There is no main effect for Task Presentation in Figure 3, overall performance was equivalent using computer presentation and using paper presentation.

What's the qualification?

This pattern is inconsistent with the interaction.

What does the qualification mean for the main effect?

In this example, Task Presentation does matter for easy tasks.

This is why interactions are so important to look at. Sometimes the main effects tell us most of what we need to know, but sometimes main effects hide what is really happening.

Try it Yourself

Let's look at the Mindset Quiz pretest scores when faculty in English and Psychology tried to improve their students' mindset. This is different from the Difference scores that we looked at earlier. This time, we are only looking at the posttests, the scores on the Mindset Quiz at the end of the semester. Review Table **10.4.4** and Figure **10.4.3**, and decide if you think that the main effects are consistent or inconsistent with the interaction.

Table **10.4.4** shows the marginal means and the cell means for our two IVs (Intervention: Yes or No; Department: English or Psychology).

Table **10.4.4**- Mindset Pretest Scores

IV Levels	**IV1: Yes Intervention**	**IV1: No Intervention**	Marginal Means for IV2:
IV2: English Department	41.15	41.43	41.29
IV2: Psychology Department	46.46	41.60	44.03
Margin Means for IV1:	43.8	41.51	(You could put the total mean here...)

How would we interpret this?

Let's first look at the main effects in Table **10.4.4**. Without doing a null hypothesis significance test, we might guess that there was a significant different between those who were going to receive the intervention and those who were not going to receive the intervention based on the means; remember, these are pretest scores so no one had received the intervention yet! So right off, that's an odd finding. Why would students already having a higher score on the Mindset Quiz? Do you think that the professor was subtly supporting growth mindset in only the classes that were going to have the intervention?

Now for the department. Without doing an null hypothesis significance testing, we might guess that there was a significant difference with the Psychology students having higher scores on the Mindset Quiz (slightly more growth mindset) than the English students. Again, this is surprising for a pretest. Why would Psychology students have a growth mindset more than English students?

On the other hand, we didn't conduct any null hypothesis significance testing, so maybe these means are actually close enough to say that they are similar. At this point, we just don't know.

But let's move on to the interaction. Looking at the four cells in the middle of the table, what do you see?

Example 10.4.2

Does one cell in the middle of Table **10.4.4** seem substantially higher? Or substantially lower?

Solution

It looks like the combination of IV levels strengthened the effect of the IVs alone. When discussing the main effects we said that it looks like the students who were going to get the intervention (IV1 = Yes) already scored higher on the Mindset Quiz, *and* the Psychology students scored higher on the pretest than English students. When looking at these in combination, it looks like the only the Psychology classes that were going to receive the intervention scored higher, not everyone who was going to

receive the intervention. This high average score on the Mindset Quiz pretest for Psychology students who were going to receive the intervention seems to have increased the marginal mean for both the Intervention IV and the Department IV.

When you look at Figure **10.4.3**, do you see these means? The top line represents those who experienced an intervention (IV1, Level: Yes), and rises sharply. The dashed bottom line represents those who did not experience an intervention (IV1, Level: No). This dashed line most stays flat. IV2 is represented by the endpoints of the lines, with the English Department on the left and the Psychology Department on the right. This means that the one group that looks really different are those in Psychology who experienced the intervention.

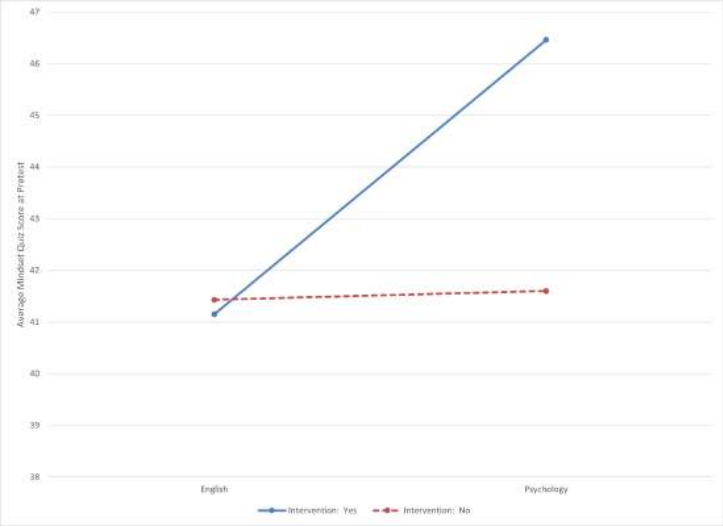


Figure **10.4.3**: Mindset Quiz Pretest Means. (CC-BY-SA; Michelle Oja)

Figure **10.4.3** makes it very clear that there is an interaction (the lines are not parallel). It also makes clear that the scores for English students weren't affected by the other IV (dots on the left), and that those who were not going to experience an intervention (dashed line) did not differ between the two departments. But, woah nelly, look at the blue dot representing students who were going to receive an intervention in a Psychology class! Based on the scaling of this chart, it looks like their mean was wildly higher than the other combinations.

What's the qualification?

The size of the effect of the Department (IV2) completely changes as a function of the levels of Intervention (IV1).

What does the qualification mean for the main effect?

It might mean that the main effect of Department doesn't really matter because the change in Mindset Quiz pretest scores only changes when in a Psychology class that will receive an intervention.

Consistent or Inconsistent?

Example 10.4.3

Are the main effects are consistent or inconsistent with the interaction?

Solution

Both main effects are consistent with the interaction.

Do main effects mean anything?

Example 10.4.4

When should you pay attention to mean differences in main effects?

Solution

When there are no interactions.

When there are no interactions, the main effects for each IV are like separate ANOVAs. An ANOVA with one IV. This is what makes interpreting main effects risky when there's an interaction. The interaction may be all that is happening, but based on the means of only one of the IVs, it looks like that IV is influencing the DV on its own.

Contributors and Attributions

Matthew J. C. Crump (Brooklyn College of CUNY)

Dr. MO (Taft College)

This page titled 10.4: Interpreting Interactions- Do Main Effects Matter? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **10.2: Interpreting main effects and interactions** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

10.5: Summary of Factorial Designs

Key Takeaways

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

What's Next?

Factorial designs are the biggest type of designs that we will cover. Up next are small-N designs, or experiments with just a few participants.

Exercises

- Practice: Create a factorial design table for an experiment on the effects of room temperature and noise level on performance on the MCAT. Be sure to indicate whether each independent variable will be manipulated between-subjects or within-subjects and explain why.
- Practice: Complete the Examples and Exercises in Oja (2021)'s 13.1.1: Factorial Notations and and Square Tables to practice identifying IVs, the DV, and the type of factorial designs.
- Practice: Review the chapter on factorial designs in any (behavioral) statistics textbook on LibreTexts. Behavioral Statistics, Ed. 1 (Oja, 2021) provided some of the information in this chapter.
- Discussion: Compare and contrast (find similarities and differences) between Introduction to Main Effects and Interactions (section 13.2) and the information presented in this chapter, then discuss which section is better and why.
- Discussion: Compare and contrast (find similarities and differences) between Graphing Main Effects and Interactions (section 13.2.2) and the information presented in this chapter, then discuss which section is better and why.

This page titled 10.5: Summary of Factorial Designs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **9.4: Factorial Designs (Summary)** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

11: Small-N Designs

11.1: Prelude to Single-Subject Research

11.2: Introduction to Small-N Research

11.3: Small-N Research Designs and Analysis

11.4: When should you use Small-N Designs?

11.5: Summary of Small-N Design

11: Small-N Designs is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

11.1: Prelude to Single-Subject Research

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, Lund, & Jackson, 1968)[1]. 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.

References

1. Hall, R. V., Lund, D., & Jackson, D. (1968). Effects of teacher attention on study behavior. *Journal of Applied Behavior Analysis, 1*, 1–12. ↵

11.1: Prelude to Single-Subject Research is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **10.1: Single-Subject Research** has no license indicated.

11.2: Introduction to Small-N Research

Learning Objectives

1. Explain what small-N research is, including how it differs from other types of research.
2. Explain who uses small-N research and why.

What Is Small-N Research?

Small-N (where N is the used to represent the sample size) research is a type of quantitative research that involves studying in behavior of each of a small number of participants alternating between a baseline (no intervention or IV) to an intervention condition. This type of research used to be called *single-subject* even though it doesn't always focus on one participant in each study, nor do we use the term "subject" to refer to study participants any more. Small-N designs typically have somewhere between one and 10 participants. Small-N research can be contrasted with large-N, or 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 small-N research is an important alternative, and it is the primary approach in some more applied areas of psychology.

Before continuing, it is important to distinguish small-N research from case studies and other more qualitative approaches that involve studying in detail a small number of participants. As described in the next chapter, case studies involve an in-depth analysis and description of an individual, which is typically primarily qualitative in nature. More broadly speaking, qualitative research focuses on understanding people’s experience by observing behavior and collecting relatively unstructured data (e.g., detailed interviews) and analyzing those data using narrative rather than quantitative techniques. Small-N research, in contrast, focuses on understanding objective behavior through experimental manipulation and control, collecting highly structured data, and analyzing those data quantitatively.

Assumptions of Small-N Research

Again, small-N research involves studying a small number of participants and focusing intensively on the behavior of each one. There are several important assumptions underlying small-N that explain the reasoning to use this type of research rather than the traditional large-N studies that we have been working with.

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. Small-N 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 particularly at-risk student. Although previous published research (both small-N and large-N 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 small-N 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, small-N research is often considered a type of experimental research with good internal validity. Recall, for example, that Hall et al. (1968) 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 small-N 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). Hall et al. (1968), 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.

Who Uses Small-N Research?

Small-N 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 small-N 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 small-N 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*. However, applied behavioral analysis, or ABA, has also come under attack. While the effectiveness of ABA is not disputed, some argue that it focuses too much on modifying behavior to be acceptable to those in charge rather than improving the experience of the participants. This is particularly relevant when using ABA to change the behavior of children with autism to fit in to neurotypical spaces, rather than considering the feelings of the children with autism and trying to change the spaces to be more welcoming and accommodating to neurodivergent people.

Although most contemporary small-N 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 small-N approach. The small-N 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).

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.

Hall, R. V., Lund, D., & Jackson, D. (1968). Effects of teacher attention on study behavior. *Journal of Applied Behavior Analysis, 1*, 1–12.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.

Skinner, B. F. (1938). T*he behavior of organisms: An experimental analysis*. New York, NY: Appleton-Century-Crofts.

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.

This page titled 11.2: Introduction to Small-N Research is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **10.2: Overview of Single-Subject Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

11.3: Small-N Research Designs and Analysis

Learning Objectives

1. Describe the basic elements of a small-N research design.
2. Design simple small-N studies using reversal and multiple-baseline designs.
3. Explain how small-N research designs address the issue of internal validity.
4. Interpret the results of simple small-N studies based on the visual inspection of graphed data.

General Features of Small-N Research Designs

Before looking at any specific small-N 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 **11.3.1**, which shows the results of a generic small-N study. First, the dependent variable (represented on the *y*-axis of the graph, the one going up and down) is measured repeatedly over time (represented by the *x*-axis, the one going side to side) 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 **11.3.1** 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.)

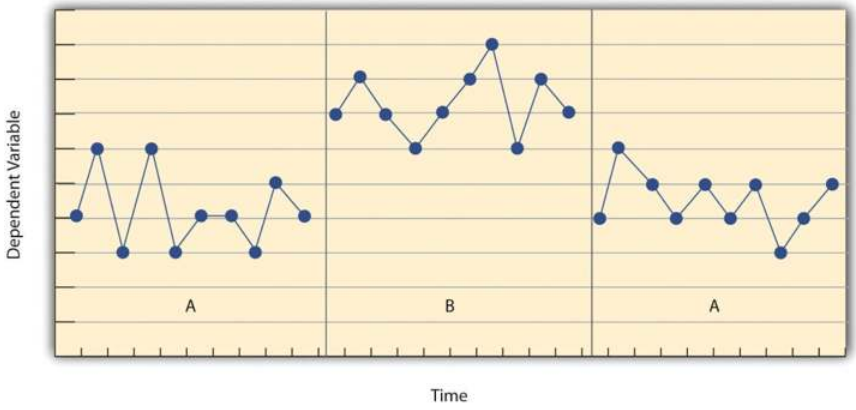


Figure **11.3.1**: Results of a Generic Small-N Study Illustrating Several Principles of Small-N Research

Another important aspect of small-N 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. As with any experimental study, the effect of an independent variable is easier to detect when the "noise" in the data is minimized.

Reversal Designs

The most basic small-N research design is the reversal design, also called the ABA design. During the first phase, A, a baseline is established for the dependent variable. A baselin is the condition in which the dependent variable is measured in the absence of any treatment. Most small-N designs begin with a baseline condition with repeated measurements of the DV, and many return to the baseline condition at least once. The baseline 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.

Hall et al. (1968) employed an ABAB reversal design. Figure **11.3.2** 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.

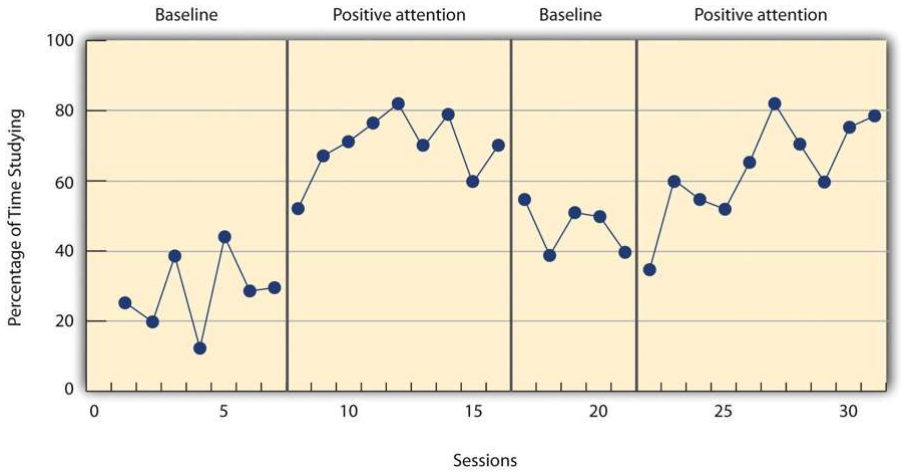


Figure **11.3.2**: An Approximation of the Results for Hall and Colleagues' Participant Robbie in Their ABAB Reversal Design

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. 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.

The key to a good ABA design is that the IV can change quickly, and the DV is easy to observe. This type of design won't work for an IV that we might expect to have long-term or permanent changes. For example, you can't teach someone to ride a bike or cope with stress, and then expect them not to do what they just learned when you stop teaching them about it. Once you learn these skills, you know them so you can never go back to baseline. Relatedly, if you were interested in improving mood as your DV, an ABA design might not be a great option because mood is hard to detect through observation alone. Surveys are typically not used in small-N designs, instead the outcome is easy to observe and easy to count.

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 mild punishment for not studying (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.

Try it yourself!

Can you draw a figure like Figure **11.3.2** that could represent a multiple-treatment reversal design (ABCACB)?

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 mild punishment for not studying 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.

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. But 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 **11.3.3**. There are three different types of multiple-baseline designs which we will now consider.

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. But 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.

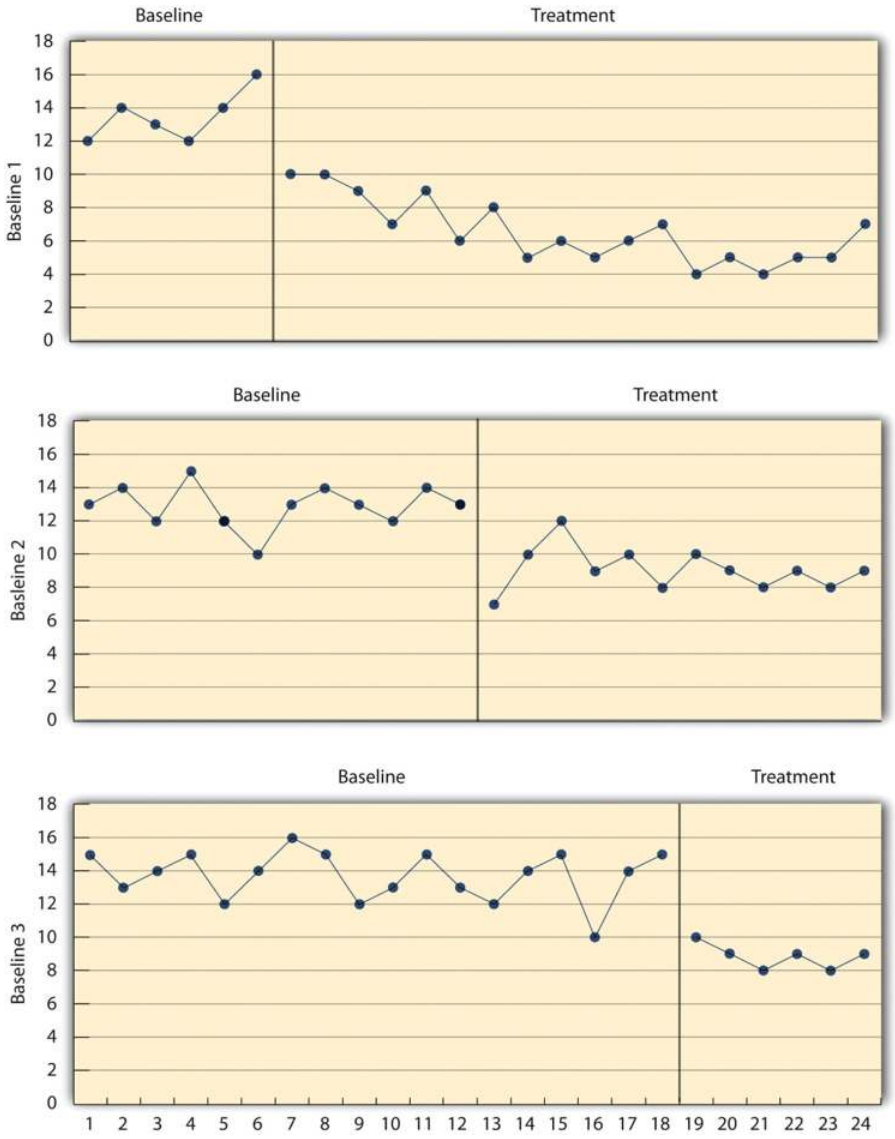


Figure **11.3.3**: Results of a Generic Multiple-Baseline Study. The multiple baselines can be for different participants, dependent variables, or settings. The treatment is introduced at a different time on each baseline.

As an example, consider a study by Scott Ross and Robert Horner (Ross & Horner, 2009). They were interested in how a school-wide bullying prevention program affected the bullying behavior of particular problem students. At each of 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 two weeks, they implemented the program at one school. After 2 more weeks, they implemented it at the second school. And after two more weeks, they implemented it at the third school. Ross and Horner (2009) 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 Ross and Horner (2009) 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). But with their multiple-baseline design, this kind of coincidence would have to happen three separate times—a very unlikely occurrence—to explain their results.

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. But 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.

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.

Data Analysis in Small-N 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. 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-subject researchers take several factors into account. One of them is changes in the level of the dependent variable from condition to condition. 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 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 third factor is latency, which 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 **11.3.4**, 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 **11.3.4**, however, the changes in level are fairly small. And 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 small-N researchers typically hope to see.

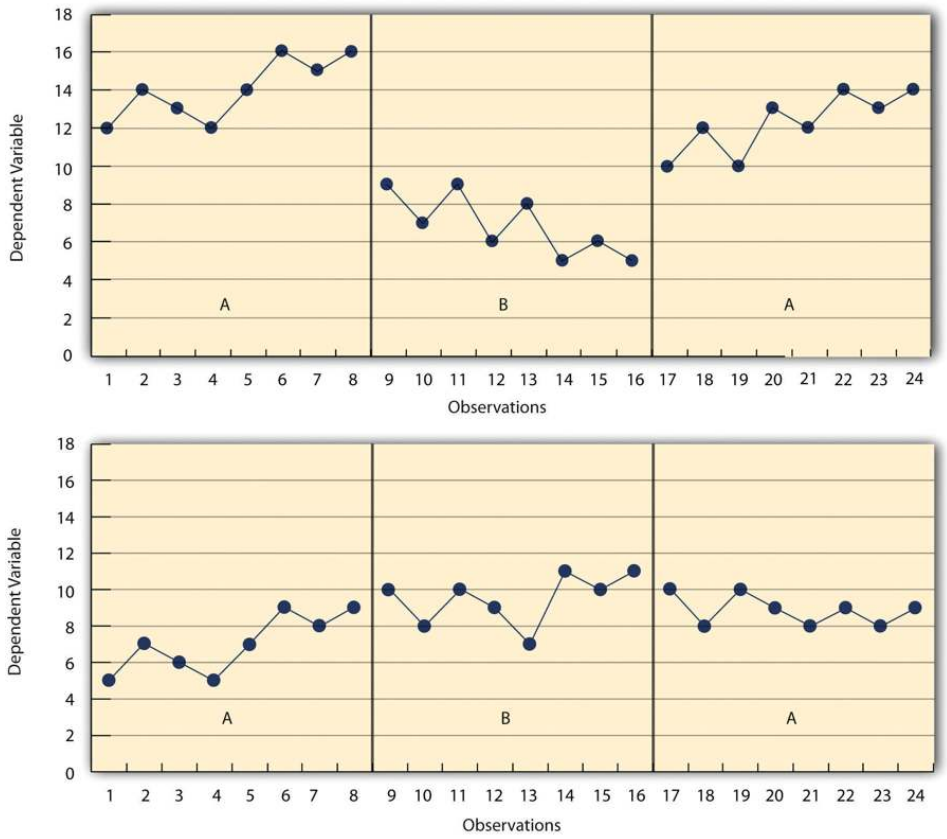


Figure **11.3.4**: Results of a Generic Single-Subject Study Illustrating Level, Trend, and Latency. Visual inspection of the data suggests an effective treatment in the top panel but an ineffective treatment in the bottom panel.

The results of small-N research can also be analyzed using statistical procedures—and this is becoming more common. There are many different approaches, and small-N 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 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 Hall et al. (1968), 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.

References

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models. *Behavioral Processes*, *54*, 137–154.

Ross, S. W., & Horner, R. H. (2009). Bully prevention in positive behavior support. *Journal of Applied Behavior Analysis*, *42*, 747–759.

Scruggs, T. E., & Mastropieri, M. A. (2001). How to summarize single-participant research: Ideas and applications. *Exceptionality*, *9*, 227–244.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. Boston, MA: Authors Cooperative.

This page titled 11.3: Small-N Research Designs and Analysis is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **10.3: Single-Subject Research Designs** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

11.4: When should you use Small-N Designs?

Learning Objectives

1. Explain some of the points of disagreement between advocates of single-subject research and advocates of group research.
2. Identify several situations in which single-subject research would be appropriate and several others in which group research would be appropriate.

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. But 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.

Data Analysis

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)[1]. 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)[2].

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 (introverts vs. extraverts).

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. But 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, Cook, & Campbell, 2002)[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 that have different strengths and weaknesses and that are appropriate for answering different kinds of research questions (Kazdin, 1982)[4]. 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 have 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 have they have experienced without actually having to count them (Dehaene, 2011)[5]. 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.

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 weakness. 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 have 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 weakness. If the result 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.

References

1. 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. ↵
2. Scruggs, T. E., & Mastropieri, M. A. (2001). How to summarize single-participant research: Ideas and applications. *Exceptionality, 9*, 227–244. ↵
3. Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin. ↵
4. Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press. ↵
5. Dehaene, S. (2011). *The number sense: How the mind creates mathematics* (2nd ed.). New York, NY: Oxford. ↵

This page titled 11.4: When should you use Small-N Designs? is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **10.4: The Single-Subject Versus Group "Debate"** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

11.5: Summary of Small-N Design

Key Takeaways

- Small-N designs, which involve experimentally testing a small number of participants and focusing intensively on the behavior of each individua, is an important alternative to group research in psychology.
- Small-N design studies must be distinguished from qualitative research (discussed in detail next) on a single person or small number of individuals. Unlike more qualitative research, Small-N design research focuses on understanding objective behavior through experimental manipulation and control, collecting highly structured data, and analyzing those data quantitatively.
- Small-N research has been around since the beginning of the field of psychology. Today it is most strongly associated with the behavioral theoretical perspective, but it can in principle be used to study behavior from any perspective.
- Small-N designs typically involve measuring the dependent variable repeatedly over time and changing conditions (e.g., from baseline to treatment) when the dependent variable has reached a steady state. This approach allows the researcher to see whether changes in the independent variable are causing changes in the dependent variable.
- In a reversal design, the participant is tested in a baseline condition, then tested in a treatment condition, and then returned to baseline. If the dependent variable changes with the introduction of the treatment and then changes back with the return to baseline, this provides strong evidence of a treatment effect.
- In a multiple-baseline design, baselines are established for different participants, different dependent variables, or different settings—and the treatment is introduced at a different time on each baseline. If the introduction of the treatment is followed by a change in the dependent variable on each baseline, this provides strong evidence of a treatment effect.
- Small-N designs are typically analyzed by graphing the data and making judgments about whether the independent variable is affecting the dependent variable based on level, trend, and latency.
- Differences between small-N designs and large-N research sometimes lead to disagreements between researchers. These disagreements center on the issues of data analysis and external validity (especially generalization to other people).
- Small-N design and large-N research are probably best seen as complementary methods, with different strengths and weaknesses, that are appropriate for answering different kinds of research questions.

What's Next?

You've read about many of the common types of experimental designs in the social sciences. Now, we will focus on nonexperimental designs. While we can't show cause-and-effect with these types of designs, they provide important information. Just like small-N designs are complementary to large-N designs, experimental and nonexperimental research can provide insight and context to each other.

Exercises

- Practice: Find and read a published article that reports new small-N research. (An archive of articles published in the Journal of Applied Behavior Analysis can be found at http://www.ncbi.nlm.nih.gov/pmc/journals/309/) Write a short summary of the study.
- Practice: Come up with three outcome behaviors (DVs) related to helping clients in your future career that would work well in a small-N design. These behaviors should be easy to observe, and would change quickly based on whether they are observed in the baseline or treatment condition.
- Practice: Come up with three treatments (IVs) related to helping clients in your future career that would work well in a small-N design. These interventions should be something that we expect to affect the outcome right away and also may quickly fade so that we can see changes in the DV based on the IV easily.
- Practice: Using one of the treatments and one of the behaviors that you listed above, design a simple small-N study (using either a reversal or multiple-baseline design) to answer a question to help clients in your future career. Be sure to specify the treatment, operationally define the dependent variable, decide when and where the observations will be made, and so on.
- Practice: Create a graph that displays the hypothetical results for the study you designed above, then write a paragraph in which you describe what the results show. Be sure to comment on level, trend, and latency.

- Discussion: Imagine you have conducted a small-N study showing a positive effect of a treatment on the behavior of a man with social anxiety disorder. Your research has been criticized on the grounds that it cannot be generalized to others. How could you respond to this criticism?
- Discussion: Imagine you have conducted a large-N study showing a positive effect of a treatment on the behavior of a group of people with social anxiety disorder, but your research has been criticized on the grounds that "average" effects cannot be generalized to individuals. How could you respond to this criticism?
- Discussion : Redesign as a group study the study by Hall et al. (1968) described at the beginning of this chapter. List the strengths and weaknesses of each study, and decide which type of study is most appropriate for this topic.
- Discussion: The generation effect refers to the fact that people who generate information as they are learning it (e.g., by self-testing) recall it better later than do people who simply review information. Design a small-N study on the generation effect applied to university students learning brain anatomy, then design a large-N study on the same topic. List the strengths and weaknesses of each study, and decide which type of study is most appropriate for this topic.

References

Hall, R. V., Lund, D., & Jackson, D. (1968). Effects of teacher attention on study behavior. *Journal of Applied Behavior Analysis*, *1*, 1–12.

This page titled 11.5: Summary of Small-N Design is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **10.5: Single-Subject Research (Summary)** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

CHAPTER OVERVIEW

12: Nonexperimental Research

12.1: Prelude to Nonexperimental Research
12.2: Introduction to Nonexperimental Research
12.3: Longitudinal and Similar Designs
12.4: Qualitative Research
12.5: Case Studies
12.6: Summary of Nonexperimental Research

12: Nonexperimental Research is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

12.1: Prelude to Nonexperimental 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 dangerous shocks to another person just because they were told to by an authority figure (Milgram, 1963)[1].
- 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)[2].
- 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)[3].
- David Rosenhan found that confederates who went to psychiatric hospitals claiming to have heard voices saying things like “empty” and “thud” were labeled as schizophrenic by the hospital staff and kept there even though they behaved normally in all other ways (Rosenhan, 1973)[4].

The answer for purposes of this chapter is that they are not experiments.

References

1. Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371–378. ↵
2. Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals, 25*, 720–725. ↵
3. Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116–131. ↵
4. Rosenhan, D. L. (1973). On being sane in insane places. *Science, 179*, 250–258. ↵

12.1: Prelude to Nonexperimental Research is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **6.1: Prelude to Nonexperimental Research** has no license indicated.

12.2: Introduction to Nonexperimental Research

Learning Objectives

1. Define non-experimental research, distinguish it clearly from experimental research.
2. Explain when a researcher might choose to conduct non-experimental research as opposed to experimental research.

What Is Non-Experimental Research?

Non-experimental research is research that lacks the manipulation of an independent variable and does not require initial equivalence between the comparison groups. Because of this, the studies lack internal validity and cannot be used to determine cause and effect relationships. However, recall the three goals of science discussed in section 2.2 (Goals of Science).

Exercise 12.2.1

Describe each of the three goals of science discussed in chapter 2.

Answer

The three goals of science are:

- Desciption: This goal is to describe reality and individuals' experiences.
- Prediction: This goal is to use one variable to predict the outcome of another variable.
- Intervention: This goal is to improve behavior and experiences based on causal relationships between variables.

Do all of these goals need to determine cause-and-effect relationships? No! The goal of description definitely does not need to determine the cause of why things are as they are. And with our knowledge of "correlation is no causation" or "What about Z?", we know that just because variables are related (and can be used to predict each other) doesn't mean that one variable caused changes in the other.

While most social science researchers consider the distinction between experimental and non-experimental research to be an extremely important one because experimental research can provide strong evidence that changes in an independent variable cause differences in a dependent variable, non-experimental research generally cannot. As we will see, however, this inability to make causal conclusions does not mean that non-experimental research is less important than experimental research. It is simply used when the goal is not (yet?) intervention.

When to Use Non-Experimental Research

As we have seen throughout this textbook, experimental research is appropriate when the researcher has a specific research question or hypothesis about a causal relationship between two variables—and it is possible, feasible, and ethical to manipulate the independent variable. It stands to reason, therefore, that non-experimental research is appropriate—even necessary—when these conditions are not met. There are many times in which non-experimental research is preferred, including when:

- The research question or hypothesis relates to a single variable rather than a statistical relationship between two variables (e.g., how accurate are people’s first impressions?).
- The research question pertains to a non-causal statistical relationship between variables (e.g., is there a correlation between verbal intelligence and mathematical intelligence?).
- The research question is about a causal relationship, but the independent variable cannot be manipulated or participants cannot be randomly assigned to conditions or orders of conditions for practical or ethical reasons (e.g., does damage to a person’s hippocampus impair the formation of long-term memory traces?).
- The research question is broad and exploratory, or is about what it is like to have a particular experience (e.g., what is it like to be a working mother diagnosed with depression?).

Again, the choice between the experimental and non-experimental approaches is generally dictated by the nature of the research question. 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).

Internal Validity Revisited

Recall that 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 **12.2.1** shows how experimental, quasi-experimental, and non-experimental 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) with initial equivalence and ongoing equivalence between the IV groups help to rule out alternative explanations for any changes found in the DV. 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 research is lowest in internal validity because these designs fail to use manipulation or control; but this doesn't mean that non-experimental studies are bad! Researchers have different topics and different goals, and sometimes non-experimental designs are the most appropriate. Quasi-experimental research falls in the middle because it contains some, but not all, of the features of a true experiment. For instance, it may fail to use random assignment to assign participants to groups or fail to use counterbalancing to control for potential order effects. Imagine, for example, that a researcher finds two similar schools, starts an anti-bullying program in one, and then finds fewer bullying incidents in that "treatment school" than in the "control school." While a comparison is being made with a control condition, the inability to randomly assign children to schools could still mean that students in the treatment school differed from students in the control school in some other way that could explain the difference in bullying (e.g., there may be a selection effect).

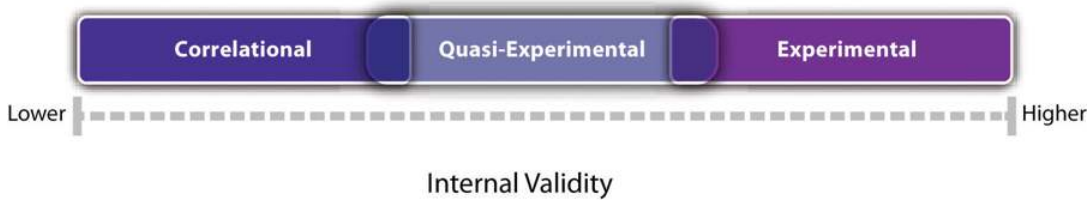


Figure **12.2.1**: Internal Validity of Correlation, Quasi-Experimental, and Experimental Studies. Experiments are generally high in internal validity, quasi-experiments lower, and correlation (non-experimental) studies lower still.

Notice also in Figure **12.2.1** that there is some overlap in the internal validity of experiments, quasi-experiments, and 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, as noted in Chapter 5.

References

Milgram, S. (1974). *Obedience to authority: An experimental view*. New York, NY: Harper & Row.

This page titled 12.2: Introduction to Nonexperimental Research is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **6.2: Overview of Non-Experimental Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

12.3: Longitudinal and Similar Designs

Learning Objectives

1. Explain why longitudinal designs are not quasi-experimental designs.
2. Provide an example of longitudinal designs compared to cross-sectional designs.

Studies of Time

When researchers wish to study change over time (for example, when developmental psychologists wish to study aging) but they do not (or cannot) manipulate an IVs), they usually take one of three non-experimental approaches: cross-sectional, longitudinal, or cross-sequential.

Longitudinal Design

To understand the other types of designs to understand changes though time, it is easier to start with longitudinal designs. Longitudinal designs follow one group of people over time, taking measurements regularly. These type of designs are often use to follow children through development in psychology, or follow the progression of diseases in medicine. As professor in a Psychology department, Dr. MO is interested in the progression of statistical literacy (the ability to understand and interpret numbers and statistics) for Psychology majors from their first course to their graduation. She hopes that taking Psychology courses will improve statistical literacy as students progress through the Psychology curriculum. To test this, Dr. MO could send a statistical literacy assessment to 100 new students who identified their major as Psychology a week before one fall semester, then continue to send this same assessment to the same participants until they graduate (or for 10 years, whichever happens first).

One benefit of this type of design is that the development of changes in thoughts, behaviors, and feelings can discovered. There are a couple of disadvantages, however. One disadvantage is that there is usually no random assignment of IVs or interventions in these types of designs. Instead, once the sample is chosen, the researcher measures the variables of interest but does not manipulate anything. While there may end up being comparisons between groups, these would be natural groups or quasi-experimental, and cannot be guaranteed. Another disadvantage is that longitudinal studies are, by definition, time consuming. It is tough to argue to a granting agency to fund a 20-year research study that may not begin providing any useful data for up to 10 years. In addition to the time and cost investment of researchers, longitudinal designs also require a large investment on the part of the participants. As participants stop participating in the study, researchers have to consider if the current sample represents the desired population.

Cross-Sectional Design

To try to avoid these disadvantages, researchers sometimes choose instead to conduct cross-sectional studies. Cross-sectional studies are a type of research design that involves measurement of the same variable for individuals who are at varying points of progress or development. This definition doesn't really make sense until you have an example, so let's go back to those Psychology majors' statistical literacy. Rather than waiting 10 years for all of the data, Dr. MO could instead send the statistical literacy assessment to 100 new students who are Psychology majors, 100 Psychology majors who are sophomores, 100 Psychology majors who are juniors, 100 Psychology majors who are seniors, and 100 Psychology majors who graduated the last semester. She could do this a week before this fall semester, and would receive the data from those who choose to participate by the start of the fall semester, rather than waiting up to 10 years! Just like with longitudinal designs, this is non-experimental because there is no manipulation of an independent variable and there's no initial equivalence or ongoing equivalence.

A benefit of this design is that the data from the participants is received quickly. A disadvantage of using this type of design is that differences between the groups other than progression or development may account for differences in the dependent variable. For instance, in our study of statistical literacy, we might find that those who graduate do have higher scores on the assessment than those sophomores, which could suggest that the curriculum improves statistical literacy. However, because we tested everyone at the same time, it could be that the students who had already graduated had taken their first behavioral statistics course from a specific instructor that helped them understand, but that instructor was assigned to teach research methods so the sophomore

students took behavioral statistics from another instructor. Was it the curriculum that improved statistical literacy, or was it something about what the graduates had experienced but none of the other cohorts would experience? Or, because cross-sectional designs are nonexperimental, could there be other variables related to the groups that affect the outcome that's not time, progress, or development? For studies looking at (physical, emotional, or social) maturation through time, differences between the groups may reflect the generation that people come from (another example of a cohort effect) rather than a direct effect of age.

Cross-Sequential Design

A third approach, known as cross-sequential studies, combines elements of both cross-sectional and longitudinal studies. Rather than measuring differences between people in different age groups or following the same people over a long period of time, researchers adopting this approach choose a smaller period of time during which they follow people in different age groups. For example, they might measure changes over a ten year period among participants who at the start of the study fall into the following age groups: 20 years old, 30 years old, 40 years old, 50 years old, and 60 years old. This design is advantageous because the researcher reaps the immediate benefits of being able to compare the age groups after the first assessment. Further, by following the different age groups over time they can subsequently determine whether the original differences they found across the age groups are due to true age effects or cohort effects.

Quasi-Experimental or Nonexperimental?

As noted throughout this section, these types of designs (longitudinal, cross-sectional, and cross-sequential) are typically nonexperimental. There is usually no manipulation of an IV because time or development is considered the "intervention". This means that there's no initial equivalence nor ongoing equivalence between the groups; in fact, the different ongoing experiences through out time is often what's being studied! This places these types of designs squarely into the "nonexperimental design" approach.

However, it would be possible to start with (non-equivalent) groups, which could make specific versions of these designs a quasi-experimental design. For example, In our example of statistical literacy in Psychology majors, we could compare the statistical literacy of students throughout their 4-10 years in college but compare students who took a behavioral statistics course with students who took a statistics course from the Math department. This would be more like an interrupted time series design now, which is quasi-experimental.

In the end, you as the researcher and you as the reader of research must decide who comfortable you are in concluding that nothing other than the IV could have affected the DV. Even if a design "looks" like nonexperimental, it might be quasi-experimental. And with comparing pretests with posttests, quasi-experimental designs may have more control than some poorly-ran experimental designs.

This page titled 12.3: Longitudinal and Similar Designs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **6.2: Overview of Non-Experimental Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

12.4: Qualitative Research

Learning Objectives

1. List several ways in which qualitative research differs from quantitative research in the social sciences.
2. Describe the strengths and weaknesses of qualitative research compared with quantitative research.
3. Give examples of qualitative research.

What Is Qualitative Research?

This textbook is primarily about quantitative research, in part because most studies conducted in psychology are quantitative in nature . Quantitative researchers typically start with a focused research question or hypothesis, collect a small amount of numerical data from a large number of individuals (large-N design), describe the resulting data using statistical techniques, and draw general conclusions about some large population. Although this method is the most common approach to conducting empirical research in psychology, qualitative research is an important alternative in psychology and other social sciences. Qualitative research originated in the disciplines of anthropology and sociology but is used to study psychological topics as well. Qualitative researchers generally begin with a research question, collect large amounts of relatively “unfiltered” data from a relatively small number of individuals, and describe their data using nonstatistical techniques. An introduction to analyzing qualitative data was considered in the chapter on analyzing. Qualitative researchers are sometimes less concerned with drawing general conclusions, especially about cause-and-effect relationships, about human behavior, and may be more interested in understanding in detail the *experience* of their research participants.

Consider, for example, a study by researcher Per Lindqvist et al. (2008), who wanted to learn how the families of teenage suicide victims cope with their loss. Lindqvist et al. (2008) did not have a specific research question such as "What percentage of family members join suicide support groups?", and there was no hypothesis. Instead, Lindqvist et al. (2008) 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, Lindqvist et al. (2008) interviewed the families of 10 teenage suicide victims in their homes in rural Sweden. The interviews were relatively unstructured, beginning with a general request for the families to talk about the victim and ending with an invitation to talk about anything else that they wanted to tell the interviewer. One of the most important themes that emerged from these interviews was that even as life returned to “normal,” the families continued to struggle with the question of why their loved one committed suicide. This struggle appeared to be especially difficult for families in which the suicide was most unexpected. Review Figure 12.4.1: Characteristics of Qualitative Research box, then consider how each of these differs (or not) from quantitative research.



Figure 12.4.1: Characteristics of Qualitative Research. (CC-BY-NC; Bunmi Malau-Aduli and Faith Alele via Introduction to Research Methods for Undergraduate Health Profession Students (4.2: Definitions and Characteristics of Qualitative Research))

For ease of reading, the content in Figure 12.4.1 is repeated here:

- Holistic account: Qualitative research seeks to develop a complex picture of the problem or issues under study by identifying the many factors.
- Reflexivity: Researchers reflect on their role in the study and how their culture, personal background and experiences may affect data interpretation.
- Emergent design: The research process is emergent; some phases of the process may shift or change once data collection begins.
- Participants meaning: Participants meaning of the problem or issue is the focus of the research your desired text here.
- Inductive and deductive data analysis: Patterns,categories, themes are built inductively. Evidence from the data to support each of them are obtained deductively.
- Multiple sources of data: Multiple forms of data such as interviews, observations, documents and audio-visual information are gathered.
- Researchers as key instrument: Researchers collect data themselves by examining documents, observing behavior or interviewing participants.
- Natural setting: Qualitative researchers typically gather data at the location where participants encounter the phenomenon under investigation.

The Purpose of Qualitative Research

Again, this textbook is primarily about quantitative research in psychology. The strength of quantitative research is its ability to provide precise answers to specific research questions, especially related to the causes of behavior. This method is how we know that people have a strong tendency to obey authority figures, for example, and that students who are women are *not* substantially more talkative than students who are men. But while quantitative research is good at providing precise answers to specific research questions, it is not nearly as good at *generating* novel and interesting research questions. Likewise, while quantitative research is good at drawing general conclusions about human behavior, it is not nearly as good at providing detailed descriptions of the behavior of particular groups in particular situations. And quantitative research is not very good at communicating what it is actually like to be a member of a particular group in a particular situation.

But the relative weaknesses of quantitative research are the relative strengths of qualitative research. Qualitative research can help researchers to generate new and interesting research questions and hypotheses. The findings of Lindqvist et al. (2008), for example, suggests that there may be a general relationship between how unexpected a suicide is and how consumed the family is with trying to understand why the teen committed suicide. This relationship can now be explored using quantitative research. But it is unclear whether this question would have arisen at all without the researchers sitting down with the families and listening to what they themselves wanted to say about their experience. Qualitative research can also provide rich and detailed descriptions of human behavior in the real-world contexts in which it occurs. Among qualitative researchers, this depth is often referred to as "thick description" (Geertz, 1973). Similarly, qualitative research can convey a sense of what it is actually like to be a member of a particular group or in a particular situation—what qualitative researchers often refer to as the "lived experience" of the research participants. Lindqvist et al. (2008), for example, describe how all the families spontaneously offered to show the interviewer the victim's bedroom or the place where the suicide occurred—revealing the importance of these physical locations to the families. It seems unlikely that a quantitative study would have discovered this detail.

Table **12.4.1**: Some contrasts between qualitative and quantitative research

Qualitative	Quantitative
In-depth information about relatively few people	Less depth information with larger samples
Conclusions are based on interpretations drawn by the investigator	Conclusions are based on statistical analyses
Global and exploratory	Specific and focused

Data Collection and Analysis in Qualitative Research

Data collection approaches in qualitative research are quite varied. In the chapter on measurement, we discussed many common methods to collect qualitative data, including:

- Interviews
- Focus Groups
- Observations
- Physical Trace
- Archival Data

While these methods of measurement might not look like they have much in common, they all may result in data that is not quantitative. Interviews and focus groups result in a large amount of narrative data, observations can result in video recordings of a variety of actions, physical trace is also visual, and archival data can be just about anything that is stored! What you might notice is that all of these methods of measurement could also be coded into quantitative data. It is up to the researcher to decide the best way to organize and (re-)code their data. 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. This results in hours of interview recordings (or pages of printed transcripts) which can be analyzed from a qualitative foundation by looking for patterns and themes, or from a quantitative foundation such that the researcher codes the data in terms of how often each participant mentions God (or a "higher power"), and then uses statistics to find out whether those who mention God more often are more successful in abstaining from alcohol then those who mention God less often. 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.

Data Analysis in Qualitative Research

Quantitative and qualitative research generally differ along several important dimensions, such as the testing of the research hypotheses, the method of data collection, as well as the method of data analysis. The chapter on analyzing has a sub-section on qualitative data to included a basic introduction to qualitative research, organizing qualitative data, options for coding qualitative data, and moving from data points to findings.

As an example, consider a study by researchers Laura Abrams and Laura Curran, who used the grounded theory approach (Glaser & Strauss, 1967) to study the experience of postpartum depression symptoms among low-income mothers (Abrams & Curran,

2009). In qualitative research using grounded theory, researchers start with the data and develop a theory or an interpretation that is "grounded in" those data. Researchers 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 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. Abrams and Curran's (2009) data were the result of unstructured interviews with 19 participants. Table **12.4.2** shows the five broad themes identified by Abrams and Curran's (2009), and the more specific repeating ideas that made up each of those themes. In the research report, Abrams and Curran (2009) provided numerous quotations from their participants as evidence of the themes, such as this one:

> 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. (Abrams & Curran, 2009, 357)

Their reporting of the results 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 **12.4.2**: Themes and Repeating Ideas in a Study of Postpartum Depression Among Low-Income Mothers. Based on Research by Abrams and Curran (2009).

Theme	Repeating ideas
Ambivalence	"I wasn't prepared for this baby," "I didn't want to have any more children."
Caregiving overload	"Please stop crying," "I need a break," "I can't do this anymore."
Juggling	"No time to breathe," "Everyone depends on me," "Navigating the maze."
Mothering alone	"I really don't have any help," "My baby has no father."
Real-life worry	"I don't have any money," "Will my baby be OK?" "It's not safe here."

The Quantitative-Qualitative "Debate"

Given their differences, it may come as no surprise that quantitative and qualitative research in the social sciences 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. However, qualitative research and analysis may have seom advantages over quantitative research because it explicitly acknowledges the dangers of not being objective, consistent, accurately measured, or generalizable. In fact, they have developed a number of frameworks for addressing these issues. A full discussion of these frameworks are beyond the scope of our discussion, but the table below shows show questions to ask yourself (as the researcher) to understand your role in the research process and to respect your participants.

Table **12.4.3**: Positionality Questions (from Secules et al., 2021)

DIMENSION OF POSITIONALITY	ORIENTING QUESTIONS	EXAMPLE TOPICS
Research topic	How does your positionality impact what research you choose to do?	• Motivation or interest • Access to topic • Emotional proximity to topic
Epistemology	How does your positionality impact how you know what you know?	• Critical lens vs. objectivity
Ontology	How does your positionality impact what you observe as a researcher?	• Insider/outsider status • Beliefs about reality
Methodology	How does your positionality impact how you make methodological choices?	• Historical traditions • Educational background
Researcher-as-instrument	How does your positionality impact how you relate to participants?	• Power and privilege • Access to participants
Communication	How does your positionality impact how you represent yourself in writing and other communication?	• Perceptions of name • Disclosures of identity • Language choices

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. 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 are complementary, and can be combined into what has come to be called mixed-methods research (Todd et al., 2004). In fact, the studies discussed previous by Lindqvist and colleagues and by Abrams and Curran both combined quantitative and qualitative approaches (Abrams & Curran, 2009; Lindvist et al., 2008). One approach to combining quantitative and qualitative research is to use qualitative research for hypothesis generation and quantitative research for hypothesis testing. Again, while a qualitative study might suggest that families who experience an unexpected suicide have more difficulty resolving the question of why, a well-designed quantitative study could test a hypothesis by measuring these specific variables in a large sample. A second approach to combining quantitative and qualitative research is referred to as triangulation. The idea is to use both quantitative and qualitative methods simultaneously to study the same general questions and to compare the results. If the results of the quantitative and qualitative methods converge on the same general conclusion, they reinforce and enrich each other. If the results diverge, then they suggest an interesting new question: Why do the results diverge and how can they be reconciled?

Using qualitative research can often help clarify quantitative results via triangulation. Trenor et al. (2008) investigated the experience of engineering students who were women. In the first phase, these engineering students were asked to complete a survey in which they rated a number of their perceptions, including their sense of belonging. Their results were statistically 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, Trenor et al. (2008) conducted interviews with the students, and in those interviews, many students from ethnic backgrounds that have been minoritized 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.

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*, 351–362.

Geertz, C. (1973). *The interpretation of cultures*. New York, NY: Basic Books.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: 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*, 26. Retrieved from http://www.biomedcentral.com/1471-244X/8/26

Secules, S., McCall, C., Mejia, J. A., Beebe, C., Masters, A. S., Sánchez-Peña, M., & Svyantek, M. (2021). Positionality practices and dimensions impact on equity research: A collaborative inquiry and call to the community. *Journal of Engineering Education, 110*, 19-43. DOI:10.1002/jee.20377

Todd, Z., Nerlich, B., McKeown, S., & Clarke, D. D. (2004) *Mixing methods in psychology: The integration of qualitative and quantitative methods in theory and practice*. London, UK: 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.

This page titled 12.4: Qualitative Research is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **6.5: Qualitative Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

12.5: Case Studies

Learning Objectives

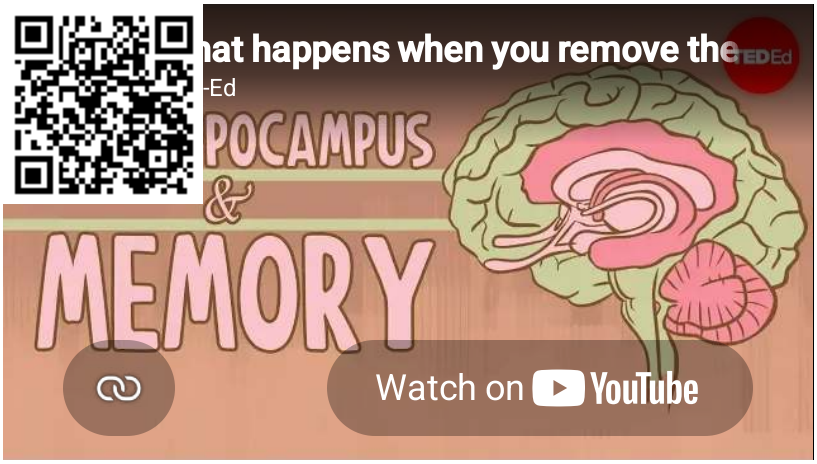
1. Describe some typical research questions that could be answered best with case studies.
2. Describe the strengths and weakness of case studies as a research method.

What are 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, 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 nonexperimental research methods, case studies tend to be more qualitative in nature. Case study methods involve an in-depth, and often a longitudinal examination of an individual. Individuals may be observed in their natural setting, or the individual may be brought into a therapist's office or a researcher's lab for study. Archival data may also be a rich source of information. Basically, because of the in-depth nature of case studies and the time commitment, almost any measurement described in the chapter on measurement can be used to gather information on the case; interviews, naturalistic observation, structured observation, psychological testing (e.g., IQ test), and/or psychobiological measurements (e.g., brain scans), and surveys could all be used to collect information on the individual. Because of this variety in the information gathered, reporting about case studies will often focus on detailed descriptions of the case rather than on statistical analyses. With that said, some quantitative data may also be included in the write-up of a case study. For instance, an individual's depression score may be compared to normative scores or their score before and after treatment may be compared.

HM is one of the most notorious case studies in psychology. HM suffered from intractable and very severe epilepsy. A surgeon localized HM's epilepsy to his medial temporal lobe and in 1953 he removed large sections of his hippocampus in an attempt to stop the seizures. The treatment was a success, in that it resolved his epilepsy and his IQ and personality were unaffected. However, the doctors soon realized that HM exhibited a strange form of amnesia, called anterograde amnesia. HM was able to carry out a conversation and he could remember short strings of letters, digits, and words. Basically, his short term memory was preserved. However, HM could not commit new events to memory. He lost the ability to transfer information from his short-term memory to his long term memory, something memory researchers call consolidation. So while he could carry on a conversation with someone, he would completely forget the conversation after it ended. This was an extremely important case study for memory researchers because it suggested that there's a dissociation between short-term memory and long-term memory, it suggested that these were two different abilities sub-served by different areas of the brain. It also suggested that the temporal lobes are particularly important for consolidating new information (i.e., for transferring information from short-term memory to long-term memory),



What happens when you remove the hippocampus? - Sam Kean. https://youtu.be/KkaXNvzE4pk

The history of psychology is filled with influential cases studies, such as Sigmund Freud's description of "Anna O." (see Note 6.1 "The Case of "Anna O."") and John Watson and Rosalie Rayner's description of Little Albert (Watson & Rayner, 1920), who

allegedly learned to fear a white rat—along with other furry objects—when the researchers repeatedly made a loud noise every time the rat approached him.

The Case of "Anna O."

Sigmund Freud used the case of a young woman he called "Anna O." to illustrate many principles of his theory of psychoanalysis (Freud, 1961). (Her real name was Bertha Pappenheim, and she was an early feminist who went on to make important contributions to the field of social work.) Anna had come to Freud's colleague Josef Breuer around 1880 with a variety of odd physical and psychological symptoms. One of them was that for several weeks she was unable to drink any fluids. According to Freud,

> She would take up the glass of water that she longed for, but as soon as it touched her lips she would push it away like someone suffering from hydrophobia.…She lived only on fruit, such as melons, etc., so as to lessen her tormenting thirst. (p. 9)

But according to Freud, a breakthrough came one day while Anna was under hypnosis.

> [S]he grumbled about her English "lady-companion," whom she did not care for, and went on to describe, with every sign of disgust, how she had once gone into this lady's room and how her little dog—horrid creature!—had drunk out of a glass there. The patient had said nothing, as she had wanted to be polite. After giving further energetic expression to the anger she had held back, she asked for something to drink, drank a large quantity of water without any difficulty, and awoke from her hypnosis with the glass at her lips; and thereupon the disturbance vanished, never to return. (p.9)

Freud's interpretation was that Anna had repressed the memory of this incident along with the emotion that it triggered and that this was what had caused her inability to drink. Furthermore, he believed that her recollection of the incident, along with her expression of the emotion she had repressed, caused the symptom to go away.

As an illustration of Freud's theory, the case study of Anna O. is quite effective. As evidence for the theory, however, it is essentially worthless. The description provides no way of knowing whether Anna had really repressed the memory of the dog drinking from the glass, whether this repression had caused her inability to drink, or whether recalling this "trauma" relieved the symptom. Like other nonexperimental research methods, case studies cannot show cause-and-effect relationships. It is also unclear from this case study how typical or atypical Anna's experience was.



Figure **12.5.1**: Anna O. "Anna O." was the subject of a famous case study used by Freud to illustrate the principles of psychoanalysis. Source: http://en.Wikipedia.org/wiki/File:Pappenheim_1882.jpg

What Case Studies are Not

As described above, case studies are different from experimental designs in several ways, including:

- Focus is on a single “case”, rather than a large sample from a population.
- Focus includes more variables collected over a longer time period. (These can be considered DVs.)
- In traditional case studies, there’s no intervention (no IV in the traditional sense).
- Case studies tend to be exploratory. The goal of case studies is in-depth understanding of an ongoing behavior. There’s no research hypothesis. There’s little control over the situation or the case’s experiences.

? Exercise 12.5.1

Based on this summary, do case studies have high or low internal validty? In other words, how good are case studies at determining cause-and-effect?

Answer

Case studies tend to have low internal validity, meaning that the information gained does not determine cause-and-effect relationships.

A quick review of case studies make them seem like small-N designs, but that's not the case. Why not? First, small-N designs tend to have only one, specific outcome (DV) that is being observed, or maybe a few. But case studies can have as many outcomes as the researcher can access. Second, small-N designs tend to have a treament, or IV, while case studies tend to gather data on what has been happening but does not intervene. Because of these differences, small-N designs are structured to attempt to provide evidence of cause-and-effect relationships, while case studies are squarely nonexperimental.

Advantages and Disadvantages of Case Studies

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 HM provided important insights into the role of the hippocampus in memory consolidation.

However, it is important to note that while case studies can provide *insights* into certain areas and variables to study, and can be useful in helping develop theories, they should never be used as *evidence* for theories. In other words, case studies can be used as inspiration to formulate theories and hypotheses, but those hypotheses and theories then need to be formally tested using more quantitative methods. The reason case studies shouldn’t be used to provide support for theories is that they are not experiments so they cannot show cause-and-effect. Case studies lack the proper controls that true experiments contain. As such, they suffer from problems with internal validity, so they cannot be used to determine causation. For instance, during HM’s surgery, the surgeon may have accidentally lesioned another area of HM’s brain (a possibility suggested by the dissection of HM’s brain following his death) and that lesion may have contributed to his inability to consolidate new information. The fact is, with case studies we cannot rule out these sorts of alternative explanations. As with all nonexperimental methods, case studies do not permit determination of causation.

Case studies suffer from problems with internal validity, but also with external validity. Because case studies are often of a single individual, and typically an individual chosen specifically because they are rare or unique in some way, researchers cannot generalize their conclusions to other individuals. With most research designs, there is a trade-off between internal and external validity. With case studies, however, there are problems with both internal validity and external validity. So there are limits both to the ability to determine causation and to generalize the results.

A final limitation of case studies is that ample opportunity exists for the theoretical biases of the researcher to bias the case description. Indeed, there have been accusations that the woman who studied HM destroyed a lot of her data that were not

published and she has been called into question for destroying contradictory data that didn't support her theory about how memories are consolidated. There is a fascinating New York Times article that describes some of the controversies that ensued after HM's death and analysis of his brain that can be found at: https://www.nytimes.com/2016/08/07/magazine/the-brain-that-couldnt-remember.html?_r=0

All research designs have situations, samples, and research questions that they may be most appropriate as the design to answer the question. It is up to you as the researcher (and you as the reader of research) to decide, for each research study, what the most appropriate research design, the most appropriate operational definition of the independent variable (or not have an independent variable), and the most appropriate operational definition of the dependent variable for this research topic on the target population.

References

Freud, S. (1961). *Five lectures on psycho-analysis*. New York, NY: Norton.

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology, 3*, 1–14.

This page titled 12.5: Case Studies is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **6.6: Observational Research** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

12.6: Summary of Nonexperimental Research

Key Takeaways

- Non-experimental research is research that lacks the manipulation of an independent variable, as well as the lack of initial equivalence between any independent variables conditions.
- There are many types of nonexperimental research designs that measure behavior to provide valuable information, even if the designs cannot show cause-and-effect relationships between variables.
- Longitudinal designs can provide useful information on the development or progress of changes in a variety of variables.
- In general, experimental research is high in internal validity, nonexperimental research is low in internal validity, and quasi-experimental research is in between.
- Qualitative research is an important alternative to quantitative research. It generally involves asking broader research questions, collecting more detailed data (e.g., interviews), and using non-statistical analyses.
- Many researchers conceptualize quantitative and qualitative research as complementary and advocate combining them.
- Case studies are typically used to collect in-depth information on a single individual; and archival research involves analyzing existing data.

Exercises

- Discussion: Is there a topic that you'd like to learn more about from a longitudinal design? How would that work?
- Discussion: With others with similar career goals, discuss what outcome you'd like to improve for your future clients. Which one of the designs that we've discussed so far makes the most sense for that outcome and your clients, describing why that design would work best?
- Discussion: What are some ways in which a qualitative study of girls who play youth baseball would likely differ from a quantitative study on the same topic? How would the data differ by interviewing girls one-on-one rather than conducting focus groups or surveys?
- Practice: Find and read a published case study in psychology. (Use *case study* as a key term in a PsycINFO search.) Then do the following:
  - Describe one problem related to internal validity.
  - Describe one problem related to external validity.
  - Generate one hypothesis suggested by the case study that might be interesting to test in a subsequent study.

This page titled 12.6: Summary of Nonexperimental Research is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- **6.7: Non-Experimental Research (Summary)** by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton is licensed CC BY-NC-SA 4.0. Original source: https://kpu.pressbooks.pub/psychmethods4e.

Index

A

ABA design
- 11.3: Small-N Research Designs and Analysis

abstract
- 5.3: How to Read a Research Article
- 7.2: Introduction to American Psychological Association (APA) Style

alpha
- 6.4: Understanding Null Hypothesis Significance Testing

alternate treatments design
- 11.3: Small-N Research Designs and Analysis

alternative hypothesis
- 6.4: Understanding Null Hypothesis Significance Testing

analysis of variance
- 6.5: Comparing Means

anonymity
- 3.4: Principles of Ethical Research

ANOVA
- 6.5: Comparing Means

APA Ethics Code
- 3.2: History of Ethics Codes
- 3.3: APA Ethics Code Standard 8

APA style
- 7.2: Introduction to American Psychological Association (APA) Style

appendix
- 5.3: How to Read a Research Article

applied behavior analysis
- 11.2: Introduction to Small-N Research

applied research
- 2.2: Goals of Science

autonomy
- 3.4: Principles of Ethical Research

B

bar graphs
- 7.5: Presenting Results

baseline
- 11.3: Small-N Research Designs and Analysis

basic research
- 2.2: Goals of Science

Behavioral measures
- 4.2: Introduction to Measuring Variables

Belmont Report
- 3.2: History of Ethics Codes
- 3.3: APA Ethics Code Standard 8

block randomization
- 8.3: Between Groups Designs and Matched Group Designs
- 8.4: Repeated Measures Design

BRUSO
- 4.3.1: Constructing Surveys

C

carryover effect
- 8.3: Between Groups Designs and Matched Group Designs
- 8.4: Repeated Measures Design

case study
- 4.6: Observations
- 4.7.2: Archival Data
- 12.5: Case Studies

categorical variable
- 2.3.3: Empirical Study

central tendency
- 6.2: Describing Variables

closing
- 5.3: How to Read a Research Article

cluster sampling
- 2.4.1: Sampling

coding
- 4.6: Observations
- 4.7.2: Archival Data
- 12.5: Case Studies

Cohen’s d
- 6.5.1: Effect Sizes

cohort effect
- 12.2: Introduction to Nonexperimental Research
- 12.3: Longitudinal and Similar Designs

complete counterbalancing
- 8.3: Between Groups Designs and Matched Group Designs
- 8.4: Repeated Measures Design

conceptual definition
- 4.2: Introduction to Measuring Variables

Concurrent validity
- 4.3.2: Reliability and Validity of Measurement

conditions
- 8.2: Initial Equivalence

confederate
- 3.4: Principles of Ethical Research

confidentiality
- 3.4: Principles of Ethical Research

confounding variable
- 8.2: Initial Equivalence

confounds
- 2.3.3: Empirical Study

consent form
- 3.2: History of Ethics Codes
- 3.3: APA Ethics Code Standard 8

Construct validity
- 4.2.1: Types of Validity

Constructs
- 4.2: Introduction to Measuring Variables

content analysis
- 4.6: Observations
- 4.7.2: Archival Data
- 12.5: Case Studies

Content validity
- 4.3.2: Reliability and Validity of Measurement

context effect
- 4.3.1: Constructing Surveys
- 8.3: Between Groups Designs and Matched Group Designs
- 8.4: Repeated Measures Design

contrast effect
- 8.3: Between Groups Designs and Matched Group Designs
- 8.4: Repeated Measures Design

control
- 8.2: Initial Equivalence

control condition
- 8.2: Initial Equivalence

Convenience sampling
- 2.3.3: Empirical Study
- 2.4.1: Sampling

Convergent validity
- 4.3.2: Reliability and Validity of Measurement

converging evidence
- 11.4: When should you use Small-N Designs?

converging operations
- 4.2: Introduction to Measuring Variables

correlation coefficient
- 2.3.4: Data Analysis

correlation matrix
- 6.6.2: Factor Analysis
- 7.5: Presenting Results

correlational research
- 12.2: Introduction to Nonexperimental Research
- 12.3: Longitudinal and Similar Designs

counterbalancing
- 8.3: Between Groups Designs and Matched Group Designs
- 8.4: Repeated Measures Design

criterion
- 4.3.2: Reliability and Validity of Measurement

criterion validity
- 4.3.2: Reliability and Validity of Measurement

criterion variable
- 6.6.2: Factor Analysis

critical values
- 6.5: Comparing Means

Cronbach’s α
- 4.3.2: Reliability and Validity of Measurement

D

debriefing
- 3.2: History of Ethics Codes
- 3.3: APA Ethics Code Standard 8

deception
- 3.2: History of Ethics Codes
- 3.3: APA Ethics Code Standard 8

Declaration of Helsinki
- 3.2: History of Ethics Codes
- 3.3: APA Ethics Code Standard 8

dependent variable
- 2.3.3: Empirical Study
- 8.2: Initial Equivalence

descriptive statistics
- 6.2: Describing Variables
- 6.3: Graphs

difference score
- 6.5: Comparing Means

diffusion of responsibility
- 8: Experimental Research
- 8.1: Prelude to Experimental Research

directionality problem
- 8.2.1: Natural Groups Research

discussion
- 5.3: How to Read a Research Article
- 7.2: Introduction to American Psychological Association (APA) Style

disguised naturalistic observation
4.6: Observations
4.7.2: Archival Data
12.5: Case Studies
disguised participant observation
4.6: Observations
4.7.2: Archival Data
12.5: Case Studies
Disproportionate stratified random sampling
2.4.1: Sampling
distribution
6.2: Describing Variables

E
Edited volumes
2.3.1: Research Literature
5.2: How to do a Literature Review
effect size
6.5.1: Effect Sizes
empirical research report
5.3: How to Read a Research Article
Empirical research reports
2.3.1: Research Literature
5.2: How to do a Literature Review
error bars
7.5: Presenting Results
ethics
3.4: Principles of Ethical Research
exempt research
3.2: History of Ethics Codes
3.3: APA Ethics Code Standard 8
expedited research
3.2: History of Ethics Codes
3.3: APA Ethics Code Standard 8
experiment
8.2: Initial Equivalence
experimental analysis of behavior
11.2: Introduction to Small-N Research
experimenter expectancy effect
8.5: Practical Considerations
external validity
2.3.3: Empirical Study
4.2.1: Types of Validity
11.4: When should you use Small-N Designs?
extraneous variables
2.3.3: Empirical Study
8.2: Initial Equivalence

F
Face validity
4.3.2: Reliability and Validity of Measurement
factor analysis
6.6.2: Factor Analysis
factorial ANOVA
6.5: Comparing Means
factorial design
10.2: Setting Up a Factorial Experiment
factorial design table
10.2: Setting Up a Factorial Experiment
faculty
For Faculty
fatigue effect
8.3: Between Groups Designs and Matched Group Designs
8.4: Repeated Measures Design
feasibility
2.3.2: Research Questions
Federal Policy for the Protection of Human Subjects
3.2: History of Ethics Codes
3.3: APA Ethics Code Standard 8
field experiments
2.3.3: Empirical Study
field study
2.3.3: Empirical Study
figures
7.5: Presenting Results
focus groups
4.4: Interviews and Focus Groups
12.4: Qualitative Research
Frequency Polygons
6.3.2.1: Graphing Quantitative Data- Line Graphs
frequency table
6.2: Describing Variables

G
greater than minimal risk research
3.2: History of Ethics Codes
3.3: APA Ethics Code Standard 8
grounded theory
4.4: Interviews and Focus Groups
12.4: Qualitative Research
group research
11.2: Introduction to Small-N Research

H
Hawthorne effect
4.6: Observations
4.7.2: Archival Data
12.5: Case Studies
histogram
6.2: Describing Variables
history
9.2: Quasi-Experimental Designs with One Group
9.4: Threats to Internal Validity

I
independent variable
2.3.3: Empirical Study
8.2: Initial Equivalence
inferential statistics
2.3.4: Data Analysis
informed consent
3.4: Principles of Ethical Research
initial manuscript
2.3.6: Sharing Your Research
Institutional Review Board
3.2: History of Ethics Codes
3.3: APA Ethics Code Standard 8
instrumentation
9.2: Quasi-Experimental Designs with One Group
9.4: Threats to Internal Validity
interaction
10.3: Interpreting the Results of a Factorial Experiment
interestingness
2.3.2: Research Questions
internal consistency
4.3.2: Reliability and Validity of Measurement
internal validity
2.3.3: Empirical Study
4.2.1: Types of Validity
interval level
4.2: Introduction to Measuring Variables
interviews
4.4: Interviews and Focus Groups
12.4: Qualitative Research
introduction
5.3: How to Read a Research Article
7.2: Introduction to American Psychological Association (APA) Style
IRB
3.2: History of Ethics Codes
3.3: APA Ethics Code Standard 8

J
justice
3.2: History of Ethics Codes
3.3: APA Ethics Code Standard 8

K
Kitty Genovese
8: Experimental Research
8.1: Prelude to Experimental Research
kurtosis
6.3.2: Describing the Shape of Graphs of Quantitative Data

L
laboratory study
2.3.3: Empirical Study
latency
11.3: Small-N Research Designs and Analysis
leptokurtic
6.3.2: Describing the Shape of Graphs of Quantitative Data
level
11.3: Small-N Research Designs and Analysis
levels of measurement
4.2: Introduction to Measuring Variables
line graph
6.3.2.1: Graphing Quantitative Data- Line Graphs
Line graphs
7.5: Presenting Results
linear relationships
6.5.1: Effect Sizes
literature review
5.3: How to Read a Research Article
longitudinal studies
12.2: Introduction to Nonexperimental Research
12.3: Longitudinal and Similar Designs

M
main effect
10.3: Interpreting the Results of a Factorial Experiment
manipulate
8.2: Initial Equivalence
manipulation check
8.5: Practical Considerations
maturation
9.2: Quasi-Experimental Designs with One Group
9.4: Threats to Internal Validity

mean
2.3.4: Data Analysis
6.2: Describing Variables
mean squares between groups
6.5: Comparing Means
mean squares within groups
6.5: Comparing Means
median
6.2: Describing Variables
Method
7.2: Introduction to American Psychological Association (APA) Style
method section
5.3: How to Read a Research Article
mixed factorial design
10.2: Setting Up a Factorial Experiment
mode
2.3.4: Data Analysis
6.2: Describing Variables
monograph
2.3.1: Research Literature
5.2: How to do a Literature Review
multiple regression
6.6.2: Factor Analysis
mundane realism
4.2.1: Types of Validity

N
naturalistic observation
4.6: Observations
4.7.2: Archival Data
12.5: Case Studies
negative relationship
8.2.1: Natural Groups Research
nominal level
4.2: Introduction to Measuring Variables
nonequivalent groups design
9.3: Quasi-Experimental Designs with Two or More Groups
Nonlinear relationships
6.5.1: Effect Sizes
null hypothesis
6.4: Understanding Null Hypothesis Significance Testing
Null hypothesis testing
6.4: Understanding Null Hypothesis Significance Testing
Nuremberg Code
3.2: History of Ethics Codes
3.3: APA Ethics Code Standard 8

O
observational research
4.6: Observations
4.7.2: Archival Data
12.2: Introduction to Nonexperimental Research
12.3: Longitudinal and Similar Designs
12.5: Case Studies
opening
5.3: How to Read a Research Article
operational definition
2.3.3: Empirical Study
4.2: Introduction to Measuring Variables
operationalization
4.2.1: Types of Validity
oral presentation
2.3.6: Sharing Your Research
order effect
8.3: Between Groups Designs and Matched Group Designs
8.4: Repeated Measures Design
ordinal level
4.2: Introduction to Measuring Variables
outcome variable
6.6.2: Factor Analysis
outlier
6.2: Describing Variables

P
p value
6.4: Understanding Null Hypothesis Significance Testing
parameters
6.4: Understanding Null Hypothesis Significance Testing
partial correlation
6.6.2: Factor Analysis
participant observation
4.6: Observations
4.7.2: Archival Data
12.5: Case Studies
Pearson’s Correlation Coefficient
8.2.1: Natural Groups Research
Pearson's r
8.2.1: Natural Groups Research
percentile rank
6.2: Describing Variables
physiological measures
4.2: Introduction to Measuring Variables
pilot test
8.5: Practical Considerations
placebo
8.2: Initial Equivalence
placebo control condition
8.2: Initial Equivalence
placebo effect
8.2: Initial Equivalence
population
2.3.3: Empirical Study
positive relationship
8.2.1: Natural Groups Research
post hoc comparisons
6.5: Comparing Means
poster
2.3.6: Sharing Your Research
poster session
2.3.6: Sharing Your Research
posttest only nonequivalent groups design
9.3: Quasi-Experimental Designs with Two or More Groups
Practical significance
6.4: Understanding Null Hypothesis Significance Testing
practice effect
8.3: Between Groups Designs and Matched Group Designs
8.4: Repeated Measures Design
Predictive validity
4.3.2: Reliability and Validity of Measurement
predictor variable
6.6.2: Factor Analysis
privacy
3.4: Principles of Ethical Research
Probability sampling
2.4.1: Sampling
professional conferences
2.3.6: Sharing Your Research
Professional journals
2.3.1: Research Literature
5.2: How to do a Literature Review
Proportionate stratified random sampling
2.4.1: Sampling
protocol
3.2: History of Ethics Codes
3.3: APA Ethics Code Standard 8
psychological realism
4.2.1: Types of Validity
psychometrics
4.2: Introduction to Measuring Variables
PsycINFO
2.3.1: Research Literature
5.2: How to do a Literature Review
Publication Manual of the American Psychological Association
7.2: Introduction to American Psychological Association (APA) Style

Q
qualitative research
4.4: Interviews and Focus Groups
12.4: Qualitative Research
quantitative research
4.4: Interviews and Focus Groups
12.4: Qualitative Research
quantitative variable
2.3.3: Empirical Study
Quota sampling
2.4.1: Sampling

R
random assignment
8.3: Between Groups Designs and Matched Group Designs
8.4: Repeated Measures Design
random counterbalancing
8.3: Between Groups Designs and Matched Group Designs
8.4: Repeated Measures Design
randomized clinical trial
8.2: Initial Equivalence
range
2.3.4: Data Analysis
6.2: Describing Variables
rating scale
4.3.1: Constructing Surveys
ratio level
4.2: Introduction to Measuring Variables
reactivity
4.6: Observations
4.7.2: Archival Data
12.5: Case Studies
reference citation
7.2: Introduction to American Psychological Association (APA) Style

references
7.2: Introduction to American Psychological Association (APA) Style
regression
6.6.2: Factor Analysis
Regression to the mean
9.2: Quasi-Experimental Designs with One Group
9.4: Threats to Internal Validity
reject the null hypothesis
6.4: Understanding Null Hypothesis Significance Testing
reliability
4.3.2: Reliability and Validity of Measurement
respect for persons
3.2: History of Ethics Codes
3.3: APA Ethics Code Standard 8
respondents
4.3: Introduction to Surveys
restriction of range
6.5.1: Effect Sizes
8.2.1: Natural Groups Research
Results
7.2: Introduction to American Psychological Association (APA) Style
results section
5.3: How to Read a Research Article
retain the null hypothesis
6.4: Understanding Null Hypothesis Significance Testing
reversal design
11.3: Small-N Research Designs and Analysis
review articles
2.3.1: Research Literature
2.3.6: Sharing Your Research
5.2: How to do a Literature Review

S

Sample
2.3.3: Empirical Study
Sampling Bias
2.4.1: Sampling
sampling error
6.4: Understanding Null Hypothesis Significance Testing
sampling frame
2.4.1: Sampling
scatterplots
7.5: Presenting Results
Scholarly books
2.3.1: Research Literature
5.2: How to do a Literature Review
Simple effects
10.3: Interpreting the Results of a Factorial Experiment
Simple random sampling
2.3.3: Empirical Study
2.4.1: Sampling
simple regression
6.6.2: Factor Analysis
single factor multi level design
8.2: Initial Equivalence
skew
6.3.2: Describing the Shape of Graphs of Quantitative Data
Skewed
6.2: Describing Variables
skewness
6.3.2: Describing the Shape of Graphs of Quantitative Data
Snowball sampling
2.4.1: Sampling
social validity
11.2: Introduction to Small-N Research
spontaneous remission
9.2: Quasi-Experimental Designs with One Group
9.4: Threats to Internal Validity
spreading interactions
10.3: Interpreting the Results of a Factorial Experiment
standard deviation
2.3.4: Data Analysis
6.2: Describing Variables
standard error
7.5: Presenting Results
statistical control
6.6.2: Factor Analysis
Statistical validity
4.2.1: Types of Validity
statistically significant
2.3.4: Data Analysis
6.4: Understanding Null Hypothesis Significance Testing
statistics
6.4: Understanding Null Hypothesis Significance Testing
steady state strategy
11.3: Small-N Research Designs and Analysis
stratified random sampling
2.4.1: Sampling
structured observation
4.6: Observations
4.7.2: Archival Data
12.5: Case Studies
subject pool
8.5: Practical Considerations
survey research
4.3: Introduction to Surveys
switching replication with treatment removal design
9.3: Quasi-Experimental Designs with Two or More Groups
symmetrical
6.2: Describing Variables

T

test statistic
6.5: Comparing Means
testing
9.2: Quasi-Experimental Designs with One Group
9.4: Threats to Internal Validity
theoretical article
2.3.1: Research Literature
5.2: How to do a Literature Review
theoretical articles
2.3.6: Sharing Your Research
theoretical narrative
4.4: Interviews and Focus Groups
12.4: Qualitative Research
title page
5.3: How to Read a Research Article
7.2: Introduction to American Psychological Association (APA) Style
to explain
2.2: Goals of Science
to predict
2.2: Goals of Science
todescribe
2.2: Goals of Science
treatment
8.2: Initial Equivalence
treatment condition
8.2: Initial Equivalence
trend
11.3: Small-N Research Designs and Analysis
triangulation
4.4: Interviews and Focus Groups
12.4: Qualitative Research
type I error
2.3.4: Data Analysis
type II error
2.3.4: Data Analysis

U

undisguised naturalistic observation
4.6: Observations
4.7.2: Archival Data
12.5: Case Studies
undisguised participant observation
4.6: Observations
4.7.2: Archival Data
12.5: Case Studies

V

validity
4.3.2: Reliability and Validity of Measurement
variability
6.2: Describing Variables
Variable
2.3.3: Empirical Study
variance
2.3.4: Data Analysis
6.2: Describing Variables
visual inspection
11.3: Small-N Research Designs and Analysis

Z

z score
6.2: Describing Variables

Glossary

Abstract | A short summary (approximately 200 words) of a research article. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Appendix | An optional section at the end of an APA Style report used to present important
material, such as tables, figures, full surveys, or pictures or screen shots. [CC-BY Michelle Oja]

Applied Behavioral Analysis | During therapy or training, the use of small-N designs to produce socially valid effects. [CC-BY Michelle Oja]

Applied Research | A type of study conducted primarily to solve a real-life problem. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Archival Data | Existing data that were collected or created for some other purpose. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Assent | Assent means a child's affirmative agreement to participate in research. Mere failure to object should not, absent affirmative agreement, be construed as assent. [Part 46 of Title 45 of the Code of Federal Regulations]

Authority | A way of gaining knowledge that relies on leaders and people in charge. [CC-BY Michelle Oja]

Baseline | A condition in which the dependent variable is measured in the absence of any treatment. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Basic Research | A type of study conducted primarily for the sake of achieving a more detailed and accurate understanding of human behavior, without necessarily trying to address any particular practical problem. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Behavioral Measure | When some aspect of a participant's behavior is observed and recorded. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Between Groups Design | An experiment in which each participant is randomly assigned into only one IV condition. [CC-BY Michelle Oja]

Case Study | A detailed description of an individual case. Cases are usually chosen because they are rare. Cases can be any single social unit, such as an individual person, a family, a specific town or city, or a specific company. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Categorical variable | Type of variable that has different values to represent different qualities or kinds This is the same as qualitative variables. [CC-BY Michelle Oja]

Closed-Ended | A question and provides a set of response options for participants to choose from. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Coding | In terms of data collection, target behaviors are specified ahead of time and observers watch for and record those specific behaviors. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Coercion | Compelling someone to act against their best interests in such a way that they feel like they have little choice but to comply. [CC-BY Michelle Oja]

Cohort Effect | The outcome of shared characteristics over time among individuals who are grouped by a shared temporal experience, such as year of birth, or common life experience, such as time of exposure to radiation. [CC-BY-SA by Wikipedia]

Complete Research Hypothesis | A prediction about how a DV will be affected by each level of the IV. [CC-BY Michelle Oja]

Conceptual definition | A description of the behaviors and internal processes that make up a construct. This contrasts with an operational definition. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Conditions | Levels of the independent variable in an experiment; the IV groups. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Confederate | In research studies, a person who plays the role of a participant but is really part of the research team or an actor; these people may interact in scripted ways with real participants. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Confound | A variable that systematically varies along with the independent variable. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Construct | A psychological variable that cannot be observed
directly. These include personality traits,
emotional states, attitudes, and abilities. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Content Analysis | A family of systematic approaches to identifying patterns in complex data, such as written or verbal responses. [CC-BY Michelle Oja]

Content Validity | To measure the accuracy of a survey, the extent to which a measure assess all target areas of the construct of interest. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Control Group | In an experiment with random assignment, this group does not receive any treatment to use as a comparison for the intervention condition. [CC-BY Michelle Oja]

Convenience Sampling | Collecting data on individuals who are easy for the researcher to access and incentivize to participate in the study. [CC-BY Michelle Oja]

Convergent Validity | The extent to which scores on a survey are related to surveys of constructs that would be expect to be related to show that a survey is accurate. [CC-BY Michelle Oja]

Copy Manuscript | A type of report that will be submitted to a professional journal for publication. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Correlation Coefficient | A statisticcal analysis that describes the strength and direction of the relationship between two quantitative variables. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Counterbalancing | Systematically varying the order of conditions across participants when participants are in all conditions. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Criterion Validity | To measure the accuracy of a survey, the extent to which scores on a measure are related to variables (criteria) that would be expect to be related. This works best when the criteria are not other surveys. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Cross-Sectional Design | A research design that involves measurement of the same variable for individuals who are at varying points of progress or development. [CC-BY Michelle Oja]

Cross-Sequential Design | A research design that combines both a longitudinal design and a cross-sectional design in which a researcher chooses a smaller time window than the desired longitudinal design to study multiple individuals of different starting ages (cross-sections). [CC-BY-SA by Wikipedia]

Debrief | The process of informing research participants *after* their participation in a study of the purpose of the study. During this process, researchers should reveal any deception, as well as minimize any harm that
might have occurred. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Dependent Variable | The measured variable that the experimenter think is the effect (the thing that the IV changes). This is the outcome variable, the thing that we're trying to improve. [CC-BY Michelle Oja]

Descriptive Statistics | A set of techniques for summarizing and displaying quantitative data. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Discriminant Validity | The extent to which scores on a survey are *not* related to surveys of constructs that would be expect to not be related to show that a survey is accurate. [CC-BY Michelle Oja]

Discussion Section | The final major section in the main body of an APA Style empirical research report, sometimes called a Conclusion. It typically includes a summary of the research question and findings, a discussion of implications of the study, limitations of the study, and suggestions for future research. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Disguised Observation | Method of data collection in which observations are made as unobtrusively as possible so that participants are not aware that they are being studied. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Distribution | The way the scores on a variable are distributed across the levels of that variable. This is often shown in a frequency chart. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Effect Size | Any number representing the strength of a statistical relationship, but especially a computation to standardize the effect. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Empirical Research Reports | A type of research literature that reports on studies in which the authors analyzed data [CC-BY Michelle Oja]

Empiricism | A way of gaining knowledge by collecting data. This differs from the scientific method in that the data collection may not be systematic or to test a hypothesis. [CC-BY Michelle Oja]

Epistemology | The branch of philosophy that examines the nature, origin, and limits of knowledge. Also called "theory of knowledge", it explores different types of knowledge. [CC-BY-SA 4.0 Wikipedia]

Experimenter Expectancy Effects | The impact that the researcher's thoughts on the results of the study (unconsciously) affects the researcher's behavior, which then influences the results in the expected direction. [CC-BY Michelle Oja]

External Validity | The degree to which the findings can be generalized to other circumstances or settings. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Extraneous variable | Any variable other than the independent variable(s) and the dependent variable(s) [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Face Validity | The extent to which a survey appears “on its face” to measure the construct of interest. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Factor Analysis | A complex statistical technique that organizes several variables into clusters where there are strong correlations among the variables within a cluster but weak correlations among the variables between clusters. Each cluster is interpreted as representing a different underlying variable or factor. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Factorial Design | A type of research design with more than one IV in which 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. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Field Study | Type of research done in the real world, in the a natural environment. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Focus Group | A small group of people who participate together in a facilitated conversation focused on a particular topic or issue. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Frequency Polygon | A graph that displays the distribution of a quantitative variable in a line. The x-axis represents the values of the variable, and the y-axis represents the frequency of each score. [CC-BY Michelle Oja]

Functional Imaging | A variety of tools that measure how the brain works through measuring electrical or chemical activity in the brain. [CC-BY Michelle Oja]

Hawthorne Effect | A situation that occurs when research participants, either intentionally or unintentionally, alter their behaviors because they know they are being studied. [CC-NC-SA]

Histogram | A graph for displaying the distribution of a quantitative variable. The x-axis represents the values of the variable, and the y-axis represents the frequency of each score. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Independent Variable | The variable that the experimenter manipulates, and believes is the cause of changes in the outcome variable. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Inferential Statistics | Analysis used to make generalizations from the sample data to the population of interest. [CC-BY Michelle Oja]

Informed Consent | The process of obtaining and documenting participants’ agreement to be in a study, having explained to them of everything that might reasonably be expected to affect their decision. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Initial Equivalence | A fundamental feature of experimental studies to show cause-and-effect by creating groups, or conditions, that start the study with similar characteristics. [CC-BY Michelle Oja]

Interaction | How the levels of two or more IVs *jointly* affect a DV; when one IV *interacts* with the other IV to affect the DV. [CC-BY Michelle Oja]

Internal Consistency | The extent to which the items on a multiple-item survey are consistent with each other. This is usually measured by Cronbach's Alpha. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Internal Validity | In terms of an experiment (not a survey), this is the degree to which we can confidently infer a causal relationship between variables. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Interrater Reliability | The extent to which different observers are consistent in their judgments. Common Synonyms include:

- inter-coder
- inter-observer
- inter-scorer
- Inter-judge

[CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Interrupted Time-Series Design | A type of quasi-experimental design that is a variant of the One-Group Pretest-Posttest Design in which a set of measurements taken at intervals over a period of time before and after a treatment is implemented. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Interrupted Time-Series Design with Nonequivalent Groups | A type of quasi-experimental design in which a set of measurements are taken at intervals over a period of time both before and after an intervention of interest in two or more nonequivalent groups. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Interval Level | Level of measurement that involves assigning scores using numerical scales in which intervals have the same interpretation throughout. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Intuition | A way of gaining knowledge through your unconscious beliefs. [CC-BY Michelle Oja]

Laboratory Study | Type of research done in a controlled environment, like a laboratory, classroom, or other situation in which the experimenter has control. [CC-BY Michelle Oja]

Likert Scale | An ordered set of response options assessing participants' level of agreement to statements. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Literature Review | Finding and reading previous research and summaries about a specific topic. [CC-BY Michelle Oja]

Longitudinal Design | A research design that involves repeated observations of the same variables over long periods of time. [CC-BY-SA by Wikipedia]

Main Effect | In a factorial design, the effect of one independent variable averaged across levels of all other independent variables. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Manipulation Check | A measure of a manipulated independent variable, isually done at the end of the
procedure, to confirm that the independent variable was successfully manipulated. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Matched Groups Design | An experiment in which each participant is paired with another participant with similar score on a variable that is not the IV nor the DV. The participants are then randomly assigned into only one of the two IV conditions. This can be modified to trios, quads, etc., if there are three, four, or more IV conditions. [CC-BY Michelle Oja]

Mean | The mathematical average of a distribution of scores, found by summing all of the scores and then dividing by the total number of scores. [CC-BY Michelle Oja]

Median | The midpoint, or exact middle, of a distribution of scores. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Method Section | The part of an APA Style empirical research report in which the method is described in detail. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Mode | The most frequently occurring score in a distribution. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Multiple Regression | A statistical technique that describes the relationship between two or more predictor variables and a single outcome variable in terms of an equation that shows the separate contribution of each predictor variable to the outcome variable. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Natural Groups Design | Research study that compares groups that were not randomly assigned into the IV conditions. This type of design is sometimes called a posttest-only nonequivalent group design. [CC-BY Michelle Oja]

Naturalistic Observation | Method of data collection that involves observing people’s behavior in the environment in which it typically occurs. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Negative skew | The scores are bunched to the right, and the thin tail is pointing to the left. [CC-BY Michelle Oja]

Neurotransmitter | Chemical messenger of the nervous system, especially between neurons. [CC-BY OpenStax]

Nominal Level | One level of measurement; see Qualitative Variable

Non-Probability Sampling | A method to develop a sample in which the researcher cannot specify the probability that each member of the population will be selected for the sample. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Non-Response Bias | The possibility that survey non-responders differ from survey responders in systematic ways, which could lead to inaccurate results. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Nonexperimental Research | Research that lacks the manipulation of an independent variable and does not have initial equivalence between groups. This reduces internal validity so that the study cannot show cause-and-effect relationships. [CC-BY Michelle Oja]

Null Hypothesis | Stating that nothing is happening statistically between variables. This can be a statement that the means of an outcome variable are similar between groups, or that there is no linear relationship between two variables, or that there is no pattern of relationship between categories. [CC-BY Michelle Oja]

Null Hypothesis Significance Testing | A formal approach that uses probability to decide between two interpretations (null hypothesis compared to a research hypothesis) of a statistical relationship in a sample. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Observational Research | Several different types studies, usually which are non-experimental, in which behavior is systematically observed and recorded. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Observer Bias | The tendency of observers to not see what is there, but instead to see what they expect or want to see. [CC-BY-SA by Wikipedia]

One-Group Pretest-Posttest Design | A quasi-experimental design in which the dependent variable is measured once before the treatment is implemented and once after it is implemented for one group. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Ongoing Equivalence | A fundamental feature of experimental studies to show cause-and-effect by attempting to have the only variable that differs between groups, or conditions, throughout the study is the independent variable. [CC-BY Michelle Oja]

Open-Ended | A question that allows participants to respond in whatever way they want. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Operational Definition | A description of the variable in terms of precisely how it is to be measured. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Order Effect | A threat to internal validity (showing cause-and-effect) when participants' responses in the conditions of a repeated measures design are affected by the order of conditions to which they were expose.d [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Ordinal Level | Level of measurement in which levels have an order; a type of variable that can be put in numerical order. The variables are in ranks (first, second, third, etc.). [CC-BY Michelle Oja]

Outlier | An extreme score, a score that seems much higher or much lower than most of the other scores (There is a technical way to calculate whether a score is an outlier or not, but you don't need to know it.) [CC-BY Michelle Oja]

p-value | A number that tells us the estimated probability that the null hypothesis is true. [CC-BY Michelle Oja]

Parameter | Statistic describing characteristics of the population (usually mean and standard deviation of the population) [CC-BY Michelle Oja]

Partial Correlation | A statistical technique in which the linear relationship between two quantitative variables is tested while statistically controlling for one or more potential additional variables. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Participant Observation | Method of data collection in which researchers become active participants in the group or situation they are studying. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Peer Review | In research publishing, a system of evaluation in which experts in the field volunteer to review research manuscript drafts to determine if the manuscript should be published, revised, or rejected. [CC-BY Michelle Oja]

Physical Trace | remnants of past behavior [CC-BY-SA by Wikipedia]

Physiological Measure | A measure that involves recording a biological variable, such as heart rate, blood pressure, galvanic skin response, or brain activity. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Placebo | A simulated or fake treatment that lacks any active ingredient or element that should make it effective. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Placebo Effect | The response of participants to a simulated or fake treatment in relation to participants who received no treatment nor any fake treatment. [CC-BY-SA by Wikipedia]

Population | The largest group that the researcher is interested in drawing conclusions about. Ideally, this is also the largest group that the sample can adequately represent. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton and CC-BY-NC-SA Michelle Oja]

Positive skew | The scores are bunched to the left, and the thin tail is pointing to the right. [CC-BY Michelle Oja]

Post-Hoc Comparison | Statistical comparison of pairs of group means following a statistically significant ANOVA result. There are several options to choose from that vary by mathematical complexity and preference for risking Type I Errors or Type II Errors. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Posttest | Measurement of an outcome variable (DV) after implementation of a treatment or event. This term is usually only used when paired with a pretest. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Posttest-Only Nonequivalent Group Desing | See "Natural Groups Design"

Practical Significance | The importance of a research result in the real-world; research results can be statistically significant without having real-world effects. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Predictor variable | The variable that the experimenter believes is the cause of changes in the outcome variable, but which the experimenter did not manipulate. This is often called an independent variable, but this is imprecise as there is no manipulation by the experimenter. [CC-BY Michelle Oja]

Pretest | Measurement of an outcome variable (DV) before implementation of a treatment or event. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Pretest-Posttest Design with Switching Replication Design | A quasi-experimental design in which nonequivalent groups are administered a pretest of the dependent variable, then one group receives a treatment while a nonequivalent 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. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Pretest-Posttest Nonequivalence Group Design | A quasi-experimental design in which a treatment group is given a pretest, receives a treatment, and then is given a posttest, while at the same time there is a nonequivalent control group that is given a pretest, does not receive the treatment, and then is given a posttest. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Primary Source | An original document; in research, an article in which the authors collected and analyzed their own data [CC-BY Michelle Oja]

Probability Sampling | A method to develop a sample in which the researcher can specify the probability that each member of the population will be selected for the sample. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Psychobiological Measures | See "physiological measures".

Psychometrics | Measurement of psychological variables and constructs. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Qualitative Research | Research that typically involves formulating broad research questions, collecting large amounts of data from a small number of participants, and summarizing the data using nonstatistical techniques. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Qualitative variable | Type of variable that has different values to represent different *categories* or kinds This is the same as the nominal scale of measurement. [CC-BY Michelle Oja]

Quantitative variable | Type of variable that is measured with some sort of scale that uses numbers that measure something. [CC-BY Michelle Oja]

Random Assignment | The placement of participants to different IV conditions based on a probabilistic method to attain a random distribution of characteristics across the different IV conditions. [CC-BY Michelle Oja]

Random Sample | Choosing participants in such a way that each individual in the population has an equal probability of being selected for the sample. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Random Selection | See "Random Sample"

Randomized Clinical Trial | An experiment designed to test the effectiveness of a psychological or medical treatment that utilizes random assignment into a treatment conditions. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Range | A measure of dispersion that measures the distance between the highest and lowest scores in a distribution. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Range | The difference between the highest score and the lowest score in a distribution of quantitative data. [CC-BY Michelle Oja]

Rating Scale | An ordered set of response options to a closed-ended question. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Ratio Level | 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. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Rationalism | A way of gaining knowledge through logical rules like "if/then" statements to reach a conclusion. [CC-BY Michelle Oja]

Reactivity | When a measure changes participants' behavior. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Reflexivity | In the context of social science research, the process of reflecting on one's own perspective and positionality and how this perspective and positionality shape "research design, data collection, analysis, and knowledge production" (Hsiung, 2008) [CC BY-NC-SA 4.0 by Mikaila Mariel Lemonik Arthu]

Regression | An advanced statistical technique that allows researchers to estimate one variable (outcome) given another variable (predictor). [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Reliability | Consistency of measurement. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Repeated Measures Design | An experiment in which each participant is tested on the DV in all conditions of the IV, with the order of the IV conditions randomized for each participant. [CC-BY Michelle Oja]

Replication | The process of conducting an empirical study again to see if the same results are observed. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Research Hypothesis | A specific prediction about how variables interact with each other that can be tested [CC-BY Michelle Oja]

Respondent | A term often used to refer to a participant in survey research. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Result Section | The part of an APA Style empirical research report in which the outcome of the analyses are
described in detail. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Review Articles | A type of research literature that summarizes others' previous research [CC-BY Michelle Oja]

Sample | People who participate in a study; the smaller group that the data is gathered from. [CC-BY Michelle Oja]

Sampling Bias | When the results of a study are less likely to be accurate because a sample is selected in such a way that it is not representative of the entire population. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Sampling Error | Random variation in a statistic from sample to sample. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Sampling Frame | A list of all the members of the population used to recruit a sample from. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Scientific Method | A way of gaining knowledge based on structured data collection that tests hypotheses or theories. [CC-BY Michelle Oja]

Secondary Sources | A variety of documents that summarize others' research or thoughts [CC-BY Michelle Oja]

Self-report measure | When participants provide information on their own thoughts, feelings, and actions. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Self-Selection Sampling | Individuals choose to take part in the research on their own accord, without being approached by the researcher directly, [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Single-Subject Design | See "Small-N Design"

Skew | A distribution in which many scores are bunched up to one side, and there are only a few scores on the other side. [CC-BY Michelle Oja]

Small-N Design | A type of quantitative research that involves recording specific behavior(s) of a small number of participants in alternating conditions of baseline to treatment and back. [CC-BY Michelle Oja]

Snowball Sampling | Choosing additional participants through referral from existing research participants. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Social Validity | The extent to which a study focuses on an intervention that has a substantial effect on an important behavior and can be implemented reliably in the real-world contexts in which that behavior occurs. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Split-Half Correlation | A measure of reliability that assesses the relationship between scores based on one half of the items on a multiple-item survey and scores based on the other half of the items. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Standard Deviation | A measure of dispersion that measures the average distance of scores from the mean. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Statistic | The results of statistical analyses [CC-BY Michelle Oja]

Statistical analysis | Procedures to organize and interpret numerical information [CC-BY Michelle Oja]

Statistical Control | In complex correlational research, accounting for third variables by measuring them and including them in the analysis. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Statistical Validity | Accurate interpretation of results of the most appropriate statistical analysis of the data. [CC-BY Michelle Oja]

Statistically Significant | When statistical analyses show that an effect is probably not due to random chance; therefore the results likely represent a real effect in the population. Traditionally, this is when the probability that the null hypothesis is true is less than 5% ($p<.05$). [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton and CC-BY-NC-SA Michelle Oja]

Stratified Random Sampling | Identifying potential participants by dividing the population into different subgroups or "strata" and then a random sample is taken from each "stratum [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Structural Imaging | A variety of tools that take pictures of the physical features of the brain. [CC-BY Michelle Oja]

Structured Observation | Method of data collection that involves observing people's behavior in an environment that the researcher has structured in some one, often adding a specific set of interventions. [CC-BY Michelle Oja]

Switching Replication with Treatment Removal design | A quasi-experimental design in which nonequivalent groups are administered a pretest of the dependent variable, then one group receives a treatment while a nonequivalent control group does not receive a treatment, the dependent variable is assessed again. The next stage is to remove the treatment from the first group and add the treatment to the control group. Finally, the dependent variable is assessed one last time.
[CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Test-Retest Reliability | A measure of consistency in which an assessment is taken twice by the same participants; consistency is found when the results do not change (when they are not expected to change). [CC-BY Michelle Oja]

Theory | A coherent explanation or interpretation of one or more phenomena [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Third-Variable Problem | The problem of knowing whether two variables, X and Y, are statistically related because one causes the other or because some third variable, Z, causes both X and Y. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Title Page | The first page of an APA-style manuscript, containing the title, author names, and affiliations. Some versions of APA Style include other components, as well, such as an author note. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

True Experiment | Type of study designed specifically to answer the question of whether there is a causal relationship between variables. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Type I Error | Based on your sample's results, you reject a true null hypothesis; you say that something is happening when there probably isn't anything happening in the population. [CC-BY Michelle Oja]

Type II Error | Based on your sample's results, you retain a false null hypothesis; you say that nothing is happening when something is probably happening in the population. [CC-BY Michelle Oja]

Undisguised Observation | Method of data collection in which participants are made aware of the researcher presence and monitoring of their behavio. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Unobtrusive Research | Methods of collecting data that don't interfere with the participants under study. [CC-NC-SA]

Validity | The accuracy in which a measure assesses the target construct; you measure the variable that you intend to measure. [CC-BY Michelle Oja]

Variable | A quantity or quality that varies across people or situations [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Wait-List Control Group | In experiments, the condition in which participants are put on a waitlist to receive the treatment after the study is completed. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Detailed Licensing

Overview

Title: Research Methods for the Social and Behavioral Sciences

Webpages: 125

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 46.4% (58 pages)
- Undeclared: 25.6% (32 pages)
- CC BY-SA 4.0: 24% (30 pages)
- Public Domain: 1.6% (2 pages)
- CC BY 4.0: 0.8% (1 page)
- Unknown License: 0.8% (1 page)
- CC BY-NC-SA 3.0: 0.8% (1 page)

By Page

- Research Methods for the Social and Behavioral Sciences - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Acknowledgements - *Undeclared*
    - Licensing - *Undeclared*
    - Licensing - *Undeclared*
    - For Faculty - *CC BY 4.0*
  - 1: Introduction to Research Methods - *CC BY-SA 4.0*
    - 1.1: Prelude to Research Methods - *CC BY-SA 4.0*
    - 1.2: Methods of Knowing - *CC BY-SA 4.0*
    - 1.3: Why are you taking this course? - *CC BY-SA 4.0*
    - 1.4: Summary of Introduction to Research Methods of the Social Sciences - *CC BY-SA 4.0*
  - 2: Overview of the Scientific Method - *CC BY-SA 4.0*
    - 2.1: Prelude to the Scientific Method - *CC BY-SA 4.0*
    - 2.2: Goals of Science - *CC BY-SA 4.0*
    - 2.3: A Model of Scientific Research - *CC BY-SA 4.0*
      - 2.3.1: Research Literature - *CC BY-SA 4.0*
      - 2.3.2: Research Questions - *CC BY-SA 4.0*
      - 2.3.3: Empirical Study - *CC BY-SA 4.0*
      - 2.3.4: Data Analysis - *CC BY-SA 4.0*
      - 2.3.5: Conclusions - *CC BY-SA 4.0*
      - 2.3.6: Sharing Your Research - *CC BY-NC-SA 4.0*
    - 2.4: Populations and Samples - *CC BY-SA 4.0*
      - 2.4.1: Sampling - *CC BY-SA 4.0*
    - 2.5: Summary of the Scientific Method - *CC BY-SA 4.0*
  - 3: Ethics in Research - *Undeclared*
    - 3.1: Prelude to Research Ethics - *Undeclared*
    - 3.2: History of Ethics Codes - *CC BY-NC-SA 4.0*
    - 3.3: APA Ethics Code Standard 8 - *CC BY-NC-SA 4.0*
    - 3.4: Principles of Ethical Research - *CC BY-NC-SA 4.0*
    - 3.5: Summary of Research Ethics - *CC BY-NC-SA 4.0*
  - 4: Measuring Variables - *Undeclared*
    - 4.1: Prelude to Psychological Measurement - *Undeclared*
    - 4.2: Introduction to Measuring Variables - *CC BY-NC-SA 4.0*
      - 4.2.1: Types of Validity - *CC BY-SA 4.0*
    - 4.3: Introduction to Surveys - *CC BY-NC-SA 4.0*
      - 4.3.1: Constructing Surveys - *CC BY-NC-SA 4.0*
      - 4.3.2: Reliability and Validity of Measurement - *CC BY-NC-SA 4.0*
    - 4.4: Interviews and Focus Groups - *CC BY-NC-SA 4.0*
    - 4.5: Measuring Behavior through Biology - *Unknown License*
    - 4.6: Observations - *CC BY-NC-SA 4.0*
    - 4.7: Unobtrusive Research - *CC BY-NC-SA 3.0*
      - 4.7.1: What Is Left Behind (Physical Trace) - *Public Domain*
      - 4.7.2: Archival Data - *CC BY-NC-SA 4.0*
    - 4.8: Summary of Measuring Variables - *CC BY-NC-SA 4.0*

- 5: Literature Review - *Undeclared*
  - 5.1: Prelude to Reviewing Literature - *CC BY-SA 4.0*
  - 5.2: How to do a Literature Review - *CC BY-NC-SA 4.0*
  - 5.3: How to Read a Research Article - *CC BY-NC-SA 4.0*
  - 5.4: Interpreting Results (p-values and NHST) - *CC BY-SA 4.0*
  - 5.5: Summary of Reviewing the Literature - *CC BY-NC-SA 4.0*
- 6: Analyzing - *Undeclared*
  - 6.1: Prelude to Analyzing - *Undeclared*
  - 6.2: Describing Variables - *CC BY-NC-SA 4.0*
  - 6.3: Graphs - *CC BY-SA 4.0*
    - 6.3.1: Graphing Qualitative Variables- Bar Graphs - *CC BY-NC-SA 4.0*
    - 6.3.2: Describing the Shape of Graphs of Quantitative Data - *CC BY-SA 4.0*
      - 6.3.2.1: Graphing Quantitative Data- Line Graphs - *CC BY-SA 4.0*
      - 6.3.2.2: Graphing Quantitative Data- Histograms - *CC BY-SA 4.0*
    - 6.3.3: Graphing Quantitative Data- Scatterplots - *CC BY-NC-SA 4.0*
    - 6.3.4: Honesty in Graphing - *CC BY-NC-SA 4.0*
  - 6.4: Understanding Null Hypothesis Significance Testing - *CC BY-NC-SA 4.0*
  - 6.5: Comparing Means - *CC BY-NC-SA 4.0*
    - 6.5.1: Effect Sizes - *CC BY-NC-SA 4.0*
  - 6.6: Introduction to Correlations - *CC BY-SA 4.0*
    - 6.6.1: Correlation versus Causation - *CC BY-NC-SA 4.0*
    - 6.6.2: Factor Analysis - *CC BY-NC-SA 4.0*
  - 6.7: The Qualitative Approach - *CC BY-NC-SA 4.0*
    - 6.7.1: Qualitative Analysis- Preparing and Managing Qualitative Data - *CC BY-NC-SA 4.0*
    - 6.7.2: Qualitative Coding - *CC BY-NC-SA 4.0*
    - 6.7.3: From Qualitative Data to Findings - *CC BY-NC-SA 4.0*
  - 6.8: Summary of Analyzing - *CC BY-NC-SA 4.0*
- 7: Formatting in APA Style - *Undeclared*
  - 7.1: Prelude to Presenting Your Research - *Undeclared*
  - 7.2: Introduction to American Psychological Association (APA) Style - *CC BY-NC-SA 4.0*
  - 7.3: General Paper Format - *CC BY-NC-SA 4.0*
  - 7.4: Formatting by Section - *CC BY-NC-SA 4.0*
  - 7.5: Presenting Results - *CC BY-NC-SA 4.0*
    - 7.5.1: Tables and Figures - *CC BY-NC-SA 4.0*
  - 7.6: Research Proposals - *CC BY-SA 4.0*
  - 7.7: Summary and APA Style Resources - *CC BY-NC-SA 4.0*
- 8: Experimental Research - *Undeclared*
  - 8.1: Prelude to Experimental Research - *Undeclared*
  - 8.2: Initial Equivalence - *CC BY-NC-SA 4.0*
    - 8.2.1: Natural Groups Research - *CC BY-NC-SA 4.0*
  - 8.3: Between Groups Designs and Matched Group Designs - *CC BY-NC-SA 4.0*
  - 8.4: Repeated Measures Design - *CC BY-NC-SA 4.0*
  - 8.5: Practical Considerations - *CC BY-NC-SA 4.0*
  - 8.6: Summary of Experimental Research - *CC BY-NC-SA 4.0*
- 9: Quasi-Experimental Research - *Undeclared*
  - 9.1: Prelude to Quasi-Experimental Research - *Undeclared*
  - 9.2: Quasi-Experimental Designs with One Group - *CC BY-NC-SA 4.0*
  - 9.3: Quasi-Experimental Designs with Two or More Groups - *CC BY-NC-SA 4.0*
  - 9.4: Threats to Internal Validity - *CC BY-NC-SA 4.0*
  - 9.5: Summary of Quasi-Experimental Research - *CC BY-NC-SA 4.0*
- 10: Factorial Designs - *Undeclared*
  - 10.1: Prelude to Factorial Designs - *Undeclared*
  - 10.2: Setting Up a Factorial Experiment - *CC BY-NC-SA 4.0*
  - 10.3: Interpreting the Results of a Factorial Experiment - *CC BY-NC-SA 4.0*
    - 10.3.1: Practice Interpreting Main Effects and Interactions - *CC BY-SA 4.0*
    - 10.3.2: Interpreting Main Effects and Interactions in Graphs - *CC BY-SA 4.0*
    - 10.3.3: Interpreting Beyond 2x2 in Graphs - *CC BY-SA 4.0*
  - 10.4: Interpreting Interactions- Do Main Effects Matter? - *CC BY-SA 4.0*
  - 10.5: Summary of Factorial Designs - *CC BY-NC-SA 4.0*
- 11: Small-N Designs - *Undeclared*
  - 11.1: Prelude to Single-Subject Research - *Undeclared*
  - 11.2: Introduction to Small-N Research - *CC BY-NC-SA 4.0*

- 11.3: Small-N Research Designs and Analysis - *CC BY-NC-SA 4.0*
- 11.4: When should you use Small-N Designs? - *CC BY-NC-SA 4.0*
- 11.5: Summary of Small-N Design - *CC BY-NC-SA 4.0*

- 12: Nonexperimental Research - *Undeclared*
  - 12.1: Prelude to Nonexperimental Research - *Undeclared*
  - 12.2: Introduction to Nonexperimental Research - *CC BY-NC-SA 4.0*
  - 12.3: Longitudinal and Similar Designs - *CC BY-NC-SA 4.0*
  - 12.4: Qualitative Research - *CC BY-NC-SA 4.0*
  - 12.5: Case Studies - *CC BY-NC-SA 4.0*
  - 12.6: Summary of Nonexperimental Research - *CC BY-NC-SA 4.0*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Common Formulas - *Public Domain*
  - Detailed Licensing - *Undeclared*

Common Formulas

The following formulas are in the order in which you learn about them in this textook. Use the Table of Contents to look for a specific equation.

Descriptive Statistics

Mean

$$\bar{X} = \frac{\sum X}{N} \tag{1}$$

Standard Deviation

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}} \tag{2}$$

Which is also: $s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}} = \sqrt{\frac{SS}{df}}$

Some instructors prefer this formula because it is easier to calculate (but more difficult to see what's happening):

$$\sqrt{\frac{\left(\sum (X^2) - \frac{(\sum X)^2}{N}\right)}{(N - 1)}} \tag{3}$$

z-score

To find the z-score when you have a raw score:

$$z = \frac{X - \bar{X}}{s} \tag{4}$$

To find a raw score when you have a z-score:

$$x = zs + \overline{X} \tag{5}$$

t-tests

One-Sample t-test

These are the same formulas, but formatted slightly differently.

$$t = \frac{(\bar{X} - \mu)}{\left(\frac{s}{\sqrt{n}}\right)} \tag{6}$$

Confidence Interval

$$\text{Margin of Error } = t \times \left(\frac{s}{\sqrt{N}}\right)$$

$$\text{Confidence Interval } = \overline{X} \pm (t \times \left(\frac{s}{\sqrt{N}}\right)) \tag{7}$$

Independent Sample t-test

Unequal N

You can always use this formula:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left[\frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}\right] \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{8}$$

Equal N

You should only use this formula when your two independent groups are the same size (N), meaning the same number of people in each group.

$$\frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{s_1^2}{N_1}\right) + \left(\frac{s_2^2}{N_2}\right)}} \tag{9}$$

Dependent Sample t-test

Conceptual Formula (symbols)

$$t = \frac{\overline{X}_D}{\left(\frac{s_D}{\sqrt{N}}\right)} \tag{10}$$

Full Formula

$$t = \frac{\left(\frac{\Sigma D}{N}\right)}{\sqrt{\left(\frac{\sum\left((X_D - \overline{X}_D)^2\right)}{(N-1)}\right)/\sqrt{N}}} \tag{11}$$

ANOVA

Sums of Squares for Between Groups Designs

Between Groups

$$SS_B = \sum_{EachGroup}\left[\left(\overline{X}_{group} - \overline{X}_T\right)^2 \times (n_{group})\right] \tag{12}$$

Within Groups

$$SS_W = \sum_{EachGroup}\left[\sum\left(\left(X - \overline{X}_{group}\right)^2\right)\right] \tag{13}$$

Total

$$SS_T = \sum\left[\left(X - \overline{X}_T\right)^2\right] \tag{14}$$

Tukey's HSD for Pairwise Comparison

$$HSD = q \times \sqrt{\frac{MSw}{n_{group}}} \tag{15}$$

Sums of Squares for Repeated Measures Designs

Between Groups

Same as above.

Participants

$$SS_{Ps} = \left[\sum\left(\frac{(\sum X_{Ps})^2}{k}\right)\right] - \frac{((\sum X)^2)}{N} \tag{16}$$

Within Groups (Error)

$$SS_{WG} = SS_T - SS_{BG} - S_P$$

Total

Same as above.

Pearson's r (Correlation)

The following formulas are the same. Use the first one when you already have the standard deviation calculated.

These are paired data, so *N is the number of pairs*.

SD Already Calculated:

$$r = \frac{\left(\frac{\sum((x_{Each} - \bar{X}_x) \times (y_{Each} - \bar{X}_y))}{(N-1)}\right)}{(s_x \times s_y)} \tag{17}$$

SD Not Calculated:

$$r = \frac{\left(\frac{\sum((x - \bar{X}_x) \times (y - \bar{X}_y))}{(N-1)}\right)}{\left(\sqrt{\frac{\sum\left((x - \overline{X_x})^2\right)}{N-1}}\right) \times \left(\sqrt{\frac{\sum\left((y - \overline{X_y})^2\right)}{N-1}}\right)} \tag{18}$$

Regression Line Equation

$$\widehat{\mathrm{Y}} = \mathrm{a} + (\mathrm{b} \times X) \tag{19}$$

a (intercept):

$$\mathrm{a} = \overline{X_y} - (\mathrm{b} \times \overline{X_x}) \tag{20}$$

b (slope):

$$\frac{\sum(Diff_x \times Diff_y)}{\sum(Diff_X{}^2)} \tag{21}$$

In which "Diff" means the differences between each score and that variable's mean.

Pearson's χ^2 (Chi-Square)

$$\chi^2 = \sum_{Each}\left(\frac{(E-O)^2}{E}\right) \tag{22}$$

Expected Frequencies

Goodness of Fit:

$$\frac{N}{k} \tag{23}$$

Test of Independence:

$$E_{EachCell} = \frac{RT \times CT}{N} \tag{24}$$

In which RT = Row Total and CT = Column Total

Detailed Licensing

Overview

Title: Research Methods for the Social and Behavioral Sciences

Webpages: 125

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 46.4% (58 pages)
- Undeclared: 25.6% (32 pages)
- CC BY-SA 4.0: 24% (30 pages)
- Public Domain: 1.6% (2 pages)
- CC BY 4.0: 0.8% (1 page)
- Unknown License: 0.8% (1 page)
- CC BY-NC-SA 3.0: 0.8% (1 page)

By Page

- Research Methods for the Social and Behavioral Sciences - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Acknowledgements - *Undeclared*
    - Licensing - *Undeclared*
    - Licensing - *Undeclared*
    - For Faculty - *CC BY 4.0*
  - 1: Introduction to Research Methods - *CC BY-SA 4.0*
    - 1.1: Prelude to Research Methods - *CC BY-SA 4.0*
    - 1.2: Methods of Knowing - *CC BY-SA 4.0*
    - 1.3: Why are you taking this course? - *CC BY-SA 4.0*
    - 1.4: Summary of Introduction to Research Methods of the Social Sciences - *CC BY-SA 4.0*
  - 2: Overview of the Scientific Method - *CC BY-SA 4.0*
    - 2.1: Prelude to the Scientific Method - *CC BY-SA 4.0*
    - 2.2: Goals of Science - *CC BY-SA 4.0*
    - 2.3: A Model of Scientific Research - *CC BY-SA 4.0*
      - 2.3.1: Research Literature - *CC BY-SA 4.0*
      - 2.3.2: Research Questions - *CC BY-SA 4.0*
      - 2.3.3: Empirical Study - *CC BY-SA 4.0*
      - 2.3.4: Data Analysis - *CC BY-SA 4.0*
      - 2.3.5: Conclusions - *CC BY-SA 4.0*
      - 2.3.6: Sharing Your Research - *CC BY-NC-SA 4.0*
    - 2.4: Populations and Samples - *CC BY-SA 4.0*
      - 2.4.1: Sampling - *CC BY-SA 4.0*
    - 2.5: Summary of the Scientific Method - *CC BY-SA 4.0*
  - 3: Ethics in Research - *Undeclared*
    - 3.1: Prelude to Research Ethics - *Undeclared*
    - 3.2: History of Ethics Codes - *CC BY-NC-SA 4.0*
    - 3.3: APA Ethics Code Standard 8 - *CC BY-NC-SA 4.0*
    - 3.4: Principles of Ethical Research - *CC BY-NC-SA 4.0*
    - 3.5: Summary of Research Ethics - *CC BY-NC-SA 4.0*
  - 4: Measuring Variables - *Undeclared*
    - 4.1: Prelude to Psychological Measurement - *Undeclared*
    - 4.2: Introduction to Measuring Variables - *CC BY-NC-SA 4.0*
      - 4.2.1: Types of Validity - *CC BY-SA 4.0*
    - 4.3: Introduction to Surveys - *CC BY-NC-SA 4.0*
      - 4.3.1: Constructing Surveys - *CC BY-NC-SA 4.0*
      - 4.3.2: Reliability and Validity of Measurement - *CC BY-NC-SA 4.0*
    - 4.4: Interviews and Focus Groups - *CC BY-NC-SA 4.0*
    - 4.5: Measuring Behavior through Biology - *Unknown License*
    - 4.6: Observations - *CC BY-NC-SA 4.0*
    - 4.7: Unobtrusive Research - *CC BY-NC-SA 3.0*
      - 4.7.1: What Is Left Behind (Physical Trace) - *Public Domain*
      - 4.7.2: Archival Data - *CC BY-NC-SA 4.0*
    - 4.8: Summary of Measuring Variables - *CC BY-NC-SA 4.0*

- 5: Literature Review - *Undeclared*
  - 5.1: Prelude to Reviewing Literature - *CC BY-SA 4.0*
  - 5.2: How to do a Literature Review - *CC BY-NC-SA 4.0*
  - 5.3: How to Read a Research Article - *CC BY-NC-SA 4.0*
  - 5.4: Interpreting Results (p-values and NHST) - *CC BY-SA 4.0*
  - 5.5: Summary of Reviewing the Literature - *CC BY-NC-SA 4.0*
- 6: Analyzing - *Undeclared*
  - 6.1: Prelude to Analyzing - *Undeclared*
  - 6.2: Describing Variables - *CC BY-NC-SA 4.0*
  - 6.3: Graphs - *CC BY-SA 4.0*
    - 6.3.1: Graphing Qualitative Variables- Bar Graphs - *CC BY-NC-SA 4.0*
    - 6.3.2: Describing the Shape of Graphs of Quantitative Data - *CC BY-SA 4.0*
      - 6.3.2.1: Graphing Quantitative Data- Line Graphs - *CC BY-SA 4.0*
      - 6.3.2.2: Graphing Quantitative Data- Histograms - *CC BY-SA 4.0*
    - 6.3.3: Graphing Quantitative Data- Scatterplots - *CC BY-NC-SA 4.0*
    - 6.3.4: Honesty in Graphing - *CC BY-NC-SA 4.0*
  - 6.4: Understanding Null Hypothesis Significance Testing - *CC BY-NC-SA 4.0*
  - 6.5: Comparing Means - *CC BY-NC-SA 4.0*
    - 6.5.1: Effect Sizes - *CC BY-NC-SA 4.0*
  - 6.6: Introduction to Correlations - *CC BY-SA 4.0*
    - 6.6.1: Correlation versus Causation - *CC BY-NC-SA 4.0*
    - 6.6.2: Factor Analysis - *CC BY-NC-SA 4.0*
  - 6.7: The Qualitative Approach - *CC BY-NC-SA 4.0*
    - 6.7.1: Qualitative Analysis- Preparing and Managing Qualitative Data - *CC BY-NC-SA 4.0*
    - 6.7.2: Qualitative Coding - *CC BY-NC-SA 4.0*
    - 6.7.3: From Qualitative Data to Findings - *CC BY-NC-SA 4.0*
  - 6.8: Summary of Analyzing - *CC BY-NC-SA 4.0*
- 7: Formatting in APA Style - *Undeclared*
  - 7.1: Prelude to Presenting Your Research - *Undeclared*
  - 7.2: Introduction to American Psychological Association (APA) Style - *CC BY-NC-SA 4.0*
  - 7.3: General Paper Format - *CC BY-NC-SA 4.0*
  - 7.4: Formatting by Section - *CC BY-NC-SA 4.0*
  - 7.5: Presenting Results - *CC BY-NC-SA 4.0*
    - 7.5.1: Tables and Figures - *CC BY-NC-SA 4.0*
  - 7.6: Research Proposals - *CC BY-SA 4.0*
  - 7.7: Summary and APA Style Resources - *CC BY-NC-SA 4.0*
- 8: Experimental Research - *Undeclared*
  - 8.1: Prelude to Experimental Research - *Undeclared*
  - 8.2: Initial Equivalence - *CC BY-NC-SA 4.0*
    - 8.2.1: Natural Groups Research - *CC BY-NC-SA 4.0*
  - 8.3: Between Groups Designs and Matched Group Designs - *CC BY-NC-SA 4.0*
  - 8.4: Repeated Measures Design - *CC BY-NC-SA 4.0*
  - 8.5: Practical Considerations - *CC BY-NC-SA 4.0*
  - 8.6: Summary of Experimental Research - *CC BY-NC-SA 4.0*
- 9: Quasi-Experimental Research - *Undeclared*
  - 9.1: Prelude to Quasi-Experimental Research - *Undeclared*
  - 9.2: Quasi-Experimental Designs with One Group - *CC BY-NC-SA 4.0*
  - 9.3: Quasi-Experimental Designs with Two or More Groups - *CC BY-NC-SA 4.0*
  - 9.4: Threats to Internal Validity - *CC BY-NC-SA 4.0*
  - 9.5: Summary of Quasi-Experimental Research - *CC BY-NC-SA 4.0*
- 10: Factorial Designs - *Undeclared*
  - 10.1: Prelude to Factorial Designs - *Undeclared*
  - 10.2: Setting Up a Factorial Experiment - *CC BY-NC-SA 4.0*
  - 10.3: Interpreting the Results of a Factorial Experiment - *CC BY-NC-SA 4.0*
    - 10.3.1: Practice Interpreting Main Effects and Interactions - *CC BY-SA 4.0*
    - 10.3.2: Interpreting Main Effects and Interactions in Graphs - *CC BY-SA 4.0*
    - 10.3.3: Interpreting Beyond 2x2 in Graphs - *CC BY-SA 4.0*
  - 10.4: Interpreting Interactions- Do Main Effects Matter? - *CC BY-SA 4.0*
  - 10.5: Summary of Factorial Designs - *CC BY-NC-SA 4.0*
- 11: Small-N Designs - *Undeclared*
  - 11.1: Prelude to Single-Subject Research - *Undeclared*
  - 11.2: Introduction to Small-N Research - *CC BY-NC-SA 4.0*

- 11.3: Small-N Research Designs and Analysis - *CC BY-NC-SA 4.0*
- 11.4: When should you use Small-N Designs? - *CC BY-NC-SA 4.0*
- 11.5: Summary of Small-N Design - *CC BY-NC-SA 4.0*

- 12: Nonexperimental Research - *Undeclared*
  - 12.1: Prelude to Nonexperimental Research - *Undeclared*
  - 12.2: Introduction to Nonexperimental Research - *CC BY-NC-SA 4.0*
  - 12.3: Longitudinal and Similar Designs - *CC BY-NC-SA 4.0*
  - 12.4: Qualitative Research - *CC BY-NC-SA 4.0*
  - 12.5: Case Studies - *CC BY-NC-SA 4.0*
  - 12.6: Summary of Nonexperimental Research - *CC BY-NC-SA 4.0*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Common Formulas - *Public Domain*
  - Detailed Licensing - *Undeclared*