

1 Psychology as a Science



VisitBritain/Jason Knott/Getty Images

Learning Outcomes

By the end of this chapter, you should be able to:

- Outline the major areas of research in the field of psychology.
- Explain the process of testing research ideas through the scientific method.
- Describe what it means to turn an idea into a testable hypothesis.
- Identify the criteria for a good theory.
- Search online databases for previous research studies.
- Summarize the key ethical principles that apply to conducting research on human and non-human animals.

In an article in *Wired* magazine, journalist Amy Wallace (2009) described her visit to the annual conference sponsored by Autism One, a nonprofit group organized around the belief that autism is caused by mandatory childhood vaccines:

I flashed more than once on Carl Sagan’s idea of the power of an “unsatisfied medical need.” Because a massive research effort has yet to reveal the precise causes of autism, pseudoscience has stepped in to the void. In the hallways of the Westin O’Hare hotel, helpful salespeople strove to catch my eye . . . pitching everything from vitamins and supplements to gluten-free cookies . . . hyperbaric chambers, and neuro-feedback machines. (p. 134)

The “pseudoscience” to which Wallace refers is the claim that vaccines generally do more harm than good and specifically that they cause children to develop autism. In fact, an extensive statistical review of epidemiological studies, including tens of thousands of vaccinated children, found *no evidence* of a link between vaccines and autism (Madsen et al., 2002). The reality is this: Research tells us that vaccines bear no relation to autism, but people still believe that they do. Because of these beliefs, increasing numbers of parents are foregoing vaccinations, and many communities are seeing a resurgence of rare diseases like measles and mumps.

So what does it mean to say that “research” has reached a conclusion? Why should we trust this conclusion over parents’ personal experience with their own child? One of the biggest challenges in starting a course on research methods is learning how to think like a scientist—that is, to frame questions in testable ways and to make decisions by weighing the evidence. The more personal these questions become, and the bigger their consequences, the harder it is to put feelings aside. However, as we will see throughout this course, in these cases precisely, listening to the evidence becomes most important.

Understanding the importance of scientific thinking matters for several reasons, even if a student never takes another psychology course. First, at a practical level, critical thinking is an invaluable skill in a wide variety of careers. Employers of all types appreciate the ability to reason through the decision-making process. Second, understanding the scientific approach tends to make people more skeptical consumers of news reports. Someone who reads in *Newsweek* that the planet is warming, or cooling, or staying the same will be able to decipher and evaluate how the author reached this conclusion and possibly reach a different one. Third, understanding science makes a person a more informed participant in debates about public policy. To know whether the planet is truly getting warmer requires carefully weighing the scientific evidence rather than trusting the loudest pundit on a cable news network.

Where does psychology fit into this picture? Objectivity can be a particular challenge in studying our own behavior and mental processes because we are intimately familiar with the processes we are trying to understand. The psychologist William C. Corning (1968) captured this sentiment over 40 years ago: “In the study of brain functions, we rely upon a biased, poorly understood, and frequently unpredictable organ in order to study the properties of another such organ; we have to use a brain to study a brain” (p. 6). (Or, in the words of comedian Emo Phillips, “I used to think that the brain was the most wonderful organ in my body. Then I realized who was telling me this” [Jarski, 2007].) The trick, then, is learning to take a step back and apply scientific thinking to issues we encounter and experience every day.

This textbook provides an introduction to the research methods used in the study of psychology. It introduces the full spectrum of research designs, from observing behavior to carefully

controlling conditions in a laboratory. The text will cover the key issues and important steps for each type of design, as well as the analysis strategies most appropriate for each one. This chapter begins with an overview of the different areas of psychological science. It then introduces the research process by discussing the key features of the scientific approach and the process of forming testable research questions. The final section discusses the importance of adhering to ethical principles at all stages of the research.

Research: Making an Impact

The Vaccines and Autism Controversy

In a 1998 paper published in the well-respected medical journal *The Lancet*, British physician Andrew Wakefield and his colleagues studied the link between autism symptoms and the measles, mumps, and rubella (MMR) vaccine in a sample of twelve children. Based on a review of these cases, the authors reported that all twelve experienced adverse effects of the vaccine, including both intestinal and behavioral problems. The finding that grabbed the headlines was the authors' report that nine of the twelve children showed an onset of autism symptoms shortly after they received the MMR vaccine.

Immediately after the publication of this paper, the scientific community criticized the study for its small sample and its lack of a comparison group (i.e., children in the general population). Unfortunately, these issues turned out to be only the tip of the iceberg (Godlee, Smith, & Marcovitch, 2011). British journalist Brian Deer (2004) conducted an in-depth investigation of Wakefield's study and discovered some startling information. First, the study had been funded by a law firm that was in the process of suing the manufacturers of the MMR vaccine, thereby threatening researchers' objectivity. Second, Deer's investigation showed clear evidence of scientific misconduct: The data had been falsified and altered to fit Wakefield's hypothesis—many of the children had shown autism symptoms before receiving the vaccine. In his report, Deer stated that every one of the twelve cases showed evidence of alteration and misrepresentation.

Ultimately, *The Lancet* withdrew the article in 2010, effectively removing it from the scientific record and declaring the findings no longer trustworthy. But in many respects, the damage was already done. Vaccination rates in Britain dropped to 80% following publication of Wakefield's article, and these rates remain below the recommended 95% level recommended by the World Health Organization (Godlee et al., 2011). Even though the article was a fraud, it made parents afraid to vaccinate their children.

Vaccinations work optimally when most members of a community receive the vaccines because this minimizes the opportunity for an outbreak. When even a small portion of a population refuses to vaccinate children, it places the entire community at risk of infection (National Institute of Allergy and Infectious Diseases, n.d.). Thus, it should be no surprise that many communities are seeing a resurgence of measles, mumps, and rubella: In 2008, England and Wales declared measles to be a prevalent problem for the first time in 14 years (Godlee et al., 2011).

This scenario highlights the importance of conducting science honestly. While disease outbreaks are the most obvious impact of Wakefield's fraud, they are not the only one. In a 2011 editorial in the *British Medical Journal* condemning Wakefield's actions, British doctor Fiona Godlee and colleagues captured this rather eloquently: "But perhaps as important as the scare's effect on infectious disease is the energy, emotion, and money that have been diverted away from efforts to understand the real causes of autism and how to help children and families who live with it."

1.1 Major Research Areas in Psychology

Psychology is a diverse discipline, encompassing a wide range of approaches to questions about why people do the things that they do. The common thread among all of these approaches is the *scientific study* of human behavior. So, while psychology might not be the only field to speculate on the causes of human behavior—philosophers have been doing this for millennia—psychology is distinguished by its reliance on the scientific method to draw conclusions. Later, the chapter will examine the meaning and implications of this scientific perspective. This section discusses the major research areas within the field of psychology, along with samples of the types of research questions asked by each one.

Biopsychology

Biopsychology, as the name implies, combines research questions and techniques from both biology and psychology. It is typically defined as the study of connections between biological systems (including the brain, hormones, and neurotransmitters) and thoughts, feelings, and behaviors. As a result, the research conducted by biopsychologists often overlaps research in other areas—but with a focus on biological processes. Biopsychologists are often interested in the way interactions between biological systems and thoughts, feelings, and behaviors affect the ability to treat disease, as the following questions reflect: What brain systems are involved in the formation of memories? Can Alzheimer's be cured or prevented through early intervention? How does long-term exposure to toxins such as lead influence our thoughts, feelings, and behaviors? How easily can the brain recover after a stroke?

In one example of this approach, Kim and colleagues (2010) investigated changes in brain anatomy among new mothers for the first three months following delivery. These authors were intrigued by the numerous changes new mothers undergo in attention, memory, and motivation; they speculated that these changes might be associated with changes in brain structure. As expected, new mothers showed increases in grey matter (i.e., increased complexity) in several brain areas associated with maternal motivation and behavior. In addition, the more these brain areas developed, the more positively these women felt toward their newborn children. Thus, Kim et al.'s study sheds light on the potential biological processes involved in the mother–infant bond.

Cognitive Psychology

Whereas biopsychology focuses on studying the brain, **cognitive psychology** studies the mind. It is typically defined as the study of internal mental processes, including the ways that people think, learn, remember, speak, perceive, and so on. Cognitive psychologists are primarily interested in the ways that people navigate and make sense of the world. Research questions in this field might ask: How do our minds translate input from the five senses into a meaningful picture of the world? How do we form memories of emotional versus mundane experiences? What draws our attention in a complex environment? What is the best way to teach children to read?

In one example of this approach, Foulsham, Cheng, Tracy, Henrich, and Kingstone (2010) were interested in what kinds of things people pay attention to in a complex social scene.

The world around us is chock-full of information, but we can only pay attention to a relatively thin slice of it. Foulsham and colleagues were particularly interested in where our attention is directed when we observe groups of people. They answered this question by asking people to watch videos of a group discussion and using tools to track eye movements. It turned out that people in this study spent most of their time looking at the most dominant member of the group, suggesting that individuals are wired to pay attention to those in positions of power. Thus, this study sheds light on one of the ways that people make sense of the world.

Developmental Psychology

Developmental psychology is defined as the systematic study of physical, social, and cognitive changes over the human life span. Although this field initially focused on childhood development, many researchers now study changes and key stages over a person's entire life span. Developmental psychologists look at a wide range of phenomena related to physical, social, and cognitive change, including: How do children bond with their primary caregiver(s)? What are our primary needs and goals at each stage of life? Why do some cognitive skills decline in old age? At what ages do infants develop basic motor skills?

In one example of this approach, Hill and Tyson (2009) explored the connection between children's school achievement and their parents' involvement with the school. In other words: Do children perform better when their parents are actively involved in school activities? The authors addressed this question by combining results from several studies into one data set. Across 50 studies, the answer to this question was yes—children do better in school if their parents are involved. Hill and Tyson's study sheds light on a key predictor of academic achievement during an important developmental period.

Social Psychology

Social psychology, which attempts to study behavior in a broader social context, is typically defined as the study of the ways humans' thoughts, feelings, and behaviors are shaped by other people. This broad perspective allows social psychologists to tackle a wide range of research questions, such as: What kinds of things do individuals look for in selecting romantic partners? Why do people stay in bad relationships? How do other people shape individuals' sense of who they are? When and why do people help in emergencies?

Norman Triplett (1898) conducted the first published social psychology study at the end of the 19th century. Triplett had noticed that professional cyclists tended to ride faster when racing against other cyclists than when competing in solo time trials. He tested this observation in a controlled



Thomas Northcut/Photodisc/Thinkstock

Social psychologist Norman Triplett's study of competition among cyclists led to conclusions about how people influence one another.

laboratory setting, asking people to do a number of tasks either alone or next to another person. His results (and countless other studies since) revealed that people worked faster in groups, suggesting that other people can have definite and concrete influences on human behavior.

Clinical Psychology

The area of **clinical psychology** focuses on understanding the best ways to treat psychological disorders. It is typically defined as the study of best practices for understanding, treating, and preventing distress and dysfunction. Clinical psychologists engage in both the assessment and the treatment of psychological disorders, as the following research questions suggest: What is the most effective treatment for depression? How can we help people overcome post-traumatic stress disorder following a traumatic event? Should anxiety disorders be treated with drugs, therapy, or a combination? What is the most reliable way to diagnose schizophrenia?

A study by Kleim and Ehlers (2008) offers an example of this approach. The study attempted to understand the risk factors for post-traumatic stress disorder, a prolonged reaction to a severe traumatic experience. Kleim and Ehlers found that assault victims who tend to form less specific memories about life in general might be more likely to develop a disorder in response to trauma than victims who tend to form detailed memories. People who tend to form vague memories may have fewer resources to draw on in trying to reconnect with their daily life after a traumatic event. This study, then, sheds light on a possible pathway contributing to the development of a psychological disorder.

Applied Research Areas

The research areas listed thus far represent the majority of basic research within psychology, but the list is not exhaustive. A great deal of additional psychological research focuses on understanding human behavior in a more *applied* context. For example, the field of **health psychology** applies psychological principles to the study of health, wellness, and illness. Health psychologists often have a background in either clinical or social psychology and use these insights toward a broader understanding of why people get sick. One major insight from this field is that the quality and quantity of our relationships with other people can actually have a dramatic impact on our physical health. Close relationships can provide practical support in times of need (e.g., making it easier to get to the doctor), as well as making stressful events seem less stressful (for review, see Newman & Roberts, 2012).

Similarly, the field of **industrial–organizational psychology** (often abbreviated as I/O psychology) applies psychological principles to the scientific study of human behavior in the workplace. I/O psychologists often have a background in social or cognitive psychology and generally help organizations function more effectively by improving employee satisfaction, performance, and safety of employees. One major insight from this field shows that people are often more productive in the workplace if given more freedom over their time. This

model started in the high-tech industry. For example, Google employees have game rooms around the office; the company requires workers to spend time each week developing “side” projects unrelated to their main responsibilities. This approach makes employees feel more valued as individuals, more dedicated to the company, and thus more industrious in completing their work.

As a final example, the field of **school psychology** applies psychological principles to the goal of helping children learn effectively. School psychologists, who are typically trained in developmental, clinical, and educational psychology, work to meet the learning and behavioral health needs of students. More so than the previous examples, school psychologists play a “practitioner” role, applying their broad knowledge base to provide psychological diagnosis, conduct health promotions, evaluate services, and conduct interventions with individual students as needed.

To learn more about all of these areas, see the American Psychological Association’s collection of web resources: <http://www.apa.org/about/division/index.aspx>.

1.2 The Research Process

With a broad understanding of the major research areas in psychology, we now turn our attention to the research process. How do psychologists conduct research? What are their goals? This section will answer these questions. This section will also compare quantitative and qualitative research, two different approaches to scientific inquiry.

The Scientific Method

What does it mean to draw conclusions based on science? Scientists across all quantitative disciplines use the same process of forming and testing their ideas. The overall goal of this research process—also known as the **scientific method**—is to draw conclusions based on empirical observations. In this section, we cover the four steps of the research process—*hypothesize*, *operationalize*, *measure*, and *explain*, abbreviated with the acronym HOME.

Step 1—Hypothesize

The first step in the research process turns an initial research question into a testable prediction, or **hypothesis**. A hypothesis is a specific statement about the relationship between two or more variables. For example, if we start with a question about the link between smoking and cancer, our *hypothesis* might be that smoking causes lung cancer. Or, if we want to know whether a new drug will be helpful in treating depression, we might *hypothesize* that drug X will lead to a reduction in depression symptoms. The next section of this chapter will cover hypotheses in more detail, but for now it is important to understand that the way a hypothesis is framed guides every other step of the research process.

Step 2—Operationalize

Once a researcher develops a hypothesis, the next step is to decide how to test it. The process of **operationalization** involves choosing measurable variables to represent the elements of the hypothesis. In the depression-drug example, we need to decide how to measure both cause and effect; in this case we define the cause as the drug and the effect as reduced symptoms of depression. That is, what doses of the drug should we investigate? How many different doses should we compare? And, how will we measure depression symptoms? Will it work to have people complete a questionnaire? Or do we need to have a clinician interview participants before and after they take the drug?

An additional complication for psychology studies is that many of research questions deal with abstract concepts. Turning these concepts into measurable variables requires some art. For example, the abstract concept of *happiness* could be defined in countless different ways—being “happy” likely means something different to one individual than it does to his neighbors. To include happiness in a research study, we need to translate it into a more concrete concept, measured by a person’s score on a happiness scale or by the number of times a person smiles in a five-minute period, or perhaps even by a person’s subjective experience of happiness during an interview. Chapter 2 (2.2) will cover this process in more detail, with a discussion of guidelines for making these important decisions about the study.

Step 3—Measure

Now that we have developed both our research question and our operational definitions, it is time to collect some data. The text will cover this process in great detail, dedicating Chapters 3 through 5 to the three primary approaches to data collection. Collection of data is a critical step in the research process, as researchers gather empirical observations that will help address their hypothesis. As Chapter 2 will explain, these observations can range from questionnaire responses to measures of brain activity, and they can be collected in a variety of ways, from online questionnaires to carefully controlled experiments. Regardless of the details of data collection, investigators will ultimately use these observations to make a decision.

Step 4—Explain

After data have been collected, the final step is to analyze and interpret the results. The goal of this step is to return full circle to the initial research question and determine whether the results support the hypothesis. Recall the hypothesis that drug X should reduce depression symptoms. If we find at the end of the study that people who took drug X showed a 70% decrease in symptoms, this result would be consistent with the hypothesis. However, the explanation stage also involves thinking about alternative explanations and planning for future studies. What if depression symptoms dropped simply due to the passage of time? How could we address this concern in a future study? As it turns out, a fairly easy way of fixing this problem exists; Chapter 5 will cover that solution.

As Table 1.1 summarizes, the research process involves four stages: forming a hypothesis, deciding how to test it, collecting data, and interpreting the results. This process is used to draw conclusions across all scientific disciplines, regardless of whether research questions involve depression drugs, reading speed, or the speed of light in a vacuum.

Table 1.1 The HOME method

Stage of Process	Main Idea	Example
Hypothesize	Take a research question, turn it into a testable prediction	<i>Question:</i> Will my new drug help depression patients? <i>Hypothesis:</i> Drug X will reduce depression symptoms.
Operationalize	Turn the key concepts from your hypothesis into measurable variables	Depression can be measured using clinician interviews
Measure	Choose and implement the best research design for your hypothesis	Compare two groups of people over time, half of whom have been given the new drug
Explain	Interpret your findings and make a decision about the state of your hypothesis	If the people who take the new drug are less depressed at the end, that supports our hypothesis

Research: Applying Concepts

Examples of the Research Process

To make the steps of the scientific method a bit more concrete, the following two examples show how they could be applied to specific research topics.

Example 1—Depression and Heart Disease

Depression affects approximately 20 million Americans, and 16% of the population will experience it at some time in their lives (NIMH, 2007). Depression is associated with a range of emotional and physical symptoms, including feelings of hopelessness and guilt, loss of appetite, sleep disturbance, and suicidal thoughts. This list has recently been expanded even further to include an increased risk of heart disease. Individuals who are otherwise healthy but suffering from depression are more likely to develop and to die from cardiovascular disease than those without depression. According to one study, patients who experience depression following a heart attack experience a fourfold increase in five-year mortality rates (research reviewed in Glassman et al., 2011).

Research Question

Based on these findings, we could ask the question, “Would it make sense to treat heart attack patients with antidepressant drugs?”

Recall that the goal of the scientific method is to take this research question, turn it into a testable hypothesis, and conduct a study that will test it. The following steps use the HOME method discussed earlier.

(continued)

Research: Applying Concepts (*continued*)

Step 1: Form a testable *hypothesis* from the research question.

We might predict that, “People who have had heart attacks and take prescribed antidepressants are more likely to survive in the years following the heart attack than those who do not take antidepressants.” We have taken a general idea about the benefits of a drug and stated it in a way that a research study can directly test.

Step 2: Decide how to *operationalize* the concepts in the study into measurable variables.

First, we would need to decide who qualifies as a “heart attack patient”: Would we include only those who had been hospitalized with severe heart attacks, or anyone with abnormal cardiac symptoms? These types of decisions will have implications for how we interpret the results.

We would also need to decide on the doses of antidepressant drugs to use and the time period to measure survival rates. How long would we need to follow patients to obtain an accurate sense of mortality rates? In this case, earlier research had focused on five-year mortality rates, so that would be a reasonable time period for this study as well.

Step 3: *Measure* the key concepts based on the decisions made in Step 2.

This step involves collecting data from participants and then conducting statistical analyses to test the hypothesis. We will cover the specifics of research designs beginning in Chapter 2 (2.1), but one good option would be to give antidepressant drugs to half of our sample and compare their survival rates with the half not given these drugs.

Step 4: *Explain* the results and tie the statistical analyses back into the hypothesis.

We would want to know whether antidepressant drugs did, indeed, benefit heart-attack patients and increase their odds of survival for five years. If so, our hypothesis is supported. If not, we would go back to the drawing board and try to determine whether a) something went wrong with the study, or b) antidepressant drugs actually do not have any benefits for this population. Answering these kinds of questions often involves conducting additional studies. Either way, the goal of this final step is to return to our research question and discuss the implications of antidepressant drug treatment for heart-attack patients.

Example 2—Language and Deception

In 1994, Susan Smith appeared on television claiming that her two young children had been kidnapped at gunpoint. Eventually, authorities discovered she had drowned her children in a lake and fabricated the kidnapping story to cover her actions. Before Smith was a suspect in the children’s deaths, she had told reporters, “My children wanted me. They needed me. And now I can’t help them” (*The Washington Post*, November 5, 1994, A15). Normally, relatives speak of a missing person in the present tense. The fact that Smith used the past tense in this context suggested to trained FBI agents that she already viewed them as dead (Adams, 1996).

(*continued*)

Research: Applying Concepts (*continued*)

Research Question

The story about Susan Smith highlights one way that people communicate differently when they are lying—they use past tense when present tense is more natural. This observation might lead us to ask, more broadly, “How do people communicate differently when they are lying versus when they are telling the truth?” We will again apply the HOME paradigm (or scientific method) to design a study that will ideally provide insight into this question.

Step 1: Form a testable *hypothesis* from the research question.

This example is somewhat more challenging because “communicating” can be defined in many ways. Thus, we need a hypothesis that will narrow the focus of our study. It turns out several studies have been conducted on the ways that people communicate when they are lying, ranging from variations in speech rate to differences in the use of certain types of words (for a review, see Depaulo et al., 2003). Based on one of these studies, we could offer the following specific prediction: “Liars communicate using more negative emotion (e.g., anger, fear) than truth-tellers do” (e.g., Newman, Pennebaker, Berry, & Richards, 2003). We have taken a general idea (“communicate differently”) and stated it in a way that can be directly tested in a research study (“use more negative emotion”).

Step 2: Decide how to *operationalize* the concepts in our study into measurable variables.

To determine measurable variables, we need to decide what counts as “using more negative emotion.” We could take the approach used in a previous study (Newman et al., 2003) and scan the words people use, looking for those reflecting emotions such as anger, anxiety, and fear. The theory behind this approach posits that the words people use reflect something about their underlying thought processes. In this case, people who are trying to lie will be more anxious and fearful as a result of the lie, and therefore use more words indicative of these negative emotions.

Step 3: *Measure* the key concepts based on the decisions made in Step 2.

To measure the variables identified in Step 2, we must devise a way to determine whether and when people are lying. One way to do this in a research study is to instruct some people to lie and others to be truthful and then compare differences in the amount of negative emotion language between these groups.

Step 4: *Explain* the results and tie the statistical analyses back into the hypothesis.

We want to know whether people who were instructed to lie indeed used more words suggestive of negative emotion. If so, this outcome supports our hypothesis. If not, we would go back to the drawing board and try to determine whether a) the study design was flawed, or b) people in fact do not use more negative emotion when they lie. Either way, the goal of this final step is to return to our research question and discuss the implications for understanding language-based indicators of deception.

Goals of Science

In addition to sharing an overall approach to answering questions, all forms of scientific inquiry tend to adopt one of four overall goals. This section provides an overview of these goals, with a focus on how they apply to psychological research. We will encounter the first three goals throughout the course and use them to organize our discussion of different research methods.

Description

One of the most basic research goals is to describe a phenomenon, including descriptions of behavior, attitudes, and emotions. Most people are probably very familiar with this type of research because it tends to crop up in everything from the nightly news to their favorite magazine. For example, if CNN reports that 60% of Americans approve of the president, it is describing a trend in public opinion. Descriptive research should always be the starting point when studying a new phenomenon. That is, before we start trying to explain why college students binge drink, we need to know how common the phenomenon is. We might, therefore, start with a simple survey that asks college students about their drinking behavior, and we might find that 29% of them show signs of dangerous binge drinking. Having described the phenomenon, we are in a better position to conduct more sophisticated research. (See Chapter 3 for more detail on descriptive research.)



Stockbyte/Thinkstock

Before a phenomenon can be explained it must first be described. For example, a survey might be used to collect information to describe the phenomenon of binge drinking.

Prediction

A second goal of research is predicting a phenomenon. This goal takes us from describing the occurrence of binge drinking among college students to attempting to understand when and why they do it. Do students give in to peer pressure? Is drinking a way to deal with the stress of school? We could address these questions by using a more detailed survey that asked people to elaborate on the reasons that they drink. The goal of this approach is to understand the factors that make something more likely to occur. (See Chapter 4 for more detail on the process of designing surveys and conducting predictive research.)

Explanation

A third, and much more powerful, goal of research is to attempt to explain a phenomenon. This goal moves from predicting relationships to drawing stronger conclusions about causal links. Whereas predictive research attempts to find associations between two phenomena (e.g., college student drinking is more likely when students are stressed), explanatory research attempts to make causal statements about the phenomenon of interest (e.g., stress causes college students to drink more). This distinction may seem subtle at this point, but it is

an important one, and closely related to the way that psychologists design their studies. (See Chapter 5 for more detail on explanatory research.)

Change

The fourth and final goal of research is generally limited to psychology and other social-science fields: When we are dealing with questions about behaviors, attitudes, and emotions, we can sometimes conduct research to try to change the phenomenon of interest. Researchers who attempt to change behaviors, attitudes, or emotions are essentially applying research findings towards the goal of solving real-world problems.

In the 1970s, Elliot Aronson, a social psychologist at the University of Texas at Austin, was interested in ways to reduce prejudice in the classroom. Research conducted at the time was discovering that prejudice is often triggered by feelings of competition; in the classroom, students competed for the teacher's attention. Aronson and his colleagues decided to change the classroom structure in a way that required students to cooperate in order to finish an assignment. Essentially, students worked in small groups, and each person mastered a piece of the material. Aronson found that using this technique, known as the "jigsaw classroom," both enhanced learning and decreased prejudice among the students (Aronson, 1978). Read the details of Aronson's study here: <http://www.jigsaw.org/>.

Aronson's research also illustrates the distinction between two categories of research. The first three goals we have discussed fall mainly under the category of **basic research**, in which the primary goal is to acquire knowledge, with less focus on how to apply the knowledge. Scientists conducting basic research might spend their time trying to describe and understand the causes of binge drinking but stop short of designing interventions to stop binge drinking. Researchers more often involve for this fourth goal of research in **applied research**, in which the primary goal is to solve a problem, with less focus on why the solution works. Scientists conducting applied research might spend their time trying to stop binge drinking without becoming caught up in the details of why these interventions are effective. Aronson's research serves as a great example of how these two categories can work together. The basic research on sources of prejudice informed his applied research on ways to reduce prejudice, which in turn informed further basic research on why this technique is so effective.

One final note on changing behavior: Any time researchers set out with the goal of changing what people do, their values enter the picture. Inherent in Aronson's research was the assumption that prejudice was a bad thing that needed to be changed. Although few people would disagree with him, he risked the difficulty of remaining objective throughout the research project. As we suggested earlier, the more emotionally involved we are in the research question, the more we have to be aware of the potential for bias, and the more closely we must pay attention to the data.

Approaches to Science: Quantitative versus Qualitative Research

Imagine for a moment that a psychologist wants to study depression across the life span. The researcher might approach this research question in one of two ways. She could design a survey that asked people to report their experiences with depression, as well as how often they had experienced various positive and negative life events. By conducting statistical

analyses of these reports, she could gain a broad understanding of the relationships between life events and the development of depression. Alternatively, the investigator could spend her resources interviewing people who had been diagnosed with depression. Her goal is trying to understand what the experience felt like and whether people believed that it started in response to some major life event. This approach would provide a very deep understanding of the experience of depression from the inside out.

These alternative approaches highlight the differences between quantitative research and qualitative research, respectively. **Quantitative research** is a systematic and empirical approach that attempts to generalize results to other contexts. By surveying the population using structured scales, our hypothetical psychologist could learn about depression and life events in general. **Qualitative research**, in contrast, is a more descriptive approach that attempts to gain a deep understanding of particular cases and contexts. By interviewing depressed people in detail, the hypothetical psychologist could learn a great deal about how individuals experience depression.

The two approaches have traditionally been popular with different social science fields. For example, much of the current research in psychology is quantitative because the research aims for generalizable knowledge about behavior and mental processes. In contrast, much of the current research in sociology and political studies tends to be qualitative because research aims for a rich understanding of a particular context. To understand why college students around the country suffer from increased depression, quantitative methods are the better choice. To understand why the citizens of Egypt revolted against their government, then qualitative methods are more appropriate. However, many psychological phenomena are best understood by starting from the ground up, with a rich, qualitative understanding of people's experiences. As later chapters will discuss, the qualitative approach has been used to gain insight into questions ranging from forming stigmatized identities to helping children cope with traumatic events.

In an ideal world, a true understanding of any phenomenon requires the use of both methods. That is, researchers can best understand depression if they both study statistical trends and conduct in-depth interviews with depressed people. Researchers can best understand binge drinking by conducting both surveys and focus groups. And investigators can best understand the experience of being bullied in school by both talking to the victims and collecting school-wide statistics. This text will discuss the ways that both approaches are used to shed light on pressing questions throughout the field of psychology. Table 1.2 compares the quantitative and qualitative approaches.

Table 1.2 Comparing quantitative and qualitative approaches

	Quantitative	Qualitative
Main Approach	Systematic, empirical, tries to generalize to other contexts	Descriptive, tries to gain rich understanding of a single context or example
Use of Hypotheses	Starting point for all quantitative research	Not necessary; hypotheses sometimes the result of qualitative study
Examples of Research	- Study depression by surveying the population - Study bullying by comparing reported incidents between schools	- Study depression by interviewing patients - Study bullying by interviewing bullies to understand their motivation

1.3 Hypotheses and Theories

The use of hypotheses is one of the key distinguishing features of quantitative research. Rather than making things up as they go along, scientists develop a hypothesis ahead of time and design a study to test this hypothesis. (Qualitative research, in contrast, often starts by gathering information and *ends* with a hypothesis for future inquiries.) This section covers the process of turning rough ideas about the world into testable hypotheses. We discuss the primary sources of hypotheses as well as several criteria for evaluating hypotheses. Watch the following video for an entertaining introduction to hypotheses and theories, which the chapter will then explore in detail: <https://www.youtube.com/watch?v=lqk3TKuGNBA>.

Sources of Research Ideas

Every study starts with an idea that researchers frame as a question. But where do all of these great ideas come from in the first place? Students are often nervous about starting a career in research for fear that they might not be able to come up with great ideas to test. In reality, though, ideas are easy to come by, a person knows where to look. The following material suggests some handy sources for developing research ideas.

Real-World Problems

A great deal of research in psychology and other social sciences is motivated by a desire to understand—or even solve—a problem in the world. This process involves asking a big question about some phenomenon and then trying to think of answers based on psychological mechanisms.

In 1961, Adolf Eichmann was on trial in Jerusalem for his role in orchestrating the Holocaust. Eichmann's repeated statements that he was only "following orders" caught the attention of Stanley Milgram, a young social psychologist who had just earned a Ph.D. from Harvard University and who began to wonder about the limits of this phenomenon. To understand the power of obedience, Milgram designed a well-known series of experiments that asked participants to help with a study of "punishment and learning." The protocol required them to deliver shocks to another participant—actually an accomplice of the experimenter—every time he got an answer wrong. Milgram discovered that two-thirds of participants would obey the experimenter's commands to deliver dangerous levels of shocks, even after the victim of these shocks appeared to lose consciousness. These results revealed that all people have a frightening tendency to obey authority.



Getty Images/Handout

Nazi Lieutenant Colonel Adolf Eichmann's claims during his trial that he was just "following orders" throughout the Holocaust inspired Stanley Milgram to conduct a groundbreaking study about obedience to authority.

We will return to this experiment in our discussion of ethics later in the chapter. Read more about Milgram and his landmark study on this website: <http://www.experiment-resources.com/stanley-milgram-experiment.html>.

Reconciliation and Synthesis

Ideas can also spring from resolving conflicts between existing ideas. The process of resolving an apparent conflict involves both **reconciliation**, or finding common ground among the ideas, and **synthesis**, or merging all the pieces into a new explanation. In the late 1980s, psychologists Jennifer Crocker and Brenda Major noticed an apparent conflict in the prejudice literature. Based on everything then known about the development of self-esteem, members of racial and ethnic minority groups would have been expected to have lower-than-average self-esteem because of the prejudice they faced. However, study after study demonstrated that, in particular, African-American college students had equivalent or higher self-esteem than European-American students. Crocker and Major (1989) offered a new theory to resolve this conflict, suggesting that the existence of prejudice actually grants access to a number of “self-protective strategies.” For example, minority group members can blame prejudice when they receive negative feedback, making the feedback much less personal and therefore less damaging to self-esteem. The results of this synthesis were published in a 1989 review paper, which many people credit with launching an entire research area on the targets of prejudice.

Learning From Failure

Kevin Dunbar, a professor at Dartmouth University, has spent much of his career studying the research process. That is, he interviews scientists and sits in on lab meetings in order to document how people actually do research in the trenches. In a 2010 interview with Jonah Lehrer, Dunbar reported the shocking statistic that approximately 50 to 75% of research results are unexpected. Even though scientists plan their experiments carefully and use established techniques, the data are surprising more often than not. But even more surprising was the tendency of most researchers to discard the data if it did not fit their hypothesis. “These weren’t sloppy people,” Dunbar commented. “They were working in some of the finest labs in the world. But experiments rarely tell us what we think they’re going to tell us. That’s the dirty secret of science.” The trick, then, is knowing what to do with data that make a particular study seem like a failure (Lehrer, 2009).

According to Dunbar, the secret to turning failure into opportunity is twofold: First, question assumptions about why the study feels like a failure in the first place. Perhaps the data contradict the hypothesis but can be explained by a new one, or perhaps the data suggest a dramatic shift in perspective. Second, seek new and diverse perspectives to help in interpreting the results. Perhaps a cognitive psychologist can shed light on reactions to prejudice. Alternatively, perhaps an anthropologist knows what to make of the surprising results of a study on aggression. Some of the best and most fruitful research ideas have sprung from combining perspectives from different disciplines. Sometimes, all that a strange dataset needs is a fresh set of eyes.

Research: Thinking Critically

The Psychology Behind Pricing

Throughout this textbook, we will use short articles about research results as a way to illustrate key points in the text. Follow the link below to an article by William Poundstone, a bestselling author and expert on the psychology of pricing decisions. In this article, Poundstone discusses the peculiar appeal of prices ending in the number “9” and reviews recent research on this appeal by a pair of consumer psychology researchers. As you read the article, consider what you have learned so far about the research process, and then respond to the questions below.

<https://www.psychologytoday.com/blog/priceless/201001/does-9-just-sound-cheap>

Think About It:

1. What hypothesis are Coulter and Coulter trying to test? Try to state this as succinctly as possible.
2. How was “perception of discounts” operationalized in their studies?
3. How were the key variables measured?
4. How do Coulter and Coulter explain their findings? Are there other possible alternative explanations?
5. Are these studies primarily aimed at description, explanation, prediction, or change? Explain.

From Ideas to Hypotheses

Once a researcher develops a research question, the next step is to translate that question into a testable hypothesis—the first step in the HOME method. Broadly speaking, hypotheses are developed in one of two ways: bottom-up and top-down. This section explores these options in more detail.

Bottom-Up—From Observation to Hypothesis

Research hypotheses are often based on observations about the world around us. For example, people may have noticed the following tendencies as they observe those around them:

- Teenagers do a lot of reckless things when their friends do them.
- Close friends and couples tend to dress alike.
- Everyone faces the front of the elevator.
- Church attendees sit and stand at the same time.

Based on this set of four observations, we could develop a general hypothesis about human behavior: People have a tendency to go along with the crowd and conform to group behaviors. This process of developing a general statement from a set of specific observations is called **induction**, and it is perhaps best understood as a “bottom-up” approach. In this case, we have developed our hypothesis about conformity from the ground up, based on observing behavioral tendencies.

The process of induction is a very common and useful way to generate hypotheses. Most notably, this process serves as a great source of ideas that are based in real-world phenomena. Induction also helps us to think about the limits of an observed phenomenon. For example, we might observe the same set of conforming behaviors and speculate whether people will also conform in dangerous situations. What if smoke started pouring into a room and no one else reacted? Would people act on their survival instinct or conform to the group and stay put? Social psychologists Bibb Latané and John Darley (1969) conducted just such an experiment with groups of college undergraduates. Participants were asked to sit in a classroom and complete a survey. Meanwhile, the experimenters piped in smoke (actually dry ice) through the air vents. They hypothesized—and found—that the pressure to conform was stronger than the instinct to flee from a potential fire.

Top-Down—From Theory to Hypothesis

The other approach to developing research hypotheses is to work down from a bigger idea. The term for these big ideas is a **theory**, which refers to a collection of ideas used to explain the connections among variables and phenomena. For example, the theory of evolution organizes knowledge about how species have developed and changed over time. One piece of this theory claims that human life originated in Africa and then spread to other parts of the planet. This idea in and of itself, however, is too big to test in a single study. Instead, researchers move from the “top down” and develop a specific hypothesis from a more general theory, a process known as **deduction**.

By developing hypotheses using a process of deduction, researchers’ biggest advantage is the ease of placing the study—and its results—in the larger context of related research. Because the hypotheses represent a specific test of a general theory, results can be combined with other research that tested the theory in different ways. For example, in the evolution example, a researcher might hypothesize that the fossils from human ancestors found in Africa would be older than those found in other parts of the world. If this hypothesis were supported, it would be consistent with the overall theory about human life originating in Africa. And as more and more researchers develop and test their own hypotheses about the origins of life, our cumulative knowledge about evolution continues to grow.

Table 1.3 presents a comparison of these two sources of research hypotheses, showcasing their relative advantages and disadvantages.

Table 1.3 Comparing sources of hypotheses

Deduction	Induction
“Top-down,” from theory to hypothesis	“Bottom-up,” from observation to hypothesis
Easy to interpret findings	Can be hard to interpret without prior research
Helps science build and grow	Helps understanding of the real world
Might miss out on new perspectives	Great way to discover new ideas

Evaluating Theories

While experiments are designed to test one hypothesis at a time, the overall progress in a field is measured by the strength and success of its theories. If we think of hypotheses as individual combat missions on the battlefield, then theories are the overall battle plan. So, how do researchers know whether their theories are any good? Next, we cover four criteria that are useful in evaluating theories.

Explains the Past; Predicts the Future

One of the most important requirements for a theory is that it be consistent with existing knowledge. If a physicist theorized that everything on earth should float off into space, that theory would conflict with millennia's worth of evidence showing that gravity exists. Similarly, if a psychologist argued that people learn better through punishment than through rewards, that theory would conflict with several decades of research on learning and reinforcement. A new theory should offer a new perspective and a new way of thinking about familiar concepts, but it cannot be so creative that it clashes with what scientists already know. On a related note, a theory also has to lead to accurate predictions about the future, meaning that it has to stand up to empirical tests. There are usually multiple ways to explain existing knowledge, but not all of them will be supported as researchers test their assumptions in new circumstances. At the end of the day, the best theory is the one that best explains both past and future data.

Testable and Falsifiable

Second, a theory needs to be stated in such a way that it leads to testable predictions. More specifically, a theory should be subject to a standard of **falsifiability**, meaning that the right set of conditions could prove it wrong (Popper, 1959). Calling something “falsifiable” does not mean it is false, only that if it were false, demonstrating its falsehood would be possible. The Darwinian theory of evolution offers an example of this criterion. One of the primary components of evolutionary theory is the idea that species change and evolve from common ancestors over time in response to changing conditions. So far, all evidence from the fossil record has supported this theory—older variants of species always appear farther down in a fossil layer. If conflicting evidence ever were to appear, however, it would deal a serious blow to the theory. The biologist J. B. S. Haldane was once asked what kind of evidence could possibly disprove the theory of natural selection, to which he replied, “fossil rabbits in the Pre-Cambrian era”—that is, a modern version of a mammal buried in a much older fossil layer (Ridley, 2004).



oodelay/iStockphoto/Thinkstock

The theory of evolution is falsifiable, meaning that it could be disproved under the right conditions, such as the discovery of fossil evidence contradicting the theory.

Research: Thinking Critically

Intelligence, Politics, and Religion

Follow the link below to an article by Daniela Perdomo, a staff writer and editor for Alternet. In this article, Perdomo reviews the controversy over a recent scientific study claiming that liberals and atheists are more intelligent. As you read the article, consider what you have learned so far about the research process, and then respond to the questions below.

http://www.alternet.org/story/145903/controversy_grows_over_study_claiming_liberals_and_atheists_are_smarter

Think About It:

1. What general theory is Kanazawa trying to test? How does the theory differ from his specific hypothesis?
2. How did Kanazawa operationalize liberalism and intelligence in his research? Are there problems with the way these constructs were operationalized? Explain.
3. What were Kanazawa's main findings? How is the strength of this evidence influenced by his research methods?
4. Why do you think this research is controversial? If Kanazawa's methodology were more rigorous, would it still be controversial?

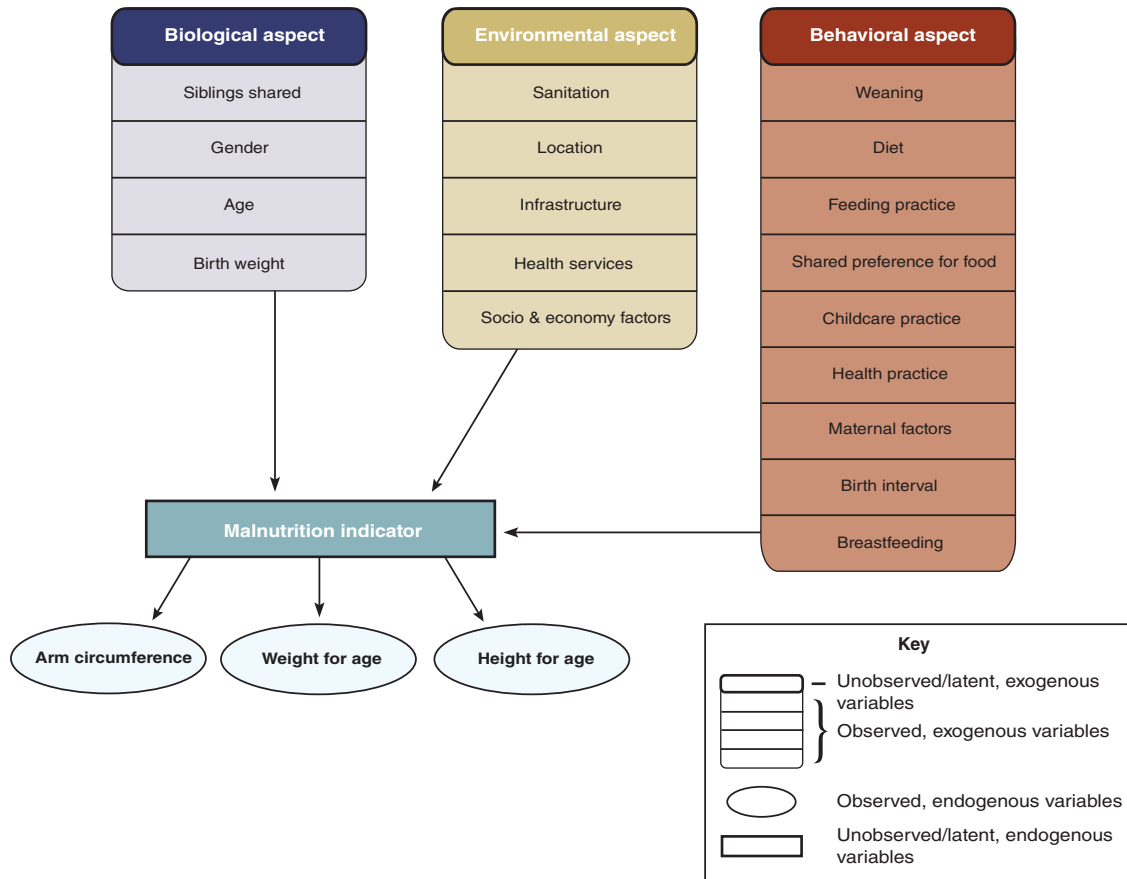
Parsimonious

Third, a theory should strive to be **parsimonious**, or as simple and concise as possible without sacrificing completeness. (Or, as Einstein [1934] famously quipped during a lecture at Oxford: "Everything should be made as simple as possible, but no simpler" [p. 165].) One helpful way to think about this criterion is in terms of efficiency. Theories need to spell out the components in a way that represents everything important but does not add so much detail that they become hard to understand. This means that theories can lack parsimony either because they are too complicated or because they are too simple.

At one end of this spectrum, Figure 1.1 presents a theoretical model of the causes of malnutrition (Cheah et al., n.d.). This theory does a superb job of summarizing all of the predictors of child malnutrition across multiple levels of analysis. The theory's potential problem, though, is that it becomes too complicated to test.

Figure 1.1: Predictors of malnutrition

Figure 1.1 presents a theoretical model of the causes of malnutrition.



At the other end of the spectrum, Figure 1.2 shows the overall theoretical perspective behind behaviorism. In the early part of the 20th century, the behaviorist school of psychology argued that everything organisms do could be represented in behavioral terms, without any need to invoke the concept of a “mind.” The overarching theory looked something like Figure 1.2, with the “black box” in the middle representing mental processes. Nevertheless, the cognitive revolution of the 1960s eventually displaced this theory, as it became clear that behaviorism was too simple. To strike an ideal balance, then, a researcher constructs a theory in a way that includes only the necessary pieces, nothing unnecessary.

Figure 1.2: The behaviorist model

Figure 1.2 presents the overall theoretical perspective behind behaviorism. The “black box” in the middle represents mental processes.



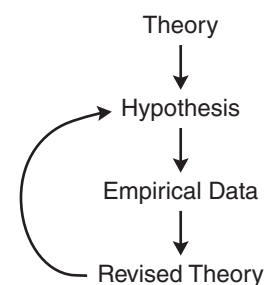
Promotes Research

Finally, science is a cumulative field, which means that a theory is really only as good as the research it generates. To state it more bluntly: A theory is essentially useless if no one follows up on it with more research. Thus, one of the best bases for evaluating a theory is whether it encourages new hypotheses. Consider the following example, drawn from real research in social psychology. Since the early 1980s, Bill Swann and his colleagues have argued that people prefer consistent feedback to positive feedback, meaning that they would rather hear things that confirm what they think of themselves. One provocative hypothesis arising from this theory proposes that people with low self-esteem are more comfortable with a romantic partner who thinks less of them than with one who thinks well of them. This hypothesis has been tested and supported many times in a variety of contexts and continues to draw people in because it offers a compelling explanation for why some people stay in bad relationships—a phenomenon that is regrettably recognizable. (For a review of this research, see Swann, Rentfrow, & Guinn, 2005.)

The Cycle of Science

Now, let us take a step back and look at the big picture. We have covered the processes of developing and evaluating both broad theories and specific hypotheses. Of course, none of these pieces occurs in isolation; science is an ongoing process of updating and revising our views based on what the data show. This overall process of quantitative research works something like the cycle depicted in Figure 1.3. Researchers start with either an overall theory or a set of observations about how concepts relate to one another and use this to generate specific, testable, and falsifiable hypotheses. These hypotheses then form the basis for research studies, which generate empirical data. Based on these data, we may have reason to suspect the overall theory needs to be refined or revised. And, so, we develop a new hypothesis, collect some new data, and either confirm or do not confirm our suspicion. The process does not end there, however: other researchers may see a new perspective on our theory and develop their own hypotheses, which lead to their own data and possibly to a revision of the theory. The scientific approach may strike some as a slow and strange approach to problem solving, but it is the most objective one available.

Figure 1.3: The cycle of science



Consider an example of how this cycle works in real life. In the 1960s, social psychologists were beginning to study the ways that people explain the behavior of others (e.g., when someone cuts me off in traffic, I tend to assume he is a jerk.) One early theory, called “correspondent inference theory,” argued that people would come up with these explanations in a rational way. For example, if we read a persuasive essay but then learn that the author was assigned a position on the topic, we should refrain from drawing any conclusions about the writer’s actual position. However, research findings demonstrated just the opposite. In a landmark 1967 study, participants actually ignored information about whether authors had chosen their own position on the issue, assuming instead that whatever they wrote reflected their true opinions (Jones & Harris, 1967). In response to these data (and similar findings from other studies), the correspondent inference theory was gradually revised to incorporate what was termed the “fundamental attribution error”—people tend to ignore situational influence

and assume that all behavior simply reflects the person's own disposition. The study's authors developed a theory, came up with a specific hypothesis, and collected some empirical data to test it. But because the data ran counter to the theory, the theory was ultimately revised to account for the empirical evidence. In this particular case, the cycle of research on understanding the fundamental attribution continues to this day, over 50 years later.

Proof and Disproof

While on the subject of adjusting theories, think about the notions of “proof” and “disproof.” Because science is a cumulative field, decisions about the validity of a theory are ultimately made based on results of several studies from several research laboratories. This means that a single research study has rather limited implications for an overall theory. This also means that a researcher must use the concepts of proof and disproof in the correct way. We will elaborate on this as we move through the course, but for now we can rely on two very simple rules:

1. If the data from one study are consistent with our hypothesis, we *support the hypothesis* rather than “prove” it. In fact, research almost never proves a theory, but statistical tests can at least suggest how confident to be in our support.
2. If the data from one study are not consistent with our hypothesis, we *fail to support the hypothesis*. As the course will discuss, many factors can cause a study to fail; these are often a result of flaws in the design rather than flaws in the overall theory.

1.4 Searching the Literature

Regardless of how a researcher develops a hypothesis, an important step in the process is to connect it with what has been done before. Scientific knowledge accumulates one study at a time, so the best studies will build on earlier studies—by extending, correcting, or contradicting them. On a practical note, to struggle over the best way to measure something when another researcher figured it out 20 years ago would be a waste of time. So, rather than reinvent the proverbial wheel, one of the first steps in a research project is to consult published relevant articles. This section will cover the process of finding these articles, followed by an overview of how to read these articles effectively.

Searching for Articles

Beginning a search for relevant research articles can seem like a daunting task, largely because of the sheer number of available sources. Should I ask a librarian? Search Wikipedia? Browse the web? Fortunately, we can use a few tricks to make sure that reference sources are both *objective* and *scholarly*. First, it is important to understand the difference between primary and secondary sources. **Primary sources** contain full reports of a research study, including information on the participants, the data collected, and the statistical analyses of these data. These types of sources appear in professional academic journals and are evaluated by a set of experts in the field before they are published—a process known as **peer review**. Thus, primary sources offer a reliable way to determine what has been done in a particular field.

Secondary sources, in contrast, consist only of summaries of primary sources. These types of sources include textbooks, some academic books, and review articles in journals such as *Psychological Bulletin*. As an analogy, think of the difference between someone telling his friends about an adventurous weekend (primary source) and one of those friends repeating the story to her roommate (secondary source). While some secondary sources undergo a process of review and evaluation (academic books), others do not (e.g., websites, friends re-telling stories).

In this day and age, people are becoming more and more comfortable searching for information via the Internet. It is particularly important, therefore, to note that websites are often not objective in their summaries of research. The vaccine/autism scare discussed at the beginning of the chapter is a good example of this point. A Google search for the terms *vaccine* and *autism* produces more than 4 million results, sorted in order of popularity. As of this writing, the top result is an article from WebMD describing the current controversy, followed by one by the Centers for Disease Control, arguing in favor of vaccines. At another time, and depending on recent events in the news, the top result might be celebrity Jenny McCarthy's website, which claims that vaccines gave her child autism. The key point is that search results in Google are not peer-reviewed, are not listed in order of reliability, and are customized to an individual's browsing history, confirming those biases. As a result, Google is a poor resource for finding trustworthy information about academic research.



anyaberkut/iStock/Thinkstock

College libraries provide students access to hard copies and digital copies of relevant research articles.

Another popular—but untrustworthy—source of information is Wikipedia. Wikipedia is a tempting resource, given its marketing as a “free online encyclopedia.” Unlike other encyclopedias, however, Wikipedia can be edited by anyone with access to the Internet. On the upside, this means that errors can be identified and corrected at any time. On the downside, this means that errors can be *made*—either accidentally or deliberately—at any time. The upshot is that there is no way to be sure that information is from a page at a time when it sticks to the facts.

What, then, does a researcher do? Fortunately, two reliable ways exist to access primary sources (research articles) that enable researchers to draw their own conclusions based on the patterns of data. First, Google Scholar (<http://scholar.google.com>) is a free resource, managed by Google, that works exactly like Google but is limited to peer-reviewed academic articles. Thus, Google Scholar provides one pipeline to access primary sources. Second, many university libraries have access to centralized databases of peer-reviewed articles. The

best-known database for psychology articles is PsycINFO; this database, maintained by the American Psychological Association, contains abstracts and citations for articles in psychology and related fields. PsycINFO is updated monthly and covers approximately 2,500 different primary-source academic journals.

Searching in PsycINFO (or Google Scholar) is as easy as typing key terms into a text box—sometimes labeled “Find,” or “Keywords.” That said, the process of choosing the best key words for a particular search can be a complex process. If search terms are too general, the search might yield too many results to be useful. If search terms are too specific, the search might yield only one or two articles and fail to represent fully prior studies. As an example, the following list of numbers represents different combinations of search terms related to the topic of self-esteem.

“self-esteem” (in all fields)	35,847 hits
“self-esteem” (title only; peer reviewed)	4,977 hits

Clearly, we need to narrow the field a bit—most of us have better things to do than review almost 5,000 abstracts. What aspect of self-esteem is most interesting? Perhaps we want to learn more about self-esteem and sexual behavior:

“self-esteem” and “condom use”	2 hits
--------------------------------	--------

Now we may have overdone the limits—two articles may not be very helpful in giving a sense of previous research. So, let us try one more combination, using a more general search term:

“self-esteem” and “sexual behavior”	133 hits
-------------------------------------	----------

This number is a bit more manageable; we could tinker a bit more, but it no longer seems overwhelming to skim through the search results and find the most useful articles. No two searches will be the same, so the real takeaway point is to try several combinations of search terms to strike a balance in the number of results.

Reading Research Articles

After assembling a collection of research articles relevant to the hypothesis, the researcher’s next step is to read them. This may sound painfully obvious, but psychological journal articles are written in a very formulaic way, which can be confusing at first glance. However, once we know what to look for, the format ultimately makes these articles easy to read (and easy to write). As a matter of fact, the format of a journal article is designed to follow the steps of the scientific method, with a section devoted to each of the four steps. This section examines each of the parts of a journal article to offer a sense of what to expect from each one. As a supplement to this discussion, the Appendix contains a copy of a real journal article, with the various sections and parts highlighted for easy reference.

The Title and the Abstract

At the top of every journal article—as well as in the search results in PsycINFO—appear both the title and an **abstract** or short summary of the article. While neither of these is a section per se, both provide the reader with a valuable first impression of the contents of the article. If a search query results in a large number of hits, a researcher can usually scan the titles to determine which ones are most likely to be useful. For example, if the research question concerns the links between depression and alcohol consumption among college students, the database might be searched for the terms “alcohol” and “depression.” Most of the results will be relevant and useful, but the researcher could probably skip an article with a title like “Fetal Alcohol Syndrome and Postpartum Depression,” since it is likely to be focused on a different population.

Once the list of results is narrowed to the most useful titles, the abstract provides additional information about the content of the article. A journal article abstract follows a standard formula of stating the objectives of the study, followed by information on the methodology, results, and conclusions. Generally, an abstract has to fit all of this information in about 150 words; as a result, it provides a concise summary that is worth reading carefully. In some cases, researchers decide after reading the abstract that this particular article is not relevant to their research.

The Introduction

The first main section of a journal article is the introduction, corresponding to the first step (i.e., hypothesize) of the four-step research process. As the name implies, the goal of this section is to introduce the research question, review background research, and state the hypothesis that was investigated. When diving into a new research area for the first time, it is a good idea to read the entire introduction carefully. This section provides the context for the rest of the paper, as well as a valuable introduction to previous work in the area.

The Method Section

The second main section of a journal article is the method section, corresponding to the second step (i.e., operationalize) of the four-step research process. The goal of this section is to explain how the hypothesis translated into a set of specific measurable variables and how the researchers gathered data to test their hypothesis. An additional—perhaps even more important—goal of this section is to provide enough detail about the study that someone could read the article and repeat the study.

The method section is typically divided into three parts: participants, materials, and procedure. The *participants* section describes the people who provided data for the study, including information about their age, gender, and other relevant information. For example, in a study on treatment of depression, the authors would specify whether the participants were “normal” college students or patients who have been hospitalized for treatment of severe, clinical levels of depression. The *materials* section describes any questionnaires or equipment that were used in the study, including both standardized measures and ones that the researchers created. The third and related section, *procedure*, provides all of the details regarding the

execution of the experiment. What did participants experience and in what order? If specific instructions were given before a task, what were they?

The materials and procedure sections are crucial for two reasons. First, they provide the necessary detail for someone else to recreate the study. While reading these sections, focus on understanding the key variables and how they were defined. Second, they allow readers to envision the study from the perspective of the participants and to decide whether the authors' interpretation of the results is the only one. For example, the authors might claim that participants were placed under stress and that the results showed a drop in concentration because of the stress. The procedure section, though, might suggest to the reader that the "stress" part of the study is more likely to invoke boredom. This could generate an idea for a follow-up study: Perhaps people actually lose concentration when they are bored.

The Results Section

The third main section of a journal article is the results section, corresponding to the third step (i.e., measure) of the four-step research process. This section aims to describe how the data were analyzed and to report the results of these analyses. The results section consists primarily of statistical analyses and, as Jordan and Zanna (1999) put it, "statistics can be intimidating" (p. 356). When students first start to read journal articles, the statistics can indeed seem overwhelming, but there are two reasons not to get discouraged. First, statistical results are always followed by a translation into plain English and almost always by tables and graphs of the data. This course will provide the opportunity to practice interpreting results in both statistical and graphical form. And this point brings us to the second reason not to become discouraged: The statistics *stop* being intimidating surprisingly quickly. The more we read journal articles and place them in the context of our own ideas, the more comfortable we become interpreting statistical analyses. In fact, as we become savvier with interpreting statistics, we may be surprised by how often authors make mistakes in either their analyses or their interpretations of them.

The Discussion Section

The fourth and final section of a journal article is the discussion section, corresponding to the fourth (i.e., explain) step of four-step research process. The goal of this section is to summarize the main findings and provide an evaluation of the hypothesis. Thus, the first few paragraphs of the discussion often supply a good summary of the entire article. Authors state whether their predictions were confirmed and speculate on the meaning of the findings. If some of the predictions were not confirmed, authors suggest explanations for this and either acknowledge or defend potential flaws in the study. In addition, to encourage others to follow up on the study, authors tie their findings to previous literature and make suggestions for future research.

Evaluating Articles

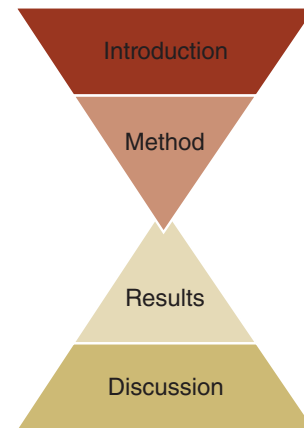
So, in sum, a journal article will follow a predictable structure: Authors first describe the problem and state their hypothesis (introduction), then explain their approach to testing the hypothesis (method), then report the findings of this test (results), and finally discuss the

meaning of these findings relative to the hypothesis (discussion). These four sections are often described as following an hourglass structure—that is, the paper starts broadly in the introduction, narrows to the specific details of the study, and ends broadly in the discussion by tying everything back into the overall problem (e.g., Bem, 1987). Figure 1.4 depicts this structure.

Before moving on, we will review some general guidelines for evaluating journal articles. After reading the paper in its entirety, use the following five questions to form an overall evaluation of the paper.

1. **What am I being asked to believe?** What is the author's main argument? Before critiquing in detail, make sure you understand the argument completely and can summarize it in a few sentences.
2. **What evidence supports this claim?** How does the author support the main argument? If it is an empirical paper, look to the data; if it is a theoretical paper, look at the literature the author summarizes.
3. **Are there alternative explanations?** Be creative here. Based on your reading of the article, what else seems plausible? To make your critique a good one, though, you should be able to test it.
4. **What additional evidence would help us test alternatives?** This question is one of the keys to performing good science. Once you identify something wrong with the original study, how can you test your alternative?
5. **What conclusions are reasonable?** Return to step 1 with your critiques in mind. What should the author reasonably conclude, given the problems with the study?

Figure 1.4: Structure of journal articles



1.5 Ethics in Research

In the summer of 1971, psychologist Phillip Zimbardo conducted an experiment at Stanford University to test the power of social roles. Zimbardo hypothesized that people would take on the characteristics and behaviors of whatever role was assigned to them, and he tested this by creating a simulated prison in the basement of the psychology building. A group of 24 psychologically healthy young men were selected from the San Francisco Bay area and randomly assigned to play the role of either “prisoner” or “guard.” Zimbardo appointed himself to the role of “warden.” The researchers gave each participant pieces of a uniform meant to reinforce their role—smocks for the prisoners, khakis and mirrored sunglasses for the guards. Almost immediately, and without instructions from the researchers, participants began to act out their roles. The guards took it upon themselves to establish control and dominate the prisoners by withholding privileges and devising clever ways to humiliate them. The prisoners, in turn, accepted all of this without much protest since it was part of their prisoner role. The experiment was scheduled to run for 14 days but was stopped after only six because the situation was out of control—prisoners were going on hunger strikes and being locked in solitary confinement, and one even suffered a serious mental breakdown. This study is known as the Stanford Prison Experiment; learn more about it and view video clips on a website designed by Zimbardo and his colleagues: <http://www.prisonexp.org/>.



©Duke Downey/San Francisco Chronicle/San Francisco Chronicle/Corbis

The Stanford Prison Experiment raised ethical concerns in the scientific community about how research is conducted.

For many, this experiment calls to mind the real-life prisoner abuse at Abu Ghraib prison in Iraq. A group of American military guards stationed at this prison during the Iraq war were caught treating the prisoners in remarkably similar ways to the “guards” at Stanford—inflicting cruel punishments and humiliations and photographing the entire ordeal. Interestingly, Zimbardo was even called to testify about the power of social roles during the trial of one of the Abu Ghraib guards. Zimbardo’s experiment also strikes many people as ethically dubious. When the research was published, it raised serious questions about the amount of distress that can be inflicted in the name of research. Although the proposal for this study was approved under ethics standards of the period, today’s more stringent standards would not allow it. But how might researchers today balance the distress of the “prisoners” with the valuable knowledge gained from the study? Should the Stanford Prison Experiment ever have been run? Does the knowledge outweigh the distress? Before moving on to the nuts and bolts of research design in the next four chapters, it is important to spend some time on the ethics of conducting research.

At the most basic level, all deliberations about the ethics of a particular study come down to the balance between a) avoiding all unnecessary discomfort for participants; and b) finding a way to capture real-world attitudes and behaviors to provide a valid test of the hypothesis. In practice, however, achieving this balance can be complicated. This section first examines some of the potential threats to participants’ well-being and then discusses how avoidance of these threats has been formalized into rules for researchers. Finally, evaluate a set of ethical dilemmas that represent the types of issues likely to arise in psychological studies.

Threats to Participants

To explain the need for ethical guidelines, this section introduces some of the possible threats to participants’ welfare in the context of research studies.

Physical Harm

We will start with the most extreme threat: Sometimes a research paradigm, or worldview, can place participants at risk for physical harm. As a general rule, these types of studies are limited to the medical field. For example, a researcher testing a new medication for heart-attack survivors faces the risk that an unexpected side effect could hasten the participants’ death. Alternatively, perhaps the participants could have benefitted more from another, more established medication, but they were not taking it because they were participating in the study to assess the new medication. Because of these risks, medical researchers are required

to perform preliminary testing—often using cell cultures and then animals—before testing drugs on a human population.

Occasionally, psychological research can pose a physical threat to participants, albeit a more minor one. As one example, for the past 25 years, Sheldon Cohen has been conducting studies in which he exposes participants to the common cold virus and measures the development of cold symptoms for several days. This work is designed to explore the link between individuals' social environment and their susceptibility to illness; learn more about it on Dr. Cohen's website: <http://www.psy.cmu.edu/~scohen/>. While the cold virus can be considered a physical threat, it is mild in comparison to the knowledge gained from these studies.

Extreme Stress

More common among psychological studies are those that introduce high levels of mental or emotional stress for participants. As the text will discuss later, the key in evaluating whether a stressful research paradigm is ethical is to think about whether—and to what extent—it exceeds the stress that participants encounter in everyday life. In the Stanford Prison Experiment, it is easy to see how the stress experienced by the “prisoners” exceeded normal levels. In 1924, Carney Landis conducted the first studies of facial expression and emotion. His goal was to map specific emotional states onto specific expressions—work that is now associated with Paul Ekman (and popularized by the television show *Lie to Me*). Landis photographed his participants as they reacted to a variety of stimuli such as smelling ammonia and viewing pornography. But the most shocking and controversial task was the final one. To measure responses to “disgust,” Landis asked his participants either to decapitate a live rat (a task they lacked the training to perform humanely), or to watch Landis behead the rat. In this case, the discomfort could not even be balanced by the knowledge gained from it; Landis found no support for his hypotheses regarding common facial expressions. Of course, this study is beyond anything deemed ethically acceptable by today's standards.

In reality, most research, and particularly psychological research, presents a much more minor degree of stress to participants. For example, one very common task used in social psychology research is to observe college students' reactions as they are asked to prepare and give a speech. Most people become anxious at the thought of public speaking, but this anxiety is mild and very much temporary. In fact, among studies that receive approval from ethics review boards, the effects of the research on overall well-being are always mild and temporary.

Deception

Finally, at the lower end of the threat spectrum, many psychological studies involve deceiving participants about the purpose of the research—at least until the study is finished. This deception is a way to ensure people's honest reactions to the experimental setting. If, for example, participants in Milgram's obedience studies had known he was studying obedience, they would have reacted very differently when asked to shock the confederate, and the study would have been pointless. As an analogy, think of how car salespeople attempt to bond and form a relationship with customers by asking personal questions. Odds are that this trick would backfire if a car salesperson said “I bet if I ask about your family, you'll start to bond with me and be more likely to buy a car!” Later chapters will elaborate, but people tend to change their behavior when they figure out the research question (as well as when they *think* they figure it out).

Deception is included here as a threat because of the potential for abuse. The history of science is rife with examples of medical research conducted on unsuspecting (and unwilling) participants. In one of the most infamous, researchers in Tuskegee, Alabama, conducted a study of the natural progression of syphilis among poor African-American farmers. The study began in 1932 under the supervision of the Public Health Service and continued until 1972. What was the deception? As it turned out, penicillin was discovered to be a reliable cure for syphilis—in 1945. The researchers not only lied about the purposes of the study (participants were never told they had syphilis), but they deliberately withheld treatment so as to continue the study. (Read more about the study here: <http://www.cdc.gov/tuskegee/timeline.htm>.)

On the one hand, these types of studies are vastly different from research that could be approved today, much less the type of research conducted in psychology. On the other hand, all researchers must be mindful at all times that they do not abuse the trust of participants. The chapter will later return to the issue of deception in the discussion of evaluating a set of research scenarios.

Ethical Guidelines

In response to public outcry over the Tuskegee Syphilis Study, the U.S. Congress formed a panel to develop guidelines that would ensure that all human subjects were treated ethically. In 1979 this committee published the Belmont Report, a document that outlined a set of basic ethical principles for the use of human subjects. (The full report is available at <http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html>.) Essentially, the Belmont Report guidelines argue for *treating participants with respect, minimizing harm, and avoiding exploitation*. These principles were formalized into a set of federal laws referred to as the **common rule**, a baseline standard of ethics for all federally funded research.

One critical part of the common rule was the creation of review boards to evaluate the ethics of every proposed research study. The common rule mandated that any institution receiving federal money must have an **institutional review board (IRB)**, which reviews and monitors all research involving humans to protect the welfare of research participants. The IRB is tasked with determining whether a study is consistent with ethical principles, and it has the authority to approve, reject, or require modification of any research proposal. To put it another way, the IRB serves as a gatekeeper for research, ensuring that something like the Tuskegee Syphilis Study, or Landis's "facial expression" studies could not be run today.

An important piece of IRB review is to assess the *degree of risk* that a study poses for participants. Based on these assessments, each proposed study undergoes one of three categories of review. The lowest risk studies are subject to **exempt review**, in which an IRB representative simply verifies the low risk and approves the study. To qualify for exempt review, a study has to fit into one of six predefined categories, including research done in educational settings (e.g., testing a new way to teach reading skills), and reanalysis of existing data (e.g., looking for patterns in polling data). The full set of guidelines is available online at <http://www.mayo.edu/research/institutional-review-board/policy-manual>.

Studies classified as medium risk—including the majority of psychological studies—are subject to **expedited review**, in which an IRB representative conducts a full review of the proposed study's procedures and ensures that participants' welfare and identity are protected.

Expedited review also requires that a study fit into one of seven predefined categories (<http://www.hhs.gov/ohrp/policy/expedited98.html>). These categories encompass most of the research that psychologists conduct, even when these studies include collection of personal information and biological specimens. The key to meeting expedited-review criteria is keeping the risk of harm and distress and the release of information to a minimum.

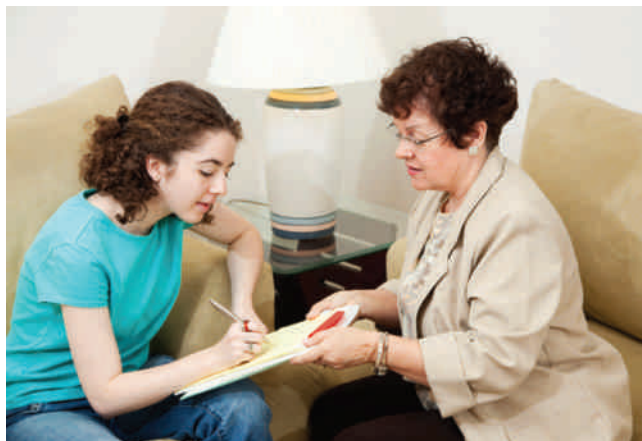
Finally, studies classified as high risk are subject to **full-board review**, in which all members of the IRB review the proposed study's procedures and then meet as a group to discuss the degree of risk and protection. This category includes studies involving medical procedures, children, prisoners, or pregnant women. Any time there is potential for physical harm, release of confidential information, or undue pressure on people to participate (e.g., prisoners), the IRB pays careful attention to the procedures for minimizing these risks.

The IRB also weighs the potential risks of the study against the potential benefits. These decisions are always made case by case, taking into account the goals of the study. For example, imagine that researchers were developing a drug that could cure previously fatal birth defects with an injection while the mother was still pregnant. Imagine also, however, that this drug could cause serious risks to the mother's health in about 1% of cases. Is this risk worth the potential benefits? Is the small risk to the mother a reason to veto the study? Questions like these are certainly not easy, but they will be at the heart of the IRB's deliberations.

The American Psychological Association (APA) has its own version of an ethical code, written specifically for the kinds of dilemmas faced by psychologists in both research and therapeutic settings. The APA ethics code lays out five specific rules for research that involves human participants. These rules, described in the sections that follow, take their inspiration from the Belmont guidelines—treat people with respect, minimize harm, and avoid exploitation. (View the full APA ethics code here: <http://www.apa.org/ethics/code/index.aspx>.)

1. Informed Consent

First and foremost, research participants must be “informed of all features of the study that would reasonably affect their decision to participate,” an ethical principle known as **informed consent**. Before people agree to take part in a study, they need to know whether it involves anything painful or uncomfortable or might reveal sensitive or embarrassing information. Participants need to be informed of the risks and benefits of participating. They also need to know how the researcher will protect the information that they provide. What if the study involves deception? This is where the phrase “reasonably affect their decision” becomes relevant. A research team pretending to study perception but actually studying conformity is under no obligation to reveal this before the study begins.



lisafx/iStock/Thinkstock

Participants in a study must indicate their informed consent before the experiment can begin.

However, if the study involves, say, running on a treadmill or taking drugs, people need to know that to make informed decisions about whether their overall health might affect their ability to participate.

2. Free Consent

Free consent forbids researchers from placing “undue pressure” on people either to participate in or to remain in a study. One lesson from the Milgram studies is that people are willing to obey seemingly strange commands from an experimenter wearing a lab coat. As researchers, we therefore have an obligation not to abuse this tendency to obey. No one probably needs to be told that it is wrong to recruit participants at gunpoint, but quite a few grey areas exist when it comes to free consent. For example, many psychology departments require students to participate in research studies or at least offer them extra credit for doing so. (There are always alternative ways to earn the credit.) Could students who are failing the class feel more compelled to agree to a research study? What about students who wait until the last minute and have fewer options? Free consent also becomes an issue when prisoners or soldiers serve as research subjects. Do these populations really feel free to say “no” to a request to participate? The answer to all of these questions depends on the context and will be weighed against the potential benefits of the study by the IRB.

3. Protection From Harm

Participants cannot be exposed to physical or emotional risk “beyond what they would encounter in real life.” The researcher who wishes to conduct a study with an aim to make participants clinically depressed is out of luck. Still, where does research ethics draw the line regarding “real life” harm? Is it acceptable to make people feel stupid or embarrassed? Is it all right to reject people from a group to observe their reactions? The answer, once again, depends on the context, more specifically on the balance of costs and benefits. If participants experience mild rejection for the sake of understanding how to cope with it, that is probably fine. However, requiring participants to experience severe verbal abuse for the sake of learning whether people like abuse is would not be acceptable by today’s standards. (If that one sounds made up, read about this study of stuttering from the 1930s: <http://www.spring.org.uk/2007/06/monster-study.php>)

4. Confidentiality

All personal information collected during the research study must be protected and prevented from being released to anyone not authorized to view it. If a researcher were to ask people about their history of drug use, such information, made available, could compromise their political prospects. Suppose employees were asked to report attitudes toward their managers; the managers who saw that information could retaliate against unfavorable ratings.

Protecting personal information involves two related options. First, whenever possible, responses should be **anonymous data**, meaning that the investigators do not collect any identifying information from participants. If participants cannot be individually identified,

the risk of retaliation or other backlash is eliminated. In some cases, however, anonymity is not possible, such as when a study needs to track people for a period of time and then link their data. In these situations, identifying information should be kept **confidential**, meaning that the information is collected but kept secret. One common way to do this is for researchers to maintain and closely guard a master list of participants matched to code numbers and identifiers, which are used during the study instead of names. Another approach, used by organizations conducting employee-satisfaction surveys, is to give one analyst in the company access to the raw data. That individual then provides aggregated summaries to the leadership teams. In this way, managers know which department is unhappy, but the responses of individual employees are kept confidential.

5. Debriefing

Finally, as the chapter earlier noted, many experiments cannot avoid using some degree of deception. In its list of ethical rules, the APA suggests a compromise regarding deception. First, researchers should employ it only when necessary, meaning that they should never create an elaborate cover story just for its own sake. Second, the study should always involve a **debriefing** of participants, in which they are informed of the true purpose once the study is concluded. In Milgram's obedience studies, participants went through a long debriefing that involved meeting the "victim" and understanding that they had not done any actual harm to another human being. If participants were under the illusion that a conformity study was focused on "perceptual processing," then researchers must tell them the truth at the end. If the study involved randomly rejecting participants from the group, then the researchers must tell them this decision was random. The goal of this disclosure is to remove possible negative effects of the study procedure and to explain why the deception was necessary to achieve the goals of the study.

Research: Applying Concepts

Ethical Dilemmas

To give you a feel for what these guidelines look like in everyday research studies, we will walk through a pair of experimental scenarios and evaluate whether they meet the APA guidelines.

Scenario 1 A cognitive psychologist wants to investigate whether different fonts are easier to skim, which could offer valuable information to website designers. She gives groups of students short articles to read in different fonts and measures how long it takes them to finish. To avoid biased responses, however, she tells students that she is interested in their reading comprehension rather than their speed.

Evaluation: The study poses no risk of physical harm or extreme stress, but participants have been deceived about the purpose of the study. APA's Rule 5 is most relevant, but any IRB is likely to approve the study, provided that participants are given a full debriefing at the end of it. The rationale for the deception seems sound; if participants were told that they were being timed for speed, they would likely read faster than normal, which would introduce bias into the data.

(continued)

Research: Applying Concepts (*continued*)

Think About It:

1. Do the benefits of this study (offer valuable information to website designers) outweigh the risks (deception about the purpose of the study)? Why or why not?
2. Could the researcher have found a way to obtain unbiased results without using deception? How?

Scenario 2 In a field experiment designed to test whether people would help more when they are alone or with others, male subjects walking alone or in a group were exposed to a simulated rape (Harari, Harari, & White, 1995). As subjects walked along, a male and female who were part of the research team acted out the rape. The man grabbed the woman around the waist, put his hand over her mouth, and dragged her into the bushes as she screamed for help. Observers stationed at various points recorded the number of subjects who offered help. Before they could actually intervene, a researcher stopped them and told them the “rape” was part of a study.

Evaluation: This study is likely to have induced extreme stress in participants and quite likely presented emotional risks beyond what participants normally encounter (Rule 3). In addition, participants did not give their consent to be in the study (Rule 1) until after their data were collected. However, this study was approved by a modern-day IRB, which means that at least one group of reviewers felt that these threats were outweighed by the benefits of the study.

Think About It:

1. Do the benefits of this study (understand helping behavior in emergencies) outweigh the risks (distress for participants)? Why or why not?
2. How might the study be redesigned to avoid extreme stress?
3. How would the study have been affected if researchers asked participants to sign a consent form in advance?

Ethics in Animal Research

So far, the discussion has focused on ethical issues in dealing with human participants. However, a significant portion of psychological research involves nonhuman animals. Studying the behavior of nonhuman animals provides an additional important avenue for understanding basic principles of behavior and ultimately for improving the welfare of both human and nonhuman animals. This is challenging terrain for many people. Some people object out of concern for animal welfare, while others feel that the benefits outweigh the discomfort caused to animals. What is more, each experiment is different in terms of the level of potential harm. Ultimately, each individual has to decide his or her position.

Worth noting for this discussion is the fact that most scientists favor the continued use of this practice, provided that the animals are treated humanely (Plous, 1996). This support centers around the argument that the benefits of animal research outweigh the costs. One of the most salient examples involves testing the effectiveness of drugs to cure cancer, depression, and so on. The first stage in testing these drugs is to examine chemical reactions in isolation, using test tubes and petri dishes. Before moving on to research involving humans, researchers are required to conduct safety testing of these drugs on animals.

To ensure that the nonhuman subjects involved in this research are treated humanely, the APA has also developed a set of guidelines, overseen by the **Committee for Animal Research and Ethics (CARE)**. (Read the CARE guidelines at <http://www.apa.org/science/leadership/care/guidelines.aspx>.) The upshot of these guidelines is to ensure that animals are *treated humanely* at all stages of the study by *well-trained personnel*, and that there is a strong *justification for their use*. And, just as research with human subjects is reviewed by an IRB, all research using animals is reviewed by the **Institutional Animal Care and Use Committee (IACUC)** to ensure that the benefits of the research outweigh any discomfort experienced by the animals.

Scientific Misconduct

Before leaving the subject of ethical conduct, we will consider a final important topic, one that has less to do with protecting participants' welfare and more to do with the overall ethics of research. Because science is a cumulative discipline, every research study contributes to the body of knowledge in that discipline. Our understanding of the development of aggression, the process of forming memories, and the mechanisms for coping with trauma all come from knowledge gained one study at a time. Therefore, when researchers do not accurately represent their data and publish dishonest results, their actions pose a serious threat to the cumulative body of knowledge. These types of violations are captured under the umbrella of **scientific misconduct**, defined as intentional or negligent distortion of the research process. To illustrate how this happens, this section describes two real cases of scientific misconduct, one probably "negligent" and the other very much intentional.

Negligent Misconduct—Race Differences in Skull Size

In the 19th century, physician Samuel Morton argued that he could measure the intelligence of a racial group by measuring its average skull size—bigger skulls would mean bigger brains and, therefore, more intelligence. (We now know that intelligence is much more complicated than this, but the science was young in the 1830s.) Morton's work is often credited with kick-starting more than a century's worth of racially tinged science by a subgroup of researchers who attempted to show that some races were superior to others. Stephen Jay Gould's 1996 book, *The Mismeasure of Man*, dissected and discredited this entire line of work, and science now takes for granted that Morton's work was terribly biased and fundamentally flawed. (For a short audio program that explains the context of Gould's book, see <http://www.uh.edu/engines/epi429.htm>.)

Gould was able to obtain access to all of Samuel Morton's laboratory notes, and the latter turns out to be a fascinating example of negligent misconduct. Morton's method of quantifying skull sizes was to pour mustard seed into the hole in the bottom and then measure the volume of mustard seed that each skull held. However, he was hardly consistent with his pouring: As he held a known European skull in his hand, he might pack with seed to make sure it was full. Yet as he held a known African skull, he might declare it full when there was still space at the top. Morton also discarded data from skulls that did not seem to fit the patterns and occasionally guessed at the race of a skull based on its size. Incredibly, he did not try to hide any of this.

Gould interpreted these facts as revealing that Morton believed so strongly in his hypothesis that his data collection was biased every step of the way. Morton's intentions were good, but the danger of this type of misconduct is that it can happen without anyone's knowledge.

Intentional Misconduct—Reactions to Discrimination

In the late 1990s, social psychologist Karen Ruggiero was interested in the way people responded to instances of discrimination and prejudice. Other researchers had documented a strange discrepancy among targets of prejudice: People perceive more discrimination directed at their group as a whole than at them personally (Taylor, Wright, Moghaddam, & LaLonde, 1990). Ruggiero argued that this tendency indicated a reluctance to admit to personal discrimination because it would mean acknowledging a lack of control over a person's own outcomes. For example, a woman might think, "I haven't personally seen any sexism because I'm in charge of my own destiny, but it's a big problem for other women."

In a compelling 1999 paper, Ruggiero reported that members of high-status groups were more likely to blame a negative event on discrimination, because that meant fewer implications for an individual's degree of long-term control. That is, if a white male law student failed to get accepted into his first choice of law school, he could claim he was discriminated against in favor of minority applicants, while still feeling in control of most other aspects of his life. Fascinating, right? Ruggiero's report had just one problem: Her data were completely fabricated. Not one of the 240 supposed participants actually existed; Ruggiero had written a piece of fiction and passed it off as a scientific journal article. This was her most egregious offense, but others surfaced as well. She fabricated partial data for another paper; she discarded participants that did not fit her hypothesis; she used federal grant money to pretend to collect these data; and she used these fake data to apply for future funding. Ruggiero was eventually caught and forced to submit retractions to several scientific journals to correct the fabricated publication. She was also forced to resign from her faculty position and barred from working on federally funded research for five years. (Read the official report of the investigation here: <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-020.html>.)

Prior to the scandal, Dr. Ruggiero had completed her Ph.D. at McGill University and had landed a prestigious faculty position at Harvard University before being wooed away to the University of Texas with a \$100,000 start-up package for setting up her laboratory. In short, she showed every sign of being a rising star in the field. So why would she take such a big risk? One of her fellow graduate students, interviewed for a 2002 article in *The Chicago Tribune*, suggested that she was motivated by a sincere belief in the work she was doing: "She was invested in proving people were denying discrimination . . . She knew what the answer ought to be." Another possible motivation has to do with the way incentives work for academic research. Science works one slow step at a time, but people are often rewarded for making a big, counterintuitive splash. Ruggiero was certainly rewarded for her efforts—at least in the short term.

This case is fascinating because it sheds real light on the scientific process. Ruggiero's deception was ultimately uncovered because other people tried to recreate her experiments. Again, this is how science works—one finding does not carry much weight until others can repeat it in their own laboratories. However, because the discrimination data were fictional, subsequent researchers could find no way to replicate them. So, people started talking at conferences, which eventually led to official questions, and the rest is history.

The silver lining to this story is that it illustrates the strength of the scientific approach. Sometimes, the approach is self-correcting, and people who attempt to cheat the system are caught. Unfortunately, other cases of misconduct will always slip through the cracks into the permanent record. One strategy that researchers can use to help sort out the truth is to place more stock in findings that are replicated by different researchers, and to be cautiously optimistic about those amazing new counterintuitive findings. Retraction Watch (<http://retractionwatch.wordpress.com/>) is a website that tracks retractions of journal articles. This blog highlights problematic research, including faked experiments and plagiarized articles.

Summary and Resources

Chapter Summary

This chapter has provided an introduction to the scientific approach to problem solving. First, it discussed the major research areas in psychology with an eye toward gaining insight into the field as a whole. It then covered the four steps of the research process: forming a research question, deciding how to test it, collecting data, and interpreting the results. The key distinguishing feature of scientific thinking is that the decision-making process is based on empirical evidence. If data run counter to a researcher's initial predictions—especially if this happens over and over again—then the researcher has to conclude that the prediction was wrong. The scientific method demands that when considering a question, any conclusions drawn must be based on facts. Do vaccines cause autism? Is the planet getting warmer? What is the best way to improve children's reading skills? In every case, researchers collect an appropriate set of data and then decide, regardless of whether the answer fits their preconceived notions or what they want to be true.

The first and most important step of the research process is to form a testable and falsifiable research hypothesis. The text explained the process of developing hypotheses and of placing them in the broader context of research in the field. Broadly speaking, hypotheses can be developed in one of two ways. Induction is a bottom-up process that involves trying to generalize from observations about the world. Deduction is a top-down process that involves trying to generate a specific prediction from a broader theoretical perspective. One of the key points from this section is that science is a cumulative discipline, meaning that knowledge in a particular field grows and accumulates with each study. The theory of evolution sprang not from a single fossil discovery but from the combined evidence of thousands of fossils and ethological studies. Thus, it is particularly important that each study be placed in the proper context of prior studies, and this requires the ability to find and digest peer-reviewed journal articles that are relevant to a given research question.

The final section of this chapter emphasized the importance of ethics in conducting research. Any time research involves human or nonhuman animals, researchers have to protect the rights of these participants. The history books are full of abuses of human participants, such as deceiving people about the diseases they had and subjecting them to extreme stress—to say nothing of the horrors inflicted by Japanese and Nazi doctors on prisoners during World War II. In response to these and countless other, more minor abuses, the U.S. federal government has mandated that all research treat participants with respect, minimize harm, and avoid exploitation. The American Psychological Association has its own guidelines governing psychological research studies: Participants must give both informed and free consent; they

must be protected from undue harm; their personal information must be protected; and they must be told the full purpose of the study at its conclusion. Finally, the chapter covered the subject of scientific misconduct, which includes all distortions of the research process. As the text explained, these distortions can be either negligent or intentional. The beauty of the scientific process is that those who attempt to commit fraud can sometimes get caught by the system that they are trying to cheat.

Key Terms

abstract A summary of a journal article that appears both at the top of the article and in search results.

anonymous data Data collected without identifying information from participants.

applied research Research in which the primary goal is to solve a problem, with less focus on why the solution works.

basic research Research in which the primary goal is to acquire knowledge, with less focus on how to apply the knowledge.

biopsychology The study of connections between biological systems (including the brain, hormones, and neurotransmitters) and our thoughts, feelings, and behaviors.

clinical psychology An applied field focused on understanding the best ways to treat psychological disorders; the study of best practices for understanding, treating, and preventing distress and dysfunction.

cognitive psychology The study of internal mental processes, including the ways that people think, learn, remember, speak, perceive, and so on.

Committee for Animal Research and Ethics (CARE) APA committee responsible for guidelines governing animal research; the upshot of these guidelines is to ensure that animals are treated humanely at all stages of the study by well-trained personnel and that there is a strong justification for their use.

common rule A set of federal laws that established the baseline standard of ethics for all federally funded research.

debriefing Ethical principle stating that participants should be informed of the study's true purpose when it is concluded.

deduction The process of developing a specific hypothesis out of a more general theory; best understood as a "top-down" approach.

developmental psychology The systematic study of physical, social, and cognitive changes over the human life span; initially focused on childhood development, though many researchers now study changes and key stages over the entire life span.

exempt review Category of IRB review reserved for low-risk studies falling into a set of predefined categories; involves having an IRB representative simply verify the low risk and approve the study.

expedited review Category of IRB review used for medium-risk studies falling into a set of predefined categories; involves having an IRB representative conduct a full review of the study procedures and ensure that participants' welfare and identity are protected.

falsifiability A concept applied to theories and hypotheses meaning that the right set of conditions could prove it wrong; calling something falsifiable does not mean it is false, only that it would be possible to demonstrate its falsehood if it were false.

free consent Ethical principle stating those involved in studies must freely agree to do so; thus, researchers are forbidden from placing undue pressure on people to participate in or remain in a study.

full-board review Category of IRB review used for high-risk studies that contain an inflated risk to participants' welfare or the potential for release of confidential information; involves having all members of the IRB review the study procedures and then meet as a group to discuss the degree of risk and protection.

health psychology The application of psychological principles to the study of health, well-being, and illness.

hypothesis A specific and falsifiable statement about the relationship between two or more variables.

induction The process of developing a general hypothesis out of a set of specific observations; best understood as a "bottom-up" approach.

industrial-organizational (I/O) psychology The application of psychological principles to the scientific study of human behavior in the workplace.

informed consent Ethical principle stating that research participants must be informed of all features of the study that would reasonably affect their decision to participate.

Institutional Animal Care and Use Committee (IACUC) Panel that reviews and monitors all research involving nonhuman animals to protect the welfare of research subjects; tasked with ensuring that the benefits of the research outweigh any discomfort experienced by the animals.

institutional review board (IRB) Panel that reviews and monitors all research involving humans to protect the welfare of

research participants; tasked with determining whether a study is consistent with ethical principles and has the authority to approve, reject, or require modification from each research proposal.

operationalization The process of choosing measurable variables to represent the components of a hypothesis.

parsimonious Term applied to theories, meaning that concepts are as simple as possible without sacrificing completeness.

peer review A process that involves having experts in the field evaluate the merits of research articles before they are published.

primary sources Sources that contain full reports of a research study, including information on the participants, the data collected, and the statistical analyses of these data; these appear in professional academic journals.

qualitative research A descriptive approach that attempts to gain a deep understanding of particular cases and contexts.

quantitative research A systematic and empirical approach that attempts to generalize results to other contexts.

reconciliation and synthesis The process of resolving an apparent conflict by finding common ground among the ideas and then merging all the pieces into one new explanation.

scientific method A method of approaching problems and drawing conclusions based on empirical observations; consists of four steps: *hypothesize*, *operationalize*, *measure*, and *explain*, abbreviated HOME.

school psychology The application of psychological principles towards the goal of helping children learn effectively.

scientific misconduct Intentional or negligent distortion of the research process.

secondary sources Sources that contain summaries of primary-source articles; these include textbooks and academic books, as well as less-than-trustworthy websites.

social psychology The study of the ways human thoughts, feelings, and behaviors are shaped by other people.

theory A collection of ideas used to explain the connections between variables and phenomena.

Apply Your Knowledge

1. For each of the following broad theoretical statements, think of a specific research hypothesis that would test the theory. Each statement has many possible hypotheses, but remember that your hypothesis needs to be both testable and falsifiable. The first one is provided as an example.

Theory: Infants look cute and helpless so that adults will take care of them.

Hypothesis: Parents will be more attentive to cute infants than to less cute infants.

Theory: People are inherently social and value the approval of others.

Hypothesis:

Theory: People prefer to feel good about themselves.

Hypothesis:

2. a. Read the following abstract of a published research study (Langer & Rodin, 1976), and identify the four components of the research process:

A field experiment was conducted to assess the effects of enhanced personal responsibility and choice on a group of nursing home residents. Researchers expected that the debilitated condition of many of the aged residing in institutional settings is, at least in part, a result of living in a virtually decision-free environment and consequently is potentially reversible. Residents who were in the experimental group were given a communication emphasizing their responsibility for themselves, whereas the communication given to a second group stressed the staff's responsibility for them. In addition, to bolster the communication, the former group was given the freedom to make choices and the responsibility of caring for a plant rather than having decisions made and the plant taken care of for them by the staff, as was the case for the latter group. Questionnaire ratings and behavioral measures showed a significant improvement for the experimental group over the comparison group on alertness, active participation, and a general sense of well-being.

Hypothesis:

Operationalization (How did researchers define variables?):

Measure (How did researchers conduct the study?):

Explain:

- b. Read the following abstract of a published research study (Swim & Hyers, 1999), and identify the four components of the research process:

Two studies illustrate women's struggle between their desire to challenge sexism and the social pressures and costs that lead to not publicly responding to sexist behavior. In Study 1, 45% of the women confronted a man who made a sexist remark and only 15% did so directly. Confronting was most likely to be chosen by women actively committed to fighting sexism in their daily lives. Private responses illustrated that a lack of responding was not necessarily indicative of complacency about the remarks or a lack of thoughts about confronting. The results from Studies 1 and 2 reveal that diffusion of responsibility, normative pressures not to respond, social pressures to be polite, and concern about retaliation likely suppressed responding.

Hypothesis:

Operationalization (How did researchers define variables?):

Measure (How did researchers conduct the study?):

Explain:

3. Read the following description of a research study, and then evaluate whether it meets the five APA ethical guidelines:

A researcher told students that their responses to an online survey on cheating were anonymous. One question asked students for their e-mail address to use in a raffle drawing. Instead, the researcher used this to locate GPAs in school files so he could correlate frequency of cheating and GPA.

Informed consent?

Free consent?

Protection from harm?

Confidentiality?

Debriefing?

Based on this evaluation, is the study likely to be approved by an Institutional Review Board? Why or why not?

Critical Thinking Questions

1. You have been asked to help determine whether watching violent television leads people to become more violent. Explain how you would approach this task using the four steps of the research process (Hint: HOME).
2. Review the guidelines for evaluating theories. Using these five criteria, evaluate and compare Freud's theory of unconscious drives. The key to this theory is that much of our behavior is driven by internal conflicts that exist outside our awareness.

Research Scenarios: Try It

Suppose you have noticed some differences among your friends in the way they use social networking sites. Surprisingly, some of your most introverted friends are the most active online. After looking into it a little more closely, you observe that whereas your extroverted friends post updates about their personal lives and social plans, your introverted friends tend to post about ideas and causes they are interested in, such as politics or special hobbies. You start to wonder if there might be a research project in this.

1. The first step in the research process is developing a testable hypothesis. Based on your observations, which of the following would be a good scientific hypothesis you could test with a research study?
 - a. Introverts and extroverts post different things on social networking sites.
 - b. Introverts post about ideas rather than people or social events, whereas extroverts post about people rather than ideas.
 - c. Introverts are much more active on social networking sites than extroverts.
 - d. Introverts dislike people.
2. The second step in the research process is to operationalize your variables. Based on your hypothesis, you will need to come up with a way to operationalize the content of social-network posts. In other words, you will need to quantify how much someone's post has to do with "ideas" and how much it has to do with "people." Which of the following would best operationalize these two variables?
 - a. Percentage of posts related to ideas and percentage of posts related to people
 - b. Percentage of posts related to literature and percentage of posts related to parties
 - c. Number of posts about news stories/politics and number of posts concerning social gatherings
 - d. Percentage of posts concerning news stories/politics/hobbies and percentage of posts concerning social gatherings/personal details
3. The third step of this process is to measure—collect your data. In addition to collecting data on the content of posts, as described above, what else will you need to measure to test your hypothesis?
 - a. Introversion/extroversion
 - b. Actual participation in politics
 - c. Total number of friends on social networking site
 - d. Total time spent on social networking site

4. The fourth step in the research process is to explain—to analyze and interpret your results. Suppose you obtained the following results:

	Introverts	Extroverts
percentage of posts on news/politics/hobbies	60%	5%
percentage of posts on social gatherings/personal details	5%	75%

What can you conclude from these results?

- a. Nothing can be concluded.
 - b. The data prove your hypothesis.
 - c. The data support your hypothesis.
 - d. The data fail to support your hypothesis.
5. If you were actually going to conduct this study, you would need to consider the ethical guidelines put forth by the APA and your institution's ethical review board. Which of the following would be the most important ethical issue to consider in conducting this particular study?
- a. protection from harm or extreme stress
 - b. deception
 - c. confidentiality
 - d. free consent

See Appendix A for answers to Research Scenarios: Try It questions.

2 Design, Measurement, and Testing Hypotheses



José Antonio Moreno/age fotostock/Superstock

Learning Outcomes

By the end of this chapter, you should be able to:

- Outline the key features of descriptive, correlational, and experimental research designs.
- Explain the importance of reliability and validity in designing research studies.
- Compare and contrast the different scaling methods for measuring variables.
- Identify the pros and cons of behavioral, physiological, and self-report measures.
- Describe the process of framing and testing hypotheses.

In the early 1950s, Canadian physician Hans Selye introduced the term *stress* into both the medical and popular lexicons. By that time, it had been accepted that humans have a well-evolved fight-or-flight response, which prepares them either to fight back or flee from danger, largely by releasing adrenaline and mobilizing the body's resources more efficiently. While working at McGill University, Selye began to wonder about the health consequences of adrenaline and designed an experiment to test his ideas using rats. Selye injected rats with doses of adrenaline over a period of several days and then euthanized the rats in order to examine the physical effects of the injections. Just as he had hypothesized, rats that were exposed to adrenaline had developed ill effects, such as ulcers, increased arterial plaques, and decreases in the size of reproductive glands—all now understood to be consequences of long-term stress exposure. But there was just one problem. When Selye took a second group of rats and injected them with a placebo, they also developed ulcers, plaques, and shrunken reproductive glands.

Fortunately, Selye was able to solve this scientific mystery with a little self-reflection. Despite all his methodological savvy, he turned out to be rather clumsy when it came to handling rats, occasionally dropping one when he removed it from its cage for an injection. In essence, the experience for both groups of rats was one that we would now call stressful, and it is no surprise that they developed physical ailments in response. Rather than testing the effects of adrenaline injections, Selye was inadvertently testing the effects of being handled by a clumsy scientist. It is important to note that if Selye ran this study in the present day, ethical guidelines would dictate much more stringent oversight of his procedures to protect the welfare of the animals.

This story illustrates two key points about the scientific process. First, as Chapter 1 discussed, researchers should always be attentive to apparent mistakes because they can lead to valuable insights. Second, it is absolutely vital that researchers actually measure what they think they are measuring—Selye ended up measuring the effects of *stress* rather than just adrenaline injections. This chapter explains what it means to do research in a more concrete way, beginning with a broad look at the three types of research design. The goal at this stage is to obtain a general sense of what these designs are, when they are used, and the main differences between them. (Chapters 3, 4, and 5 are each dedicated to one type of research design and will elaborate on each one.) Following the overview of designs, this chapter covers a set of basic principles that are common to all research designs. Regardless of the particulars of a given design, all research studies involve making sure measurements are accurate and consistent and that they are captured using the appropriate type of scale. Finally, the chapter will discuss the general process of hypothesis testing, from laying out predictions to drawing conclusions.

2.1 Overview of Research Designs

As Chapter 1 explained, scientists can have a wide range of goals when they begin a research project, everything from describing a phenomenon to changing people's behavior. It turns out that these goals will dictate different approaches to answering a research question. That is, researchers will approach the problem of describing voting patterns differently than they would approach the problem of how to increase voter turnout. These approaches are called

research designs, or the specific methods that are used to collect, analyze, and interpret data. The choice of a design is not one to be made lightly; the way an investigator collects data trickles down to decisions about how to analyze the data and about the kinds of conclusions that can be drawn from the results. This section provides a brief introduction to the three main types of design—descriptive, correlational, and experimental.

Descriptive Research

Recall from Chapter 1 that a research study can have the basic goal of describing a phenomenon. If a research question centers around description, then the research design falls under the category of **descriptive research**, in which the primary goal is to describe thoughts, feelings, or behaviors. Descriptive research provides a static picture of what people are thinking, feeling, and doing at a given moment in time, as the following examples of research questions illustrate:

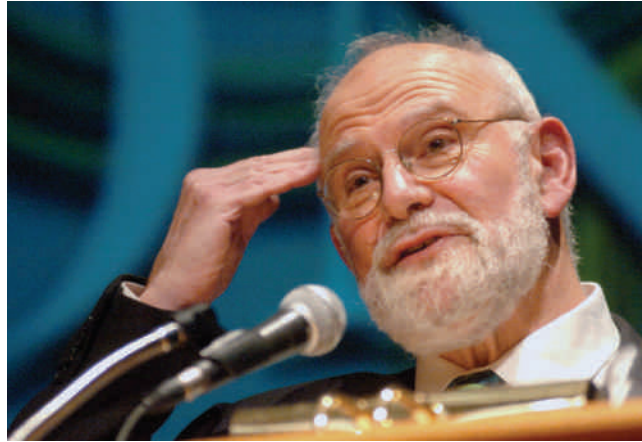
- What percentage of doctors prefer Xanax for the treatment of anxiety? (thoughts)
- What percentage of registered Republicans vote for independent candidates? (behaviors)
- What percentage of Americans blame the president for the economic crisis? (thoughts)
- What percentage of college students experience clinical depression? (feelings)
- What is the difference in crime rates between Beverly Hills and Detroit? (behaviors)

What these five questions have in common is an attempt to get a broad understanding of a phenomenon without trying to delve into its causes.

The crime-rate example highlights the main advantages and disadvantages of descriptive designs. On the plus side, descriptive research is a good way to achieve a broad overview of a phenomenon and may inspire future research. It is also a good way to study things that are difficult to translate into a controlled experimental setting. For example, crime rates can affect every aspect of people's lives, and this importance would likely be lost in an experiment that staged a mock crime in a laboratory. On the downside, descriptive research provides a static overview of a phenomenon and cannot explore the reasons for it. A descriptive design might tell us that Beverly Hills residents are half as likely as Detroit residents to be assault victims, but it would not reveal the underlying *reasons* for this discrepancy. (If we wanted to understand why this was true, we would use one of the other designs.)

Descriptive research can be either qualitative or quantitative; in fact, the large majority of qualitative research falls under the category of descriptive designs. Descriptions are *quantitative* when they attempt to make comparisons or to present a random sampling of people's opinions. The majority of our example questions above would fall into this group because they quantify opinions from samples of households, or cities, or college students. Good examples of quantitative description appear in the "snapshot" feature on the front page of *USA Today*. The graphics represent poll results from various sources; the snapshot for May 15, 2015, reported that 90% of Americans crave more "variety" in their home-cooked meals (i.e., thoughts). View a current gallery of these snapshot graphs here: <http://www.usatoday.com/services/snapshots/gallery/>

Descriptive designs are qualitative when they attempt to provide a rich description of a particular set of circumstances. A powerful example of this approach can be found in the work of the late neurologist Oliver Sacks. Sacks wrote several books exploring the ways that people with neurological damage or deficits are able to navigate the world around them. In one selection from *The Man Who Mistook His Wife for a Hat*, Sacks (1998) relates the story of a man he calls William Thompson. As a result of chronic alcohol abuse, Thompson developed Korsakov's syndrome, a brain disease marked by profound memory loss. The memory loss was so severe that Thompson had effectively "erased" himself and could remember only scattered fragments of his past.



Johnathon Henninger/Connecticut Post/AP Images

Dr. Oliver Sacks studied how people with neurological damage formed and retained memories.

Whenever Thompson encountered people, he would frantically try to determine who he was. He would develop hypotheses and test them, as in this excerpt from one of Sacks's visits:

I am a grocer, and you're my customer, right? Well, will that be paper or plastic? No, wait, why are you wearing that white coat? You must be Hymie, the kosher butcher. Yep. That's it. But why are there no bloodstains on your coat? (p. 112)

Sacks concluded that Thompson was "continually creating a world and self, to replace what was continually being forgotten and lost" (p. 113). With this story, Sacks helps illuminate Thompson's experience and fosters readers' gratitude for the ability to form and retain memories. This story also illustrates the trade-off in these sorts of descriptive case studies: Despite all its richness, we cannot generalize these details to other cases of brain damage; we would need to study and describe each patient individually.

Correlational Research

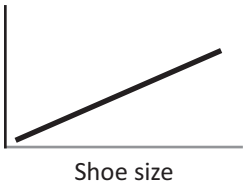
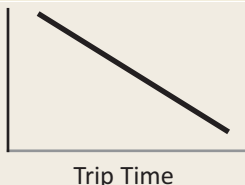
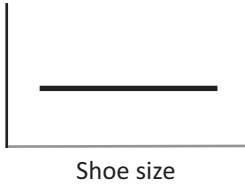
Recall from Chapter 1 that research studies can also have the goal of trying to *predict* a phenomenon. If a research question centers around prediction, then the research design falls under the category of **correlational research**, in which the primary goal is to understand the *relationships* among various thoughts, feelings, and behaviors. Examples of correlational research questions include:

- Are people more aggressive on hot days?
- Are people more likely to smoke when they are drinking?
- Is income level associated with happiness?
- What is the best predictor of success in college?
- Does television viewing relate to hours of exercise?

What these questions have in common is the goal of *predicting* one variable based on another. If we know the temperature, can we predict aggression? If we know a person's income, can we predict her level of happiness? If we know a student's SAT scores, can we predict his college GPA?

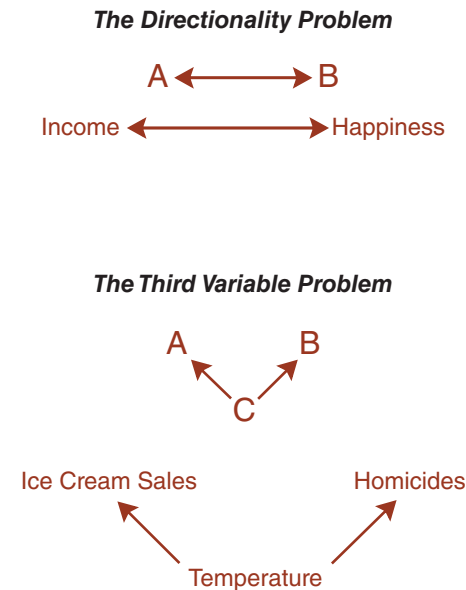
These predictive relationships can turn out in one of three ways (Chapter 4 will provide more detail about each): A **positive correlation** means that higher values of one variable predict higher values of the other variable. For instance, more money is associated with higher levels of happiness, and less money is associated with lower levels of happiness. The key is that these variables move up and down together, as the first row of Table 2.1 shows. A **negative correlation** means that higher values of one variable predict lower values of the other variable. For example, more television viewing is associated with fewer hours of exercise, and fewer hours of television is associated with more hours of exercise. The key is that one variable increases while the other decreases, as the second row of Table 2.1 illustrates. Finally, worth noting is a third possibility, which is *no correlation* between two variables, meaning that we cannot predict one variable based on another. In brief, changes in one variable are not associated with changes in the other, as seen in the third row of Table 2.1.

Table 2.1: Three possibilities for correlational research

Outcome	Description	Visual
Positive Correlation	Variables go up and down together. For example: Taller people have bigger feet, and shorter people have smaller feet.	Height  Shoe size
Negative Correlation	One variable goes up, and the other goes down. For example: As a driver's speed goes up, the time it takes to finish the trip decreases.	Speed  Trip Time
No Correlation	The variables have nothing to do with one another. For example: Shoe size and number of siblings are completely unrelated.	# Siblings  Shoe size

Correlational designs are about testing predictions, but we are still unable to make causal, explanatory statements (that comes next). A common mantra in the field of psychology is that correlation does not equal causation. In other words, just because variable A predicts variable B does not mean that A causes B. This is true for two reasons, which we refer to as the **directionality problem** and the **third variable problem**. (See Figure 2.1.)

Figure 2.1: Correlation is not causation



First, when we measure two variables at the same time, we have no way of knowing the direction of the relationship. Take the relationship between money and happiness: It could be true that money makes people happier, because they can afford nice things and fancy vacations. It could also be true that happy people have the confidence and charm to obtain higher-paying jobs, resulting in more money. In a correlational study, we are unable to distinguish between these possibilities. Or, take the relationship between television viewing and obesity: It could be that people who watch more television get heavier, because TV makes them snack more and exercise less. It could also be that people who are overweight lack the energy to move around and end up watching more television as a consequence. Once again, we cannot identify a cause–effect relationship in a correlational study.

Second, when we measure two variables as they naturally occur, a third variable that actually causes both of them is always a possibility. For example,

imagine we find a correlation between the number of churches and the number of liquor stores in a city. Do people build more churches to offset the threat of liquor stores? Do people build more liquor stores to rebel against churches? Most likely, the link involves a third variable, population size, that causes changes in both variables: The more people who are living in a city, the more churches and liquor stores they can support. As another example, imagine a correlation between ice cream sales and homicide rates is discovered. Does ice cream lead people to commit murder? Do murderers like to buy ice cream on the way home from the scene of the crime? Most likely, the link involves a third variable, temperature, that causes changes in both variables: The hotter it gets outside, the more people want ice cream, and the greater likelihood that disagreements will turn violent.

Experimental Research

Finally, recall that research projects can have the goal of attempting to explain a phenomenon. When the research goal involves causal explanations, then research design falls under the category of **experimental research**, in which the primary goal is to explain thoughts, feelings, and behaviors and to make causal statements. Examples of experimental research questions include:

- Does smoking cause cancer?
- Does drinking alcohol make people more aggressive?
- Does loneliness cause alcoholism?
- Does stress cause heart disease?
- Can meditation make people healthier?

Research: Making an Impact

Helping Behaviors

The 1964 murder of Kitty Genovese in plain sight of her neighbors, none of whom helped, drove numerous researchers to investigate why people may not help others in need. Are individuals selfish and bad, or does a group dynamic lead to inaction? Is there something wrong with our culture, or are situations more powerful than we think?

Among the body of research conducted in the late 1960s and 1970s was one pivotal study that revealed why people may not help others in emergencies. Darley and Latané (1968) conducted an experiment with various individuals in different rooms who communicated with each other via intercom. In reality, the study included just one participant and a number of confederates, one of whom pretended to have a seizure. Among participants who thought they were the only other person listening over the intercom, more than 80% helped, and they did so in less than 1 minute. However, among participants who thought they were one of a *group* of people listening over the intercom, less than 40% helped, and even then only after more than 2.5 minutes. This phenomenon—that the more people who witness an emergency are present, the less likely any of them is to help—has been dubbed the “bystander effect.” One of the main reasons that this tendency occurs is that responsibility for helping gets “diffused” among all of the people present, so that each one feels less personal responsibility for taking action.

Darley and Latané’s research can be seen in action and has influenced safety measures in today’s society. For example, when someone witnesses an emergency, no longer does it suffice to simply yell to the group, “Call 911!” Because of the bystander effect, we know that most people will believe someone else will do it, and the call will not be made. Instead, it is necessary to designate a specific person to make the call. In fact, part of modern-day CPR training involves making individuals aware of the bystander effect and best practices for getting people to help and be accountable.

Although the bystander effect may be the rule, there are always exceptions. For example, on September 11, 2001, the fourth hijacked airplane was overtaken by a courageous group of passengers. Most people on the plane had heard about the twin tower crashes and recognized that their plane was heading for Washington, D.C. Despite being amongst nearly 100 other people, a few people chose to help the intended targets in D.C. Risking their own safety, this heroic group chose to help to prevent others from experiencing death and suffering. So, although we may see events that remind us of the reality of the bystander effect, we also see moments where people are willing to help, no matter the number of people that surround them.

Think About It:

1. What type of research design best describes Darley & Latane’s (1968) study?
2. What practical applications have resulted from research on people’s reluctance to help in emergencies?

What these five questions have in common is a focus on understanding *why* something happens. Experiments move beyond, for example, the question of whether alcoholics are more aggressive to whether alcohol actually causes an increase in aggression.

Experimental designs are able to address the shortcomings of correlational designs because the researcher has more control over the environment. Chapter 5 will cover this in great detail, but the basic process of conducting an experiment is relatively simple: A researcher has to control the environment as much as possible so that all participants in the study have the same experience. This helps eliminate other third variables that might influence the results. Researchers will then manipulate, or change, one key variable and then measure outcomes in another key variable. The variable manipulated by the experimenter is called the **independent variable (IV)**. The outcome variable that is measured by the experimenter is called the **dependent variable (DV)**. The combination of controlling the setting and changing one aspect of this setting at a time allows the experimenter to state with some certainty that the changes caused something to happen.

Think of this in a little more concrete way. Imagine that a researcher wanted to test the hypothesis that meditation improves health. In this case, meditation would be the independent variable, and health would be the dependent variable. One way to test this hypothesis would be to take a group of people and have half of them meditate 20 minutes per day for several days while the other half did something else for the same amount of time. The group that meditates would be called the **experimental group** because it provides the test of the hypothesis. The group that does not meditate would be called the **control group** because it provides a basis of comparison for the experimental group.

The researcher would want to make sure that these groups spent the 20 minutes in similar conditions so that the only difference would be the presence or absence of meditation. One way to accomplish this would be to have all participants sit quietly for the 20 minutes but give the experimental group specific instructions on how to meditate. Then, to test whether meditation led to increased health and happiness, the researcher would give both groups a set of outcome measures at the end of the study—perhaps a combination of survey measures and a doctor's examination. If differences were found between the dependent measures for the two groups, the experimenter could be fairly confident that meditation caused them to happen. One way we can operationalize health outcomes in this study would be to measure blood pressure, as higher levels of blood pressure put people at risk for developing cardiovascular disease. So, for example, the researcher might find lower blood pressure in the experimental (meditation) group, which would suggest that meditation causes blood pressure to drop.

Choosing a Research Design

The choice of a research design is guided first and foremost by a researcher's finding the best fit to the research question and then adjusting it depending on practical and ethical concerns. At this point, a nagging question may come to mind: If experiments are the most powerful type of design, why not use them every time? Why would anyone give up the chance to make causal statements? One reason is that we are often interested in variables that cannot be manipulated, for ethical or practical reasons, and that therefore have to be studied as they occur naturally. In one example, Matthias Mehl and Jamie Pennebaker (2003) happened to start a weeklong study of college students' social lives on September 10, 2001. Following the

terrorist attacks on the morning of September 11, Mehl and Pennebaker were able to track changes in people's social connections and use this to understand how groups respond to traumatic events. Of course, it would have been unthinkable to manipulate a terrorist attack for this study experimentally, but since it occurred naturally, the researchers were able to conduct a correlational study of coping.

Another reason to use descriptive and correlational designs is that these are useful in the early stages of a research program. For example, before a psychologist can start to think about the causes of binge drinking among college students, it is important to understand how common is this phenomenon. Likewise, before a researcher designs a time- and cost-intensive experiment on the effects of meditation, it is a good idea to conduct a correlational study to test whether meditation even predicts health. In fact, this latter example comes from a series of real research studies conducted by psychiatrist Sara Lazar and her colleagues at Massachusetts General Hospital. This research team first discovered that experienced practitioners of mindfulness meditation had more development in brain areas associated with control over attention and emotion. But this study was correlational at best; perhaps meditation caused changes in brain structure or perhaps people who were better at integrating emotions were drawn to meditation. In a follow-up study, researchers randomly assigned people either to meditate or to perform stretching exercises for two months. These experimental findings confirmed that mindfulness meditation actually caused structural changes to the brain (Hölzel et al., 2011). This series of studies is a prime example of how a research program can progress from correlational to experimental designs.

Table 2.2 summarizes the main advantages and disadvantages of these three types of design. In addition, the bottom of the table includes two examples of research topics—meditation and health, and temperature and aggression—to showcase the similarities and differences between the designs.

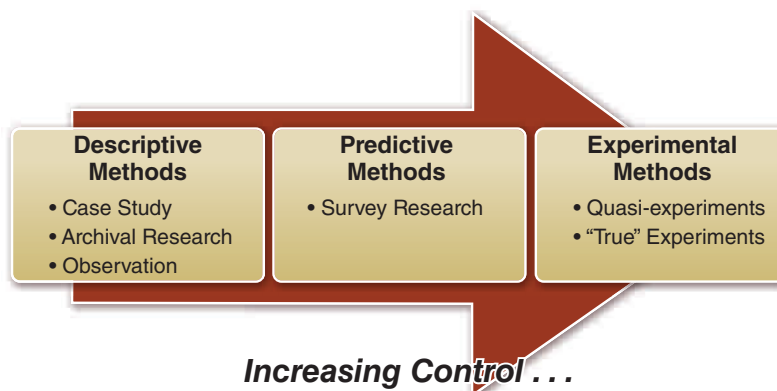
Table 2.2: Summary of research designs

Research Design	Descriptive	Correlational	Experimental
Goal	<i>Describe</i> characteristics of an existing phenomenon	<i>Predict</i> behavior; assess strength of relationship between variables	<i>Explain</i> behavior; assess impact of IV on DV
Advantages	Provides a complete picture of what is occurring at a given time	Allows testing of expected relationships; predictions can be made	Allows conclusions to be drawn about causal relationships
Disadvantages	Does not assess relationships; no explanation for phenomenon	Cannot draw inferences about causal relationships	Cannot manipulate many important variables
Example #1: Studying Meditation	What percentage of college students meditate at least once a week?	Are regular meditators happier and healthier?	If we randomly assign people to start meditating, do they become happier and healthier?
Example #2: Temperature and Aggression	How many violent crimes are committed in the summer?	Are crime rates higher in the summer than in the winter?	If we turn up the temperature in the laboratory, do people become more aggressive?

Designs on the Continuum of Control

Before leaving the design overview behind, we will consider how these designs relate to one another. The best way to think about the differences between the designs is in terms of the amount of **control** a researcher has. That is, experimental designs are the most powerful because the researcher controls everything from the hypothesis to the environment in which the data are collected. Correlational designs are less powerful because the researcher is restricted to measuring variables as they occur naturally. However, with correlational designs, the researcher does maintain control over several aspects of data collection, including the setting and the choice of measures. Descriptive designs are the least powerful because researchers have difficulty controlling outside influences on data collection. For example, when people answer opinion polls over the phone, they might be sitting quietly and pondering the questions or they might be watching television, eating dinner, and dealing with a fussy toddler. As a result, researchers are more limited as to the conclusions they can draw from these data. Figure 2.2 shows an overview of where research designs fall on the **continuum of control** in order of increasing control: from descriptive, to predictive, to experimental. Chapters 3, 4, and 5 will cover variations on these designs in more detail.

Figure 2.2: The continuum of control framework



2.2 Reliability and Validity

Each of the three types of research designs—descriptive, correlational, and experimental—has the same basic goal: to take a hypothesis about some phenomenon and translate it into measurable and testable terms. That is, whether researchers use a descriptive, correlational, or experimental design to test predictions about income and happiness, they still need to translate (or *operationalize*) the concepts of income and happiness into measures that will be meaningful for the study. Unfortunately, the sad truth is that research measurements will always be influenced by factors in addition to the conceptual variable of interest. Answers to any set of questions about happiness will depend both on actual levels of happiness and the ways people interpret the questions. The meditation experiment may have different effects, depending on people’s experience with meditation. Even describing the percentage of Republicans voting for independent candidates will vary according to characteristics of a particular candidate.

These additional sources of influence can be grouped into two categories: random and systematic errors. **Random error** involves chance fluctuations in measurements, such as when a participant misunderstands the question, or shows up in a terrible mood after walking through a blizzard to get to the study. Although random errors can influence measurement, they generally cancel out over the span of an entire sample. That is, some people may overreact to a question while others underreact. Heavy snowfall might put one person in a terrible mood and make another appreciate the joy of winter. While both of these examples would add error to a dataset, they would cancel each other out in a sufficiently large sample.

Systematic errors, in contrast, are those that systematically increase or decrease along with values of the measured variable. For example, people who have more experience with meditation may consistently show more improvement in a meditation experiment than those with less experience. Or, people who have higher self-esteem may score higher on a measure of happiness than those with lower self-esteem. In this case, the happiness scale does not do a good job of homing in on the concept of “happiness” and will end up instead assessing a combination of happiness and self-esteem. These types of errors can cause more serious trouble for a researcher’s hypothesis tests because they interfere with the attempts to understand the link between two variables.

In sum, the measured values of a variable reflect a combination of the true score, random error, and systematic error, as the following conceptual equation shows:

$$\text{Measured Score} = \text{True Score} + (\text{Random Error} + \text{Systematic Error})$$

For example:

$$\text{Happiness Score} = \text{Actual Happiness} + (\text{Misreading the Question} + \text{Self-Esteem})$$

So, if our measurements are also affected by outside influences, how do we know whether our measures are meaningful? Occasionally, the answer to this question is straightforward; if we ask people to report their weight or their income level, these values can be verified using objective sources. Many research questions within psychology, however, involve more ambiguity. How do we know that our happiness scale is accurate? The problem is that we have no way to objectively verify happiness beyond people’s self-reports of their own happiness. What researchers need, then, are ways to assess *how close* they are to measuring happiness in a meaningful way. This assessment involves two related concepts: **reliability**, or the consistency of a measure; and **validity**, or the accuracy of a measure. This section examines both of these concepts in detail.

Reliability

The consistency of time measurement by watches, cell phones, and clocks reflects a high degree of reliability. People think of a watch as reliable when it keeps track of the time consistently—an hour should take the same amount of time to pass, 24 times per day. Likewise, the scale is reliable when it gives the same value for weight in back-to-back measurements—an individual’s weight should be the same if he steps off the scale and right back on, provided he stays away from the fridge in the meantime.

Reliability is defined as the extent to which a measured variable is free from random errors, and it is best understood as the degree of *consistency* in research measurements. As the chapter discussed previously, researchers' measures are never perfect, and five main sources of random error threaten reliability:

1. *Transient states*, or temporary fluctuations in participants' cognitive or mental state; for example, some participants may complete a study after an exhausting midterm or after a fight with their significant others.
2. *Stable individual differences* among participants; for example, some participants are habitually more motivated or happier than other participants.
3. *Situational factors* in the administration of the study; for example, an experiment conducted in the early morning may make everyone tired or grumpy.
4. *Bad measures* that add ambiguity or confusion to the measurement; for example, participants may respond differently to a question about "the kinds of drugs you are taking." Some may take this to mean illegal drugs, whereas others interpret it as prescription or over-the-counter drugs.
5. *Mistakes in coding responses* during data entry; for example, a handwritten "7" could be mistaken for a "4." (Happily, these types of errors have been minimized by the increasing role of computers in data collection. If someone clicks the number "7" in an online survey, the computer will record it as a "7" almost every time.)

Researchers naturally want to minimize the influence of all of these sources of error, and the text will touch on techniques for doing so throughout. However, researchers are also resigned to the fact that all measurements contain a degree of error. The goal, then, is to develop an estimate of how reliable measures are. Researchers generally estimate reliability in three ways.

1. **Test-retest reliability** refers to the consistency of the measure over time—much like the examples of a reliable watch and a reliable scale. A fair number of research questions in the social and behavioral sciences involve measuring stable qualities. For example, if someone were to design a measure of intelligence or personality, both of these characteristics should be relatively stable over time. An individual score on an intelligence test today should be roughly the same as the score when tested again in five years. A person's level of extroversion today should correlate highly with his or her level of extroversion in 20 years. The test-retest reliability of these measures is quantified by simply correlating measures at two time points. The higher these correlations are, the higher the reliability will be. This makes conceptual sense as well; if measured scores reflect the true score more than they reflect random error, then this will result in increased stability of the measurements.
2. **Inter-item reliability** refers to the internal consistency among different items on a measure. Think back to the last time you completed a survey. Did it seem to ask the same questions more than once? (Chapter 4 [4.1] will discuss this technique.) The repetition is included because a single item is more likely to contain measurement error than the average of several items will—remember that small random errors

tend to cancel out each other. Consider the following items from Sheldon Cohen's *Perceived Stress Scale* (Cohen, Kamarck, & Mermelstein, 1983):

- In the last month, how often have you felt that you were unable to control the important things in your life?
- In the last month, how often have you felt confident about your ability to handle your personal problems?
- In the last month, how often have you felt that things were going your way?
- In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?

Each of these items taps into the concept of feeling “stressed out,” or overwhelmed by the demands of life. One standard way to evaluate a measure like this is by computing the average correlation between each pair of items, a statistic referred to as *Cronbach's alpha*. The more these items tap into a central, consistent construct, the higher the value of this statistic is. Conceptually, a higher alpha means that variation in responses to the different items reflects variation in the “true” variable being assessed by the scale items. Alpha levels range from zero to one, with higher numbers indicating more internal consistency. As a general rule, researchers want this index to be above 0.70 to have any confidence in the measure.

3. **Interrater reliability** refers to the consistency among judges observing participants' behavior. The previous two forms of reliability were relevant in dealing with self-report scales; interrater reliability is more applicable when the research involves **behavioral measures**, which involve direct and systematic recording of observable behaviors. Imagine a researcher is studying whether alcohol consumption makes people behave more aggressively. One way to tackle this hypothesis would be to have a group of judges observe participants after drinking and rate their levels of aggression. In the same way that using multiple scale items helps to cancel out the small errors of individual items, using multiple judges cancels out the variations in each individual's ratings. In this case, people could have slightly different ideas and thresholds for what constitutes aggression. To determine how much these differences matter, the researcher can evaluate the judges' ratings by calculating the average correlation among the ratings. The higher the alpha values, the more the judges agree in their ratings of aggressive behavior. Conceptually, a higher alpha value means that variation in the judges' ratings reflects real variation in levels of aggression.

Validity

Recall the watch and scale examples. Perhaps some people set their watch 10 minutes ahead to avoid being late. Or perhaps certain individuals adjust their scale by 5 pounds to boost either their motivation or self-esteem. In these cases, the watch and the scale may produce consistent measurements, but the measurements are not accurate. It turns out that the reliability of a measure is a necessary but not sufficient basis for evaluating it. Put bluntly, measures can be (and have to be) consistent, but they might still be worthless. The additional piece of the puzzle is the validity of measures, or the extent to which they accurately measure what they are designed to measure.

Whereas reliability is threatened more by random error, validity is threatened more by systematic error. If the measured scores on the happiness scale reflect, say, self-esteem more than they reflect happiness, this would threaten the validity of the scale. The previous section explained that a test designed to measure intelligence ought to be consistent over time. And, in fact, these tests do show very high degrees of reliability. However, several researchers have cast serious doubts on the validity of intelligence testing, arguing that even scores on an official IQ test are influenced by a person's cultural background, socioeconomic status (SES), and experience with the process of test-taking (for discussion of these critiques, see Daniels et al., 1997; Gould, 1996). For example, children growing up in higher SES households tend to have more books in the home, spend more time interacting with one or both parents, and attend schools that have more time and resources available—all of which correlate with scores on IQ tests. Thus, because all of these factors could increase scores on an intelligence test, they amount to systematic error in the measure of intelligence and, therefore, threaten the validity of a measured score on an intelligence test.

Researchers have two primary ways to discuss and evaluate the validity, or accuracy, of measures: construct validity and criterion validity.

Researchers evaluate **construct validity** based on how well the measures capture the underlying conceptual ideas (i.e., the constructs) in a study. These constructs are equivalent to the “true score” discussed in the previous section. That is, how accurately does the bathroom scale measure the construct of weight? How accurately does an IQ test measure the construct of intelligence relative to other things? The validity of measures can be assessed in a couple of ways. On the subjective end of the continuum, researchers can evaluate construct validity by assessing the **face validity** of the measure, or the extent to which it simply seems like a good measure of the construct. The items from the Perceived Stress Scale have high face validity because the items match what we intuitively mean by “stress” (e.g., “How often have you felt difficulties were piling up so high that you could not overcome them?”). However, if we were to measure an individual's speed at eating hot dogs and then state it was a stress measure, the participant might be skeptical because hot-dog eating speed would lack face validity as a measure of stress.

Although face validity is nice to have, it can sometimes (ironically) reduce the validity of the measures. Imagine seeing the following two measures on a survey of attitudes:

1. Do you dislike people whose skin color is different from yours?
2. Do you ever beat your children?

On the one hand, these are extremely face-valid measures of attitudes about prejudice and corporal punishment—the questions very much capture our intuitive ideas about these concepts. On the other hand, even people who do support these attitudes may be unlikely to answer honestly because they recognize that neither attitude is popular. In cases like this, a measure low in face validity might end up being the more accurate approach. Chapter 4 will discuss ways to strike this balance.

On the less subjective end, researchers can evaluate construct validity by examining measures' empirical connections to both related and unrelated constructs. Imagine for a moment that

we are developing a new measure of liberal political attitudes. If we think about a person who describes herself as liberal, she is likely to support gun control, equal rights, and a woman's right to choose. And, she is less likely to be pro-war, anti-immigration, or anti-gay rights. Therefore, we would expect our new liberalism measure to correlate positively with existing measures of attitudes toward guns, affirmative action, and abortion. This pattern of correlations taps into the metric of **convergent validity**, or the extent to which our measure overlaps with conceptually similar measures. But, we would want to ensure that the new measure captures something distinct from other constructs. In this case, we might want to demonstrate that we have developed a true measure of *political* attitudes, which does not simply correlate with *religious* beliefs. That is, we would want to show that liberal political views could be independent of religion. This hypothesized lack of correlations taps into the metric of **discriminant validity**, or the extent to which a measure diverges from unrelated measures.

To take another example, imagine someone wanted to develop a new measure of narcissism, usually defined as an intense desire to be liked and admired by other people. Narcissists tend to be self-absorbed but also very attuned to the feedback they receive from other people—especially feedback about the extent to which people admire them. Narcissism somewhat resembles self-esteem but differs enough; perhaps it is best viewed as high and unstable self-esteem. So, given these facts, a researcher might assess the discriminant validity of the measure by making sure it does not overlap too closely with measures of self-esteem or self-confidence. This approach would establish that the narcissism measure stands apart from these different constructs. The researcher might then assess the convergent validity of the measure by making sure that it does correlate with things like sensitivity to rejection and need for approval. These correlations would place the measure into a broader theoretical context and help to establish it as a valid measure of the construct of narcissism.

Criterion validity involves evaluating the validity of measures based on the association between measures and relevant *behavioral* outcomes. The “criterion” in this case refers to a measure that can be used to make decisions. For example, if someone developed a personality test to assess an individual's management style, the most relevant metric of its validity is whether it predicts a person's actual behavior as a manager. That is, we might expect people scoring high on this scale to be able to increase the productivity of their employees and to maintain a comfortable work environment. Likewise, if someone developed a measure that predicted the best careers for graduating seniors based on their skills and personalities, then criterion validity would be assessed using people's actual success in these various careers. Whereas construct validity is more concerned with the underlying theory behind the constructs, criterion validity is more concerned with the practical application of measures. As might be expected, researchers are more likely to use this approach in applied settings.



Jupiterimages/Stockbyte/Thinkstock

That said, criterion validity is also a useful way to supplement validation of a new questionnaire. For example, a questionnaire about generosity should

Criterion validity can be used to predict a future behavioral outcome like management success.

be able to predict people's annual giving to charities, and a questionnaire about hostility ought to predict hostile behaviors. To supplement the construct validity of the narcissism measure, a researcher might examine its ability to predict the ways people respond to rejection and approval. Based on the definition of the construct, the researcher might hypothesize that narcissists would become hostile following rejection and perhaps become eager to please following approval. If these predictions were supported, it would mean further validation that the measure was capturing the construct of narcissism.

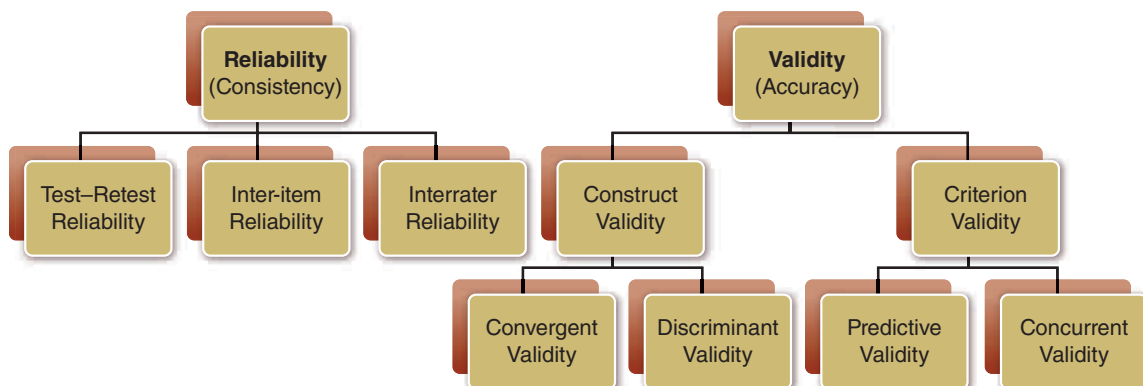
Criterion validity falls into one of two categories, depending on whether the researcher is interested in the present or the future. **Predictive validity** involves attempting to predict a future behavioral outcome based on the measure, as in the examples of the management-style and career-placement measures. Predictive validity is also at work when researchers (and colleges) try to predict graduates' likelihood of school success based on SAT or GRE scores. The goal here is to validate the construct via its ability to predict the future.

In contrast, **concurrent validity** involves attempting to link a self-report measure with a behavioral measure collected at the same time, as in the examples of the generosity and hostility questionnaires. The phrase "at the same time" is used vaguely here; these self-report and behavioral measures may be separated by a short time span. In fact, concurrent validity sometimes involves trying to predict behaviors that occurred before completion of the scale, such as trying to predict students' past drinking behaviors from an "attitudes toward alcohol" scale. The goal in this case is to validate the construct via its association with similar measures.

Comparing Reliability and Validity

This section has discussed how both reliability (consistency) and validity (accuracy) are ways to evaluate measured variables and to assess how well these measurements capture the underlying conceptual variable. In establishing estimates of both of these metrics, researchers essentially examine a set of correlations with their measured variables. But while reliability involves correlating variables with themselves (e.g., happiness scores at week 1 and week 4), validity involves correlating variables with other variables (e.g., happiness scale with the number of times a person smiles). Figure 2.3 displays the relationships among types of reliability and validity.

Figure 2.3: Types of reliability and validity



We learned earlier that reliability is necessary but not sufficient to evaluate measured variables. That is, reliability has to come first and is an essential requirement for any variable—no one would trust a watch that was sometimes five minutes fast and other times ten minutes slow. If we cannot establish that a measure is reliable, then there is really no chance of establishing its construct validity because every measurement might be a reflection of random error. However, just because a measure is consistent does not make it accurate. Someone's watch might consistently be ten minutes fast; a scale might always be five pounds under the person's actual weight. For that matter, a test of intelligence might result in consistent scores but actually be capturing respondents' cultural background. Reliability tells us the extent to which a measure is free from random error. Validity takes the second step of telling us the extent to which the measure is also free from systematic error.

Finally, it is worth pointing out that establishing validity for a new measure is hard work. Reliability can be tested in a single step by correlating scores from multiple measures, multiple items, or multiple judges within a study. But testing the construct validity of a new measure involves demonstrating both convergent and discriminant validity. In developing our narcissism scale, we would need to show that it correlated with things like fear of rejection (convergent) but was reasonably different from things like self-esteem (discriminant). The latter criterion is particularly difficult to establish because it takes time and effort—and multiple studies—to demonstrate that one scale is distinct from another. However, an easy way exists to avoid these challenges: Use existing measures whenever possible. Before creating a brand-new happiness scale, or narcissism scale, or self-esteem scale, check the research literature to see if one exists that has already gone through the ordeal of being validated.

2.3 Scales and Types of Measurement

One of the easiest ways to decrease error variance and thereby increase reliability and validity is to make smart choices when designing and selecting measures. Throughout this book, we will discuss guidelines for each type of research design and ways to ensure that measures are as accurate and unbiased as possible. This section examines some basic rules that apply across all three types of design. We first review the four scales of measurement and discuss the proper use of each one; we then turn our attention to three types of measurement used in psychological research studies.

Scales of Measurement

Whenever researchers perform the process of translating conceptual variables into measurable variables (i.e., operationalization; see Chapter 1, section 1.2), they must ensure that their measurements accurately represent the underlying concepts. In Chapter 1, the discussion of validity explained that this accuracy is a critical piece of hypothesis testing. For example, if researchers develop a scale to measure job satisfaction, then they need to verify that this is actually what the scale is measuring.

However, measurement accuracy has an additional, subtler dimension: We also need to be sure that the numbers used in our chosen measurement accurately reflect the underlying mathematical properties of the concept. In many cases in the natural sciences, this process

is automatically precise. When we measure the speed of a falling object or the temperature of a boiling object, the underlying concepts (speed and temperature) translate directly into scaled measurements. In the social and behavioral sciences, though, this process is trickier; researchers have to decide carefully how best to represent abstract concepts such as happiness, aggression, and political attitudes. As researchers take the step of **scaling** variables, or specifying the relationship between a conceptual variable and numbers on a quantitative measure, they have four different scales to choose from, presented below in order of increasing statistical power and flexibility.

Nominal Scales

Nominal scales are used to label or identify a particular group or characteristic. For example, we can label a person's gender as male or female, and we can label a person's religion as Catholic, Protestant, Buddhist, Jewish, Muslim, Hindu, etc. In experimental designs, researchers can also use nominal scales to label the condition to which a person has been assigned (e.g., experimental or control groups). The assumption in using these labels is that members of the group have some common value or characteristic, as defined by the label. For example, everyone in the Catholic group should have similar religious beliefs, and everyone in the female group should be of the same gender.

Research studies commonly represent these labels using numeric codes in a data file, such as 1 to indicate females and 2 to indicate males. However, these numbers are completely arbitrary and meaningless—that is, males do not have more gender than females. We could just as easily replace the 1 and the 2 with another pair of numbers or with a pair of letters or names. Thus, the primary limitation of nominal scales is that the scaling itself is arbitrary, which prevents us from using these values in mathematical calculations. One helpful way to appreciate the difference between this scale and the next three is to think of nominal scales as qualitative, because they label and identify, and to think of the other scales as quantitative, because they indicate the extent to which someone possesses a quality or characteristic. The next sections explore these quantitative scales in more detail.

Ordinal Scales

Researchers use **ordinal scales** to represent ranked orders of conceptual variables, such that higher numbers reflect increasing magnitude of the underlying variable. For example, beauty contestants, horses, and Olympic athletes are all ranked by the order in which they finish—first, second, third, and so on. Likewise, movies, restaurants, and consumer goods are often rated using a system of stars (i.e., 1 star is poor; 5 stars is excellent) to represent their quality. In these examples, we can draw conclusions about the relative speed, beauty, or deliciousness of the rating target. Even so, the numbers used to label these rankings do not necessarily map directly to differences in the conceptual variable. The fourth-place finisher in a race is rarely twice as slow as the second-place finisher; the beauty-contest winner is not three times as attractive as the third-place finisher; and the boost in quality between a four-star and a five-star restaurant is not the same as the boost between a two-star and three-star restaurant. Ordinal scales represent rank orders, but the numbers do not have any absolute value of their own. This type of scale, then, is more powerful than a nominal scale but still limited in that it does not allow performance of mathematical operations. For example, if an Olympic athlete finished first in the 800-meter dash, third in the 400-meter hurdles, and second in the

400-meter relay, we might be tempted to calculate her average finish as second place. Unfortunately, the properties of ordinal scales prevent us from doing this sort of calculation, because the conceptual distance between first, second, and third place would be different in each case. (That is, the runner might have won the 800-meter dash by 5 seconds, but the 400-meter relay by less than a second.) To perform any mathematical manipulation of variables requires one of the next two types of scale.

Interval Scales

Interval scales represent cases where the numbers on a measured variable correspond to equal distances on a conceptual variable. For example, temperature increases on the Fahrenheit scale represent equal intervals—warming from 40 to 47 degrees is the same increase as warming from 90 to 97 degrees. Interval scales share the key feature of ordinal scales—higher numbers indicate higher relative levels of the variable—but interval scales go an important step further. Because these numbers represent equal intervals, we are able to add, subtract, and compute averages. That is, whereas we could not calculate the athlete's average finish, we can calculate the average temperature in San Francisco or the average age of participants.



HasseChr/iStock Editorial/Thinkstock

An ordinal scale can place these three women in first, second, and third, but it cannot tell you how far apart they finished in their race.

Ratio Scales

Ratio scales go one final step further, representing interval scales that also have a true *zero point*, that is, the potential for a complete absence of the conceptual variable. Physical measurements, such as length, weight, and time represent ratio scales, because it is possible to have a complete absence of any of these. Most behavioral measures also represent ratio scales, as it is possible to have zero drinks per day, zero presses of a reward button, or zero symptoms of the flu. Temperature in degrees Kelvin is measured on a ratio scale because 0 degrees Kelvin indicates an absence of molecular motion. (In contrast, 0 degrees Fahrenheit is merely a center point on the temperature scale.) Contrast these measurements with many of the conceptual variables featured in psychology research—no such things as zero attitude toward gun control or zero self-esteem exist. The big advantage of having a true zero point is that it allows us to add, subtract, multiply, and divide scale values. When we measure weight, for example, it makes sense to say that a 300-pound adult weighs twice as much as a 150-pound adult. Likewise, it makes sense to say that having two drinks per day is only one-fourth as many as having eight drinks per day.

Choosing and Using Scales of Measurement

The take-home point from the discussion of these four scales of measurement is twofold. First, researchers should always use the most powerful and flexible scale possible for their conceptual variables. In many cases, no choice is possible; time is measured on a ratio scale

and gender is measured on a nominal scale. But some cases permit researchers a bit more freedom in designing their study. For example, if someone were interested in correlating weight with happiness, the researcher could capture weight in a few different ways. One option would be to ask people their satisfaction with their current weight on a seven-point scale. However, the resulting data would be on an ordinal or interval scale (see discussion below), and the degree to which the researcher could manipulate the scale values would be limited. Another, more powerful option, would be to measure people's weight on a bathroom scale, resulting in ratio-scale data. Whenever possible, it is preferable to incorporate physical or behavioral measures. But it is also preferable—actually, required—to represent data accurately. Most variables in the social and behavioral sciences do not have a true zero point and must therefore be measured on nominal, ordinal, or interval scales.

Second, researchers should always be aware of the limitations of their measurement scale. As discussed above, these scales lend themselves to different amounts of mathematical manipulation. It is not possible to calculate statistical averages with anything less than an interval scale and not possible to multiply or divide anything less than a ratio scale. What does this mean for researchers? If they have collected ordinal data, they are limited to discussing the rank ordering of the values (e.g., the critics liked Restaurant A better than Restaurant B). If they have collected nominal data, they are limited to describing the different groups (e.g., percentages of Catholics and Protestants).

One prominent grey area for both of these points is the use of attitude scales in the social and behavioral sciences. If we were to ask people to rate their attitudes about the death penalty on a seven-point rating scale, would the scale be ordinal or interval? This consideration turns out to be a contentious issue in the field. From the conservative point of view, these attitude ratings constitute only *ordinal* scales. We know that a 7 indicates more endorsement than a 3 but cannot say that moving from a 3 to a 4 is equivalent to moving from a 6 to a 7 in people's minds. From the more liberal point of view, these attitude ratings can be viewed as interval scales. A researcher's perspective is often driven by practical concerns—treating these as equal intervals allows us to compute totals and averages for our variables. Chapter 4 will return to this issue in discussing the creation of questionnaire items. For now, a good guideline is to assume that these individual attitude questions represent ordinal scales by default.

Types of Measurement

Each of the four scales of measurement can be used across a wide variety of research designs. In this section, we shift gears slightly and discuss measurement at a more conceptual, less mathematical level. The types of dependent measures used in psychological research studies can be grouped into three broad categories: behavioral, physiological, and self-report.

Behavioral Measurement

As mentioned earlier, behavioral measures are those that involve direct and systematic recording of observable behaviors. If a research question involves the ways that married couples deal with conflict, the researcher could include a behavioral measure by observing the way participants interact during an argument. Do they cut one another off? Listen attentively?

Express hostility? Behaviors can be measured and quantified in one of four primary ways, as the scenario of observing married couples during conflict situations illustrates:

- **Frequency** measurements involve counting the number of times a behavior occurs. For example, researchers could count the number of times each member of the couple rolled his or her eyes as a measure of dismissive behavior.
- **Duration** measurements involve measuring the length of time a behavior lasts. For example, researchers could quantify the length of time the couple spends discussing positive versus negative topics as a measure of emotional tone.
- **Intensity** measurements involve measuring the strength or potency of a behavior. For example, researchers could quantify the intensity of anger or happiness in each minute of the conflict using ratings by trained judges.
- **Latency** measures involve measuring the delay before onset of a behavior. For example, researchers could measure the time between one person's provocative statement and the other person's response.

John Gottman, a psychologist at the University of Washington, has been conducting research along these lines for several decades, observing body language and interaction styles among married couples as they discuss an unresolved issue in their relationship (read more about this research and its implications for therapy on Dr. Gottman's website, <http://www.gottman.com/>). What all of these behavioral measures provide is an unobtrusive way to measure the health of a relationship. That is, the major strength of behavioral responses is that they are typically more honest and unfiltered than responses to questionnaires. As Chapter 4 will discuss, people are sometimes dishonest on questionnaires to convey a more positive (or less negative) impression.

Behavioral responses offer a particular benefit for researchers interested in unpopular attitudes, such as prejudice and discrimination. If we were to ask people the extent to which they dislike members of other ethnic groups, they might not admit to these prejudices. Alternatively, a researcher could adopt the approach used by Yale psychologist Jack Dovidio and colleagues and measure how close people sat to people of different ethnic and racial groups, using this distance as a subtle and effective behavioral measure of prejudice (see <http://www.yale.edu/intergroup/> for more information). But the primary downside to using behavioral measures may be evident: We end up having to infer the reasons that people behave as they do. Suppose that in one of these experiments, European-American participants, on average, sit farther away from African-Americans than from other European-Americans. This could—and often does—indicate prejudice; however, for the sake of argument, the farthest seat from the minority group member might also be the comfortable recliner with great lighting next to the window. To understand the reasons for behaviors, researchers have to supplement the behavioral measures with either physiological or self-report measurements.

Physiological Measurement

Physiological measures are those that involve quantifying bodily processes, including heart rate, brain activity, and facial muscle movements. If we were interested in the experience of test anxiety, we could measure heart rates as people complete a difficult math test. If we wanted to study emotional reactions to political speeches, we could measure heart rate, facial

muscles, and brain activity as people view video clips. These types of measures' big advantage is that they are the least subjective and controllable. It is incredibly difficult for people to control their heart rate or brain activity consciously, making these a great tool for assessing emotional reactions. However, as with behavioral measures, we also need some way to contextualize physiological data.

The best example of this shortcoming is the use of the polygraph, or lie detector, to detect deception. The lie-detector test involves connecting a variety of sensors to the body to measure heart rate, blood pressure, breathing rate, and sweating. All of these are physiological markers of the body's fight-or-flight stress response, and the test's goal is to measure whether someone shows signs of stress while being questioned. But here is the problem: Being falsely accused is also stressful. A trained polygraph examiner must place all of the accused's physiological responses in the proper context. Is the individual stressed throughout the exam or only stressed when asked whether he pilfered money from the cash box? Is the person stressed when asked about her relationship with her spouse because she killed him or because she was having an affair? The examiner has to be extremely careful to avoid false accusations based on misinterpretations of physiological responses. (For a recent commentary on the use of the polygraph in the courtroom, see <http://www.thedailybeast.com/articles/2015/02/04/the-polygraph-has-been-lying-for-90-years.html>). The same cautions apply to using these measures in psychological research: Does heart rate increase because participants are stressed by a political message, or because the experiment is taking too long, and they are late to another appointment? The researcher should always include additional measures in the study to help sort out the reasons behind physiological change.

Self-Report Measurement

Self-report measures are those that involve asking people to report on their own thoughts, feelings, and behaviors. If we were interested in the relationship between income and happiness, we could simply ask people to report their income and their level of happiness. If we wanted to know whether people were satisfied in their romantic relationships, we could



Digital Vision/Photodisc/Thinkstock

A self-report measure might be used to determine how likely voters are to support a candidate.

simply ask them to rate their degree of satisfaction. The major advantage of these measures is that they provide access to internal processes. That is, if we want insight into *why* people voted for their favorite political candidate, the only option is to ask them. However, as the text has suggested already, people may not necessarily be honest and forthright in their answers, especially when dealing with politically incorrect or unpopular attitudes. Chapter 4 will return to this tension and discuss ways to increase the likelihood of honest self-reported answers.

Two broad categories of self-report measures can be used. One of the most common approaches is to ask for people's responses using a **fixed-format** scale, which asks them to indicate their opinion on a preexisting scale. For example, a researcher might ask people, "How likely are you to vote for the Republican candidate for president?" on a scale

from 1 (not likely) to 7 (very likely). The other broad approach is to obtain responses using a **free-response** format, which asks people to express their opinion in an open-ended format. For example, researchers might ask people to explain, “What are the factors you consider in choosing a political candidate?” The trade-off between these two categories is essentially a choice between data that is easy to code and analyze and data that is rich and complex. In general, fixed-format scales are used more in quantitative research, while free-response formats are used more in qualitative research. Chapter 4 will discuss these categories further in a discussion of survey research.

Research: Thinking Critically

Neuroscience and Addictive Behaviors

Follow the link below to read an article by journalist Christian Nordqvist. In this article, Nordqvist reviews recent research suggesting that food addiction might involve brain mechanisms similar to those involved in drug addiction. As you read the article, consider what you have learned so far about the research process, and then respond to the questions below.

<http://www.medicalnewstoday.com/articles/221233.php>

Think About It:

1. Is the study described here descriptive, correlational, or experimental? Explain.
2. Can we conclude from this study that food addiction causes brain abnormalities? Why or why not?
3. The authors of the study concluded: “The current study also provides evidence that objectively measured biological differences are related to variations in YFAS (Yale Food Addiction Scale) scores, thus providing further support for the validity of the scale.” What type(s) of validity are they referring to? Explain.
4. What types of measures are included in this study (e.g., behavioral, self-report)? What are the strengths and limitations of these measures in this study?

Converging Operations: The Best of All Worlds

As these descriptions show, each type of measurement has its strengths and flaws. So, how do researchers decide which one to use? This question has to be answered for every case, and the answer involves consideration of three factors: *fit* with the research question; *insights* from previous research; and *practical* considerations like budget and equipment availability. However, in an ideal world, a program of research will use a wide variety of measures and designs. The term for this approach is **converging operations**, or the use of multiple research methods to solve a single problem. In essence, over the course of several studies—perhaps spanning several years—a researcher would address a research question using different designs, different measures, and different levels of analysis.

One good example of converging operations comes from the research of psychologist James Gross and his colleagues at Stanford University. Gross and his team study the ways that people regulate their emotional responses and has conducted this work using everything from questionnaires to brain scans (see <http://spl.stanford.edu/projects.html>).

One branch of Gross's research has examined the consequences of trying to either suppress emotions (pretend they are not happening) or reappraise them (think of them in a different light). Gross's team studies suppression by asking people to hold in their emotional reactions while watching a graphic medical video. The researchers study reappraisal by asking people to watch the same video while trying to view it as a medical student, thus changing the meaning of what they see. When people try to suppress their emotional responses, they experience an ironic *increase* in physiological and self-reported emotional responses, as well as deficits in cognitive and social functioning. When reappraising emotions, on the other hand, people experience lower levels of both reported and physiological emotion, without any loss of other functioning. In another branch of the research, Gross and colleagues have examined the neural processes at work when people change their perspective about an emotional event. In yet another branch of the research, they have used self-report measures to examine individual differences in emotional responses, with the goal of understanding why some people are more capable of managing their emotions than others. Taken together, these studies all converge into a more comprehensive picture of the process of emotion regulation than would be possible from any single study or method.

2.4 Hypothesis Testing

Regardless of the details of a particular study, be it correlational, experimental, or descriptive, all quantitative research follows the same process of testing a hypothesis. This section provides an overview of this process, including a discussion of the statistical logic, the five steps of the process, and the two ways we can make mistakes during our hypothesis test. Some of this material may be a review from statistics class, but it forms the basis of our scientific decision-making process and thus warrants repeating.

The Logic of Hypothesis Testing

Chapter 1 discussed several criteria for identifying a “good” theory, one of which is that theories have to be *falsifiable*. In other words, research questions should have the ability to be proven wrong under the right set of conditions. Why is this so important? This will sound counterintuitive at first, but by the standards of logic, when data run counter to a researcher's theory, that is more meaningful than when data support the theory.

For example, suppose we hypothesize that growing up in a low-income family puts children at higher risk for depression. If the data fit this pattern, our prediction might very well be correct. It is also possible, however, that these results are due to a third variable—perhaps low-income families grow up in more stressful neighborhoods, and stress turns out to increase a person's depression risk. Or, perhaps our sample accidentally contained an abnormal number of depressed people. This is why we are always cautious in interpreting positive results from a single study. Yet now, imagine that we test the same hypothesis and find that those who grew up in low-income families show a lower rate of depression. This is still a single study, but it suggests that our hypothesis may have been off-base.

Another way to think about this is from a statistical perspective. As the chapter discussed earlier, all measurements contain some amount of random error, which means that any pattern of data could be caused by random chance. This is the primary reason that research is never able to “prove” a theory. We will learn (or recall) from the study of statistics that at the end of any hypothesis test, we calculate a ***p* value**, representing the *probability* of observing our results—or results that are even more extreme—due entirely to random chance. Conceptually, we are calculating the probability that we are wrong rather than the probability that we are right in our predictions. And the bigger the effect, the smaller this probability will generally be. So, as strange as it seems, the ideal result of hypothesis testing is to have a small probability of being wrong.

This focus on falsifiability carries over to the way we test our hypotheses, in that the goal is to reject the possibility of results being due to chance. The starting point of a hypothesis test is to state a **null hypothesis**, or the assumption that the variables have no real effect in the overall population. This is another way of saying that observed patterns of data are due to random chance. In essence, we propose this null in hopes of minimizing the odds that it is true. Then, as a counterpoint to the null hypothesis, we propose an **alternative hypothesis** that represents the predicted pattern of results. This part is a little confusing, because the word *alternative* actually refers to the hypothesis in which we are interested. The term is employed because, in statistical jargon, the alternative hypothesis represents the predicted deviation from the null. These alternative hypotheses can be **directional**, meaning that we specify the direction of the effect, or **nondirectional**, meaning that we simply predict an effect.

Say we want to test the hypothesis that people like cats better than dogs. We would start with the null hypothesis, that people like cats and dogs the same amount (i.e., no difference). The next step is to state the alternative hypothesis (that is, our actual hypothesis), which in this case is that people will prefer cats. Because we are predicting a direction (cats more than dogs), this hypothesis is directional. The other option would be a nondirectional hypothesis, or simply stating that people’s cat preferences differ from their dog preferences. (Note that we have avoided predicting *which* one people like better, what makes it nondirectional.)

Finally, these three hypotheses can also be expressed using logical notation, as shown below. The letter *H* is used as an abbreviation for “hypothesis,” and the Greek letter μ is a common abbreviation for the mean, or average.

Conceptual Hypothesis: People like cats better than dogs.

Null Hypothesis: $H_0: \mu_{\text{cat}} = \mu_{\text{dog}}$

the “cat” mean is equal to the “dog” mean;

people like cats and dogs the same

Nondirectional Alternative Hypothesis: $H_1: \mu_{\text{cat}} \neq \mu_{\text{dog}}$

the “cat” mean is not equal to the “dog” mean;

people like cats and dogs different amounts

Directional Alternative Hypothesis: $H_1: \mu_{\text{cat}} > \mu_{\text{dog}}$

the “cat” mean is greater than the “dog” mean;

people like cats more than dogs

Why distinguish between directional and nondirectional hypotheses? A statistics class provides a more detailed answer, but it is important to note that this decision will have implications for the level of statistical significance. In essence, nondirectional hypotheses are less precise: “I think there is a difference,” versus “I believe cats are the preferred pet!” Because we always want to minimize the risk of coming to the wrong conclusion, we have to be more conservative with a nondirectional test. In this context, being conservative means needing a bigger group difference to feel confident in the results.

In the cats-versus-dogs example, a larger difference in ratings would be needed to support the claim that people like cats and dogs different amounts than would be needed to support the claim that people like cats more than dogs. The goal of all this statistical and logical jargon is to place hypothesis testing in the proper frame. The most important thing to remember is that hypothesis testing is designed to reject the null hypothesis, and statistical tests tell us how confident to be in this rejection.

Five Steps to Hypothesis Testing

Now that we understand how to frame a hypothesis, what does a researcher do with this information? Framing a hypothesis is the first step of a five-step process of testing a hypothesis. This section walks through an example of hypothesis testing from start to finish, that is, from an initial hypothesis to a conclusion about the hypothesis. Using a fictitious study, we will test the prediction that married couples without children are happier than those with children in the home. This example is inspired by an actual study by Harvard social psychologist Dan Gilbert and his colleagues, described in a news article at <http://www.telegraph.co.uk/news/1941195/Marriage-without-children-the-key-to-bliss.html>. The hypothesis may seem counterintuitive, but Gilbert’s research suggests that people tend to both overestimate the extent to which children will make them happy and underestimate the added stress and financial demands of having children in the house.

Step 1—State the Hypothesis

The first step in testing this hypothesis is to spell it out in logical terms. Remember that we want to start with the null hypothesis that the presence of children in a home has no effect. So, in this case, the null hypothesis would be that couples are equally happy with and without children. Or, in logical notation, $H_0: \mu_{\text{children}} = \mu_{\text{no children}}$ (i.e., the mean happiness rating for couples with children equals the mean happiness rating for couples without children). From there, we can spell out our alternative hypothesis; in this case, we predict that having children will make couples less happy. Because this is a directional hypothesis, we write $H_1: \mu_{\text{children}} < \mu_{\text{no children}}$ (i.e., the mean happiness rating for couples with children is lower than the mean happiness rating for couples without children).

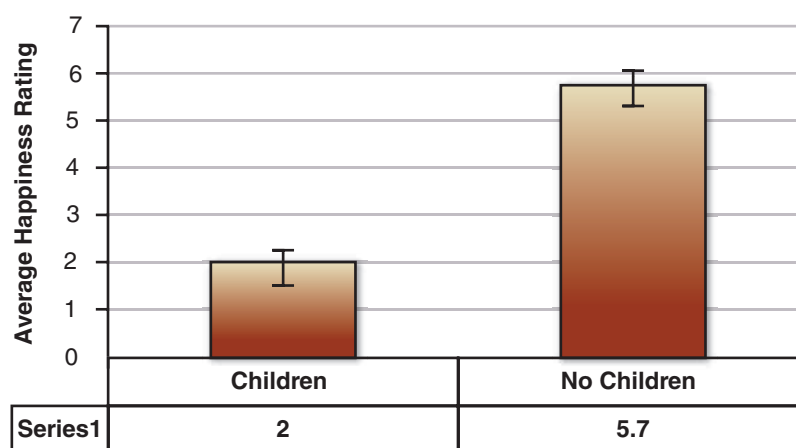
Step 2—Define Variables

Once we have an idea of the conceptual relationship that we want to test, we need to translate these concepts into measurable variables. As the chapter has discussed more than once, the decisions we make at this stage will trickle down and influence every subsequent step of the research process. For our current example, we will need to find a way to define the concept of “happiness,” as well as decide our criteria for “couples with / without children.” We have encountered happiness as an example before, so it seems fairly straightforward to define it based on participants’ responses to a happiness scale. But what does it mean for a couple to have children? Do the children need to be of a certain age, or would the study include everyone from parents of newborns to empty-nesters whose children are away at college? These types of decisions need to be made carefully, to ensure that we are controlling outside influences that might interfere with our hypothesis test. For example, couples who survive the trials and tribulations of raising a toddler without getting divorced may come to develop a more realistic set of expectations for their everyday happiness, compared to the parents of newborns or the parents of college students.

Step 3—Collect Data

The next step is to design and conduct a study that will test our hypothesis. The next three chapters will elaborate on this process in great detail, but the general idea is the same regardless of the design. In this case, the most appropriate design would be correlational because we want to predict happiness based on whether people have children. It would be impractical and unethical to randomly assign people to have children, so an experimental design is not possible in this case. One way to conduct our study would be to survey married couples about whether they had children and ask them to rate their current level of happiness with the marriage. Suppose we conduct this study and end up with the data in Figure 2.4.

Figure 2.4: Sample data for the “children and happiness” study



As the figure shows, the results suggest an average happiness rating of 5.7 for couples without children, compared to an average happiness rating of 2.0 for couples with children. These groups certainly look different—and encouraging for our hypothesis—but we need to be sure that the difference is big enough that we can reject the null hypothesis.

Step 4—Calculate Statistics

The next step in our hypothesis test is to calculate statistical tests to decide how confident we can be that the results are meaningful. Researchers have a wide variety of statistical tools at their disposal and different ways to analyze all manner of data. These tools can be broadly grouped into **descriptive statistics**, which describe the patterns and distribution of measured variables, and **inferential statistics**, which attempt to draw inferences about the population from which the sample was drawn. Researchers use inferential statistics to make decisions about the significance of the data. Statistics courses cover many of these in detail, and we will discuss a few examples throughout this book. All of these different techniques share a common principle: They attempt to make inferences by comparing the relationship among variables to the random variability of the data. As the chapter discussed earlier, people's measured levels of everything from happiness to heart rate can be influenced by a wide range of variables. The hope in testing our hypotheses is that differences in our measurements will primarily reflect differences in the variables we are studying. In the current example, we would want to see that differences in happiness ratings of the married couples were influenced more by the presence of children than by random fluctuations in happiness. Regardless of which statistic a researcher chooses to test the hypothesis, the resulting value will be translated into a measure of statistical significance, and this provides a key piece of information for the final decision.

Step 5—Make a Decision

Finally, we are able to draw a conclusion about our experiment. Based on the outcome of our statistical test (i.e., step 4), we will make one of two decisions about our null hypothesis:

Reject null: decide that the probability of the null being correct is sufficiently small; that is, results are due to differences in groups

or

Fail to reject null: decide that the probability of the null being correct is too big; that is, results are due to chance

Given the mean difference in Figure 2.4, and the small amount of error, our statistical test would certainly be significant, and we could be confident in rejecting the null hypothesis. At long last, we can express our findings in plain English: Couples with children are less happy than couples without children.

Having walked through this five-step process, we note an important fact. When it comes to analyzing data, to test hypotheses, researchers actually rely on a computer program for part of this process—Step 4 in particular. In these modern times, computing even a simple means comparison by hand is rare. Software programs such as SPSS, SAS, and Microsoft Excel can take a table of data, compute the mean difference, compare it to the variability, and calculate the probability that the results are due to chance. However, because these calculations happen behind the scenes, it is very important to understand the process. By understanding how the software operates, researchers can reach informed conclusions about their research questions. Otherwise, they risk making one of two possible errors in the hypothesis test, discussed in the next section.

Errors in Hypothesis Testing

In the children and happiness study, we concluded with a reasonable amount of confidence that our hypothesis was supported. Still, what if we made the wrong decision? Because our conclusions are based on interpreting probability, there is always a chance that we draw the wrong conclusion. In interpreting our hypothesis tests, we risk two potential errors, referred to as Type I and Type II errors.

Type I errors occur when the results are due to chance, but the researcher mistakenly concludes that the effect is significant. In other words, no effect of the variables exists in the population, but some quirk of the sample makes the effect appear significant. This error can be viewed as a false positive—researchers get excited over results that are not actually meaningful. In our children and happiness study, a Type I error would occur if children had no effect on happiness in the real world, but some quirk of chance made our “no children” group happier than the “children” group. For example, our sample of childless couples might accidentally contain a greater proportion of people with happy personalities or greater job stability or simply more marital satisfaction from the start.

Fortunately—although this error seems worrisome—we can generally compute the probability of making it. Our **alpha level** sets the bar for how extreme our data must be to reject the null hypothesis. At the end of the statistical calculation, a p value tells us how extreme the data actually are. When we set an alpha threshold of, say, 0.05, we are attempting to avoid a Type I error; our results will only be statistically significant if the effect outweighs the random variability by a big-enough amount. If the p value falls below our predetermined alpha level, we decide that the risk of a Type I error is sufficiently small and can therefore reject the null hypothesis. If, however, the p value is greater than (or even equal to) our alpha cutoff, we decide that the risk of Type I error is too high to ignore and will therefore fail to reject the null hypothesis.

Type II errors occur when the results are significant, but the researcher mistakenly concludes that they are due to chance. In other words, an effect of the variables does exist in the population, but some quirk of the sample makes the effect appear nonsignificant. This error can be viewed as a false negative—researchers miss results that actually could have been meaningful. In our children and happiness experiment, a Type II error would occur if couples without children really were happier than couples with children but some flaw in the experiment kept us from detecting the difference. For example, if our measures of happiness were poorly designed, people might vary in how they interpreted the items, and this source of error could make it difficult to spot an overall difference between the groups.

Although this error sounds disappointing, the good news is researchers have some fairly easy ways to avoid or minimize it. The key factor in reducing Type II error is to maximize the **power** of the statistical test, or the probability of detecting a real difference. In fact, power is **inversely related** to the probability of a Type II error—the higher the power, the lower the chance of Type II error. Power is analogous to the sensitivity, or accuracy, of the hypothesis test; it is under the researcher’s control in three main ways. First, as the section Reliability and Validity discussed it is important to make sure that measures are capturing what the researcher thinks they are. If the happiness scale actually captures something like narcissism, then this will cause problems for the hypothesis about the predictors of happiness. Second, it is important to be careful throughout the process of coding and analyzing data. Small

mistakes can occur at every step, from entering data, to calculating scale totals, to choosing an inappropriate analysis. And third, statistical tests generally have more power when the sample is larger. We will discuss each of these factors in more detail as we move through the course.

Research: Thinking Critically

The Truth About Cats and Dogs

Follow the link below to a press release on the website of the American Psychological Association. This press release describes a compelling research finding, from the social psychologist Allen McConnell, that examines the benefits of pet ownership for people's mental health. As you read the article, consider what you have learned so far about the research process, and then respond to the questions below.

<http://www.apa.org/news/press/releases/2011/07/cats-dogs.aspx>

Think About It:

1. In the first study described, 217 people answered surveys about well-being, and the researchers compared responses of pet owners to those of nonowners.
 - a. Is this study descriptive, correlational, or experimental?
 - b. Can we infer a causal relationship from this study? Explain.
 - c. Is there a possible directionality problem or third variable problem? Explain.
2. In the third study, what is the independent variable? What is the dependent variable?
3. What are the null hypotheses being tested in each of these studies? What are the alternate hypotheses?
4. What would a Type I decision error be in these studies? A Type II decision error?

Summary of Correct and Incorrect Decisions

In the real world, at the level of the entire population, our null hypothesis is either true or false. That is, if we could test our hypothesis by surveying every married couple in the world, we could say with 100% certainty whether or not the hypothesis was true. However, in each individual study, at the level of our sample, we have to decide either to reject the null or fail to reject it. Table 2.3 summarizes the four possible outcomes of a decision about a hypothesis test. In the top left and bottom right cells, we make the right decision—either rejecting a null hypothesis that is false or failing to reject one that is true in the population. In the bottom left cell of the table, we make a Type I error, rejecting a null hypothesis that is actually true, and mistakenly thinking our hypothesis is supported (i.e., a false positive). In the top right cell of the table, we make a Type II error, failing to reject a null hypothesis that is actually false, and mistakenly thinking our hypothesis should be rejected (i.e., a false negative).

Table 2.3: Errors and correct decisions in hypothesis testing

	Researcher's Decision	
	Reject Null	Fail to Reject Null
Null is FALSE	Correct Decision	Type II Error
Null is TRUE	Type I Error	Correct Decision

Chapter 1 (section 1.3) explained the process of drawing conclusions about “proof” and “disproof,” suggesting that neither one is ever possible in a single study. Now that we have covered the hypothesis-testing process, the reasoning behind rules regarding proof and disproof should be clearer. In fact, Type I and Type II errors are possible in every research study. Rejecting the null hypothesis in one study does not automatically mean that it is false, only that the null hypothesis could not explain the pattern of data in the study. Moreover, failing to reject the null in one study does not automatically mean that it is true, only that the pattern of data in the study does not support rejecting it. Science accumulates knowledge over the course of several related studies. It is only when these studies start to suggest the same conclusion that we can feel more confident in our decisions about the status of the null hypothesis.

Effect Size

So far, our discussion about hypothesis testing has been focused on statistical significance, and we have been concerned with the probability that our results might be due to random chance. However, keep in mind an additional piece of the puzzle of interpreting results. Imagine that someone has been placed in charge of testing a new drug that might help cure depression. The researcher might start by collecting a large sample of depressed patients and giving half of them the new drug and half of them a placebo. Now imagine that the new drug reduced symptoms by 20%, compared to a 10% reduction with the placebo. Is this effect big enough to become excited? If the new drug costs twice as much as existing ones, is it worth recommending? These questions revolve around the issue of **effect size**, a statistic used to represent the size, or magnitude, of an effect.

Size may be calculated in several ways, but as a general rule, bigger values mean a stronger effect. One of these statistics, **Cohen’s d** , is calculated as the difference between two means divided by their pooled variability. In this case, our variability measure is something called the **standard deviation**, which represents the *average* deviation of individual scores from the mean of the group. A larger standard deviation indicates that the scores are dispersed more widely around the mean. When we use this number in calculating Cohen’s d , the resulting values can therefore be expressed in terms of standard deviations; a d of 1 indicates that the means are one standard deviation apart. How big should we expect our effects to be? Based on his analyses of typical effect sizes in the social sciences, Cohen suggests the following benchmarks: $d = 0.20$ is a small effect; $d = 0.40$ is a moderate effect; and $d = 0.60$ is a large effect. In other words, a “large” effect in social and behavioral sciences accounts for a little over half of a standard deviation. For comparison purposes, the effect of the polio vaccine on reducing polio symptoms was a $d = 2.72$ (almost three standard deviations; Oshinsky, 2006). Our children and happiness study produces a $d = 3.82$, but fake data are always more impressive than real data.



diego_cervo/iStock/Thinkstock

Effect size can be used to help determine the effectiveness of a particular drug.

Effect size is useful in two primary ways. First, at the end of an experiment, we can calculate the exact size of the effect in our particular sample. This is a useful supplement to our test of statistical significance because it is less dependent on sample size. If we fail to reject the null

hypothesis in a small sample, the effect size might tell us whether the effect is big enough to test again with a larger sample. And, if we support our research hypothesis, the effect size provides valuable information about the usefulness of our findings. Imagine testing two different diabetes drugs in two different studies. Say both show a statistically significant reduction in symptoms, but Drug A has an effect size of $d = 0.50$, and Drug B has an effect size of $d = 2.5$. This tells us that Drug B has a larger effect and could therefore offer diabetes patients a bigger benefit.

The second use for effect size is in deciding on our sample size before the study begins. We learned earlier that our statistical tests generally have more power in a larger sample size. So why not run 10,000 participants in every single research study? The problem is that participants take time, money, and other resources, and not every study needs 10,000 people to detect an effect. Rather than striving for perfect power in every study, researchers usually compromise and hope for 80% power, which equates to only a 20% chance of Type II error. It turns out that we also have more power when the underlying effect is larger. Thus, we can take our estimates of effect size and determine the number of people we need to achieve at least 80% power.

The best way to perform these calculations is by using any of the power calculators available over the Internet. Figure 2.5 presents an annotated example using the calculator available at <http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html>. The values entered represent the means from our children and happiness study, plus the pooled standard deviation of 1.25. This calculation results in the previously mentioned d of 3.82. According to this calculator, we would only need two people per group to detect this effect in a future study—much cheaper and easier than 10,000.

Figure 2.5: Example of using effect size to estimate sample size

Inference for Means: Comparing Two Independent Samples

Choose which calculation you desire, enter the relevant population values for μ_1 (mean of population 1), μ_2 (mean of population 2), and σ (common standard deviation) and, if calculating power, a sample size (assumed the same for each sample). You may also modify α (type 1 error rate) and the power, if relevant. After making your entries, hit the calculate button at the bottom.

The screenshot shows a web-based calculator for comparing two independent samples. It has two main options: 'Calculate Sample Size (for specified Power)' and 'Calculate Power (for specified Sample Size)'. The first option is selected. Below this, there are input fields for μ_1 (5.7), μ_2 (2), and σ (1.25). There are also radio buttons for '1 Sided Test' and '2 Sided Test', with '1 Sided Test' selected. Further down, there are input fields for α (default .05) and desired power (default .80). At the bottom, there is a field for 'The sample size (for each sample separately) is:' which shows the value 2. A 'Calculate' button is at the bottom left. Three red boxes with arrows provide annotations: the first box points to the two main options; the second box points to the input fields for μ_1 , μ_2 , and σ ; the third box points to the sample size output field.

• ☒ Calculate Sample Size (for specified Power)
 • ☐ Calculate Power (for specified Sample Size)

Enter a value for μ_1 :
 Enter a value for μ_2 :
 Enter a value for σ :

• ☒ 1 Sided Test
 • ☐ 2 Sided Test

Enter a value for α (default is .05):
 Enter a value for desired power (default is .80):
 The sample size (for each sample separately) is:

*Note: In this example, the calculated sample size is too small for the approximate calculation here to be reliable and has been used for illustrative purposes only.

Reference: The calculations are the customary ones based on normal distributions. See for example *Hypothesis Testing: Two-Sample Inference- Estimation of Sample Size and Power for Comparing Two Means* in Bernard Rosner's **Fundamentals of Biostatistics**.

Summary and Resources

Chapter Summary

This chapter has discussed several basic principles of research design and emphasized the importance of ensuring that a study uses the best and most accurate measures available. We first examined the three main types of research design—descriptive, correlational, and experimental. These designs increase the amount of control that a researcher has. Descriptive designs can provide rich descriptions of various phenomena, from brain tumors to voting preferences, but are unable to delve into why these things happen. Correlational designs allow researchers to predict variables from other variables but are still unable to identify a causal relationship. This limitation in correlational designs occurs for two reasons: We do not know the direction of the relationship, and it is always possible that a third variable is causing both of them. Finally, experimental designs allow researchers to state with some certainty that one variable causes another because these designs involve systematically testing the impact of variables while controlling the environment. The downside of experimental designs is that they often have to sacrifice some realism to establish control.

The chapter focused on the importance of the accuracy and consistency of measures. In every research study, researchers start with an abstract variable and operationalize it into a measured variable. “Happiness” becomes a seven-point scale; “time” becomes the reading on a stopwatch, and so on. A researcher’s job is to evaluate the extent to which these measured variables capture the underlying concepts. One metric for evaluating this is the reliability, or consistency of the measures. Measures are more reliable when they are free from random error; we can assess this level of reliability by comparing multiple measures within the study. A second metric is the validity, or accuracy of the measures. Measures are more valid when they are free from systematic error, meaning that they measure what they claim to measure. We can generally assess validity by examining correlations with other measures, either to test the theoretical construct or to predict a behavioral criterion.

The chapter next discussed the different options for scaling and measuring variables. In addition to ensuring the accuracy and consistency of measures, it is critical to use a scaling method that matches the mathematical properties of the variable. Nominal scales represent arbitrary labels for categories; ordinal scales represent rank ordering of values; interval scales represent scales with equal intervals; and ratio scales represent variables with true zero points. A researcher should use the most powerful scale available—for example, by using behavioral counts rather than labels when possible. Nevertheless, researchers must also be aware of the limitations of the scale that they choose. While ratio scale values can be added, subtracted, divided, and multiplied, ordinal scale values cannot be manipulated. The chapter also identified three primary types of measurement. Behavioral measures involve observation and systematic recording of behavior; self-report measures involve asking people to report their own thoughts; and physiological measures involve measurements of bodily processes. Because each approach has advantages and disadvantages, many researchers use converging operations over the course of a research program, making use of all three to address a broad question.

Finally, this chapter discussed the process of hypothesis testing. Regardless of the question asked, the design used, and the way data are measured, all studies involve the same process of testing hypotheses using statistical results. The text explained the five steps of this process:

(1) Lay out the null and alternative hypotheses, (2) define variables, (3) collect data, (4) calculate the appropriate statistics, and (5) make a decision about the original hypothesis. Despite our best efforts, a hypothesis test occasionally leads to incorrect conclusions. A Type I error occurs when the researcher rejects the null but should not have; a Type II error occurs when the researcher fails to reject the null but could have under better conditions. As later chapters will discuss, we can reduce the odds of both errors through careful research design and analysis. The next three chapters will cover the specifics of the three types of research design: descriptive (Chapter 3), correlational (Chapter 4), and experimental (Chapter 5).

Key Terms

alpha level Predetermined probability cutoff for a hypothesis test; usually set as $p < 0.05$.

alternative hypothesis The predicted pattern of results or predicted deviation from the null.

behavioral measure A type of measure that involves direct and systematic recording of observable behaviors.

Cohen's d An effect size measure calculated as the difference between two means divided by their pooled standard deviation; the resulting values are expressed in terms of standard deviations.

concurrent validity The extent to which a self-report measure is able to predict a behavioral measure collected at the same time.

construct validity An assessment of how well the measures capture the underlying conceptual ideas (i.e., constructs) in a study.

continuum of control A framework for organizing and discussing research designs in terms of the amount of control the researcher has over the design.

control group Group that provides a basis of comparison for the experimental group.

convergent validity The extent to which a measure overlaps conceptually similar measures.

converging operations The use of multiple research methods to solve a single problem.

correlational research Research designed to predict thoughts, feelings, or behaviors.

criterion validity An assessment of validity based on the association between measures and relevant behavioral outcomes.

dependent variable Outcome variable that is measured by the experimenter.

descriptive research Research designed to describe thoughts, feelings, or behaviors.

descriptive statistics Statistics that describe the patterns and distribution of measured variables.

directional hypothesis Alternative hypothesis that specifies the direction of the effect; also called a one-tailed test.

directionality problem Limitation of correlational research; when we measure two variables at the same time, we have no way of knowing the cause of the relationship.

discriminant validity The extent to which a measure diverges from unrelated measures.

duration The length of time a behavior lasts.

effect size A statistic that represents the size, or magnitude, of an effect.

experimental group Group that receives the treatment of interest that provides the test of the hypothesis.

experimental research Research designed to explain thoughts, feelings, and behaviors and to make causal statements.

face validity The extent to which a measure *seems like* a good measure of the construct.

fixed-format A response format for self-report measures that asks people to indicate their opinions on a preexisting scale.

free-response A response format for self-report measures that asks people to express their opinions in an open-ended format.

frequency The number of times a behavior occurs.

independent variable Variable in an experimental design that is manipulated by the experimenter.

inferential statistics Statistics that attempt to draw inferences about the population from which a sample was drawn.

intensity The strength or potency of a behavior.

inter-item reliability The internal consistency among different questions on a questionnaire measure.

interrater reliability The consistency among judges' observations of participants' behavior.

interval scale A scaling method used to represent cases where the numbers on a measured variable correspond to equal distances on a conceptual variable.

latency The length of delay before onset of a behavior.

negative correlation Relationship between two variables such that as one variable increases, the other variable decreases.

nominal scale A scaling method used to label or identify a particular group or characteristic.

nondirectional hypothesis Alternative hypothesis that predicts only an effect, without specifying its direction; also called a two-tailed test.

null hypothesis The assumption that variables have no real effect on the overall population.

ordinal scale A scaling method used to represent ranked order of conceptual variables, such that higher numbers reflect increasing magnitude of the underlying variable.

physiological measure A type of measure that quantifies bodily processes, including heart rate, brain activity, and facial muscle movements.

positive correlation Relationship between two variables such that as one variable increases, so does the other variable.

power The probability of detecting a real difference; inversely related to the probability of a Type II error.

predictive validity The extent to which a self-report measure is able to predict a behavioral outcome.

p value A statistic representing the probability that results are due to random chance.

random error Chance fluctuations in the measurements.

ratio scale A scaling method used to represent interval scales that also have a true zero point, that is, a complete absence of the conceptual variable.

reliability Consistency of measurement; the extent to which a measured variable is free from random errors.

research design The specific method used to collect, analyze, and interpret data.

scaling The process of specifying the relationship between a conceptual variable and numbers on a quantitative measure.

self-report measure A type of measure that involves asking people to report their thoughts, feelings, and behaviors.

standard deviation An estimate of variability that represents the average deviation from the mean of the group; calculated by subtracting each score from the mean, adding up those differences, and dividing by the number of scores.

systematic errors Errors that systematically increase or decrease values on the measured variable.

test-retest reliability The consistency of the measure at different time points.

third variable problem Limitation of correlational research; when we measure two variables as they naturally occur, the possibility exists of a third variable that causes both of them.

Type I error A hypothesis testing error that occurs when results are due to chance but the conclusion mistakenly states that the effect is significant; a false positive.

Type II error A hypothesis testing error that occurs when the results are significant but the conclusion mistakenly states that they are due to chance; a false negative.

validity The accuracy of measurements; the extent to which they accurately measure what they are designed to measure and are free from systematic error.

Apply Your Knowledge

1. For each of the following research questions, tell whether the most appropriate strategy involves descriptive, correlational, or experimental research.
 - a. Are students more likely to cheat on exams in their first or last year of college?
 - b. Does writing about a traumatic experience result in better health?
 - c. What personality variables predict success in school?
2. Dr. Blutarsky is interested in predicting the link between poor parenting and teen alcohol abuse. To investigate this, he has parents fill out questionnaires about their parenting styles and then waits to see how likely their children are to abuse alcohol.
 - a. Identify the independent and dependent variables in this study.

Independent:

Dependent:

- b. What type of research design is Dr. Blutarsky using?
- c. Give an operational definition of “poor parenting” and “alcohol abuse.”

Poor parenting:

Alcohol abuse:

3. For each of the following, identify the scale of measurement:
 - a. placing children in gifted and special-needs programs based on ability
 - b. an “attitudes toward the president” scale, measured from 1 to 7
 - c. height measured in inches
 - d. the number of drinks consumed per day
4. For each of the following abstract concepts, suggest a way to measure it using a behavioral and self-report measure:

	Behavioral	Self-Report
Conformity		
Enjoyment of reading		
Leadership ability		
Paranoia		
Independence		

Critical Thinking Questions

1. Can a measure be reliable but not valid? Explain why or why not.
2. Explain the trade-off between Type I and Type II errors. Why might attempts to minimize one of these inflate the other?

Research Scenarios: Try It

Psychologists have long observed that the emotion of fear consists of multiple distinct components. Imagine encountering a bear while hiking through the woods. Your experience of fear can be broken down into cognitive components, such as the subjective feeling of fear accompanied by the thought, “Uh oh,” physiological components, such as a racing heart and a spike in blood pressure, and behavioral components, like freezing in place or running away.

Suppose you are assisting a researcher working on a new therapy for arachnophobia, and you are developing a comprehensive battery to measure fear of spiders. You want to create a set of measures that will be able to measure fear quickly and accurately so that you can determine whether the therapy works.

1. Which of the following measures would be the fastest way to measure the cognitive component of arachnophobia?
 - a. A free-response questionnaire
 - b. A fixed-format questionnaire
 - c. A brain scan
 - d. A measure of response time to subliminal presentations of spiders on a computer screen

2. One way of obtaining a behavioral index of arachnophobia is by using a behavioral approach test, or BAT. In this test, the experimenter measures in feet and inches how close a person is willing to get to a live spider in a glass box—this distance is called the BAT index. The closer the person is willing to get to the spider, the less afraid the individual is considered to be, and the smaller the BAT index. Which of the following would indicate convergent validity between your behavioral and self-report measures of arachnophobia?
 - a. Your questionnaire correlates positively with the BAT index.
 - b. Your questionnaire correlates negatively with the BAT index.
 - c. Your questionnaire does not correlate with the BAT index.
 - d. Your questionnaire seems like it measures the same thing as the BAT index.
3. Suppose your physiological measure of fear is a measure of perspiration. Psychologists measure perspiration by measuring how well the skin conducts tiny (and harmless) electrical impulses, a phenomenon called skin conductance. However, you observe that your participants seem to get anxious when you have only started setting up the electrodes used to measure skin conductance—before they even see a spider. You suspect that the apparatus used to measure skin conductance itself makes people fearful. What problem might result from this?
 - a. Low discriminant validity
 - b. Unreliable measurement
 - c. Random error in measurement
 - d. Systematic error in measurement
4. After developing your cognitive, physiological, and behavioral measures of arachnophobia, you and your colleague conduct the experimental trial of arachnophobia treatment. You administer your three measures before and after the therapy and are pleased to discover that all participants undergoing the new therapy show statistically significant reductions in all three measures of arachnophobia. In this case, what is the null hypothesis that you rejected?
 - a. Arachnophobia was greater before therapy than after.
 - b. Arachnophobia was greater after therapy than before.
 - c. Arachnophobia was the same before and after therapy.
 - d. The measures of arachnophobia were not affected equally by therapy.
5. After analyzing your data, you find that the effect size of therapy on self-reported fear was $d = 0.80$, and the effect size of therapy on the behavioral index was $d = 0.20$. Which of the following could you conclude?
 - a. Therapy had little to no effect on either subjective fear or fearful behavior.
 - b. Therapy had a large effect on both subjective fear and fearful behavior.
 - c. Therapy had a large effect on the subjective experience of fear, but only a small effect on fearful behavior.
 - d. It is impossible to compare how therapy influenced these two measures of fear.

See Appendix A for answers to Research Scenarios: Try It questions.

3 Descriptive Designs— Observing Behavior



Alexander Macfarlane/Axiom Photographic/Design Pics/Superstock

Learning Outcomes

By the end of this chapter, you should be able to:

- Explain the distinguishing features of qualitative research.
- Distinguish the key features, pros, and cons of case studies.
- Distinguish the key features, pros, and cons of archival research.
- Distinguish the key features, pros, and cons of observational research.
- Outline best practices for describing data, both graphically and numerically.

In the fall of 2009, Phoebe Prince and her family relocated from Ireland to South Hadley, Massachusetts. Phoebe was immediately singled out by bullies at her new high school and subjected to physical threats, insults about her Irish heritage, and harassing posts on her Facebook page. This relentless bullying continued until January of 2010, ending only because Phoebe elected to take her own life in order to escape her tormentors (“Report of plea deal,” 2011). Tragic stories like this one are all too common, and it should come as no surprise that the Centers for Disease Control and Prevention (2012) has identified bullying as a serious problem facing our nation’s children and adolescents.

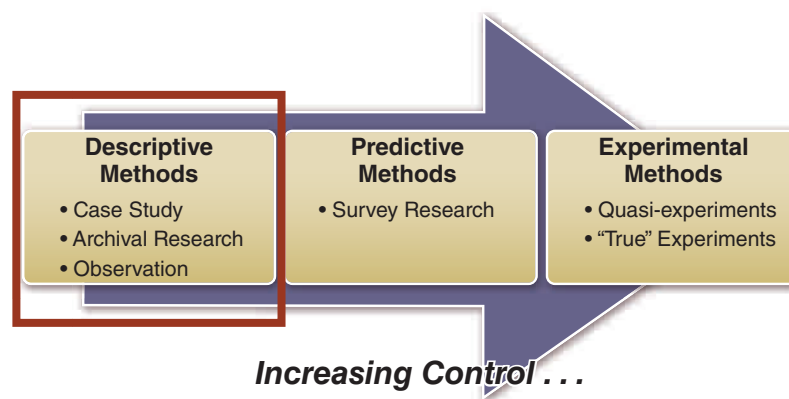
Scientific research on bullying began in Norway in the late 1970s in response to a wave of teen suicides. Work begun by psychologist Dan Olweus—and since continued by many others—has documented both the frequency and the consequences of bullying in the school system. Thus, we know that approximately one third of children are victims of bullying at some point during development, with between 5% and 10% bullied on a regular basis (Griffin & Gross, 2004; Nansel et al., 2001). Victimization by bullies has been linked with a wide range of emotional and behavioral problems, including depression, anxiety, self-reported health problems, and an increased risk of both violent behavior and suicide (for a detailed review, see Griffin & Gross, 2004). Recent research even suggests that bullying during adolescence may have a lasting impact on the body’s physiological stress response (Hamilton, Newman, Delville, & Delville, 2008).

Nevertheless, most of this research has a common limitation: It has studied the phenomenon of bullying using self-report survey measures. That is, researchers typically ask students and teachers to describe the extent of bullying in the schools. In many studies, researchers will also have students fill out a collection of survey measures, describing both bullying experiences and psychological functioning in their own words. These studies are conducted rigorously, and the measures they use certainly meet the criteria of reliability and validity discussed in Chapter 2 (2.2). However, as Wendy Craig, Professor of Psychology at Queen’s University, and Debra Pepler, a Distinguished Professor at York University, suggested in a 1997 article, this questionnaire approach cannot capture the full context of bullying behaviors. As we have already discussed, self-report measures are fully dependent on people’s ability and willingness to answer honestly and accurately. It is easy to imagine scenarios in which reports of bullying experiences might be downplayed out of fear, or perhaps misremembered simply due to the stress of the experience itself.

To address this limitation, Craig and Pepler (1997) decided to observe bullying behaviors as they occurred naturally on the playground. Among other things, the researchers found that acts of bullying occurred approximately every 7 minutes, lasted only about 38 seconds, and tended to occur within 120 feet of the school building. They also found that peers intervened to try to stop the bullying more than twice as often as adults did (11% versus 4%, respectively). These findings add significantly to scientific understanding of when and how bullying occurs. For our purposes, the most notable thing about them is that none of the findings could have been documented without directly observing and recording bullying behaviors on the playground. By using this technique, the researchers were able to gain a more thorough understanding of the phenomenon of bullying and, as a result, to provide real-world advice to teachers and parents.

One recurring theme in this book is that it is absolutely critical for researchers to pick the right research design to address their hypothesis. The next three chapters will discuss three specific categories of research designs, proceeding in order of increasing control over elements of the design (see Figure 3.1). This chapter focuses on descriptive research designs, in which the primary goal is to describe attitudes or behavior. We will begin by contrasting qualitative and quantitative approaches to description. We will then discuss three approaches to descriptive designs—studying single cases, mining existing archives, and observing behavior directly—covering the basic concept and the pros and cons of each. Finally, the chapter concludes with a discussion of guidelines for presenting descriptive data in both graphical and numeric form.

Figure 3.1: Descriptive designs on the continuum of control



3.1 Qualitative and Quantitative Methods

Chapter 1 explained that researchers generally take one of two broad approaches to answering their research questions. *Quantitative* research is a systematic and empirical approach that attempts to generalize results to other contexts, whereas *qualitative* research is a more descriptive approach that attempts to gain a deep understanding of particular cases and contexts. Before we discuss specific examples of descriptive designs, it is important to understand that these can represent either quantitative *or* qualitative perspectives. This section contrasts the two approaches in more detail.

Chapter 1 used the analogy of studying traffic patterns to contrast qualitative and quantitative methods—a qualitative researcher would likely study a single busy intersection in detail. This example illustrates a key point about this approach: Qualitative researchers are focused on interpreting and making sense out of what they observe rather than trying to simplify and quantify these observations. In general, qualitative research involves a collection of interviews and observations made in a natural setting. Regardless of the overall approach (qualitative or quantitative), data collection in the real world results in less control and structure than it does in a laboratory setting. But whereas quantitative researchers might view reduced control as a threat to reliability and validity, qualitative researchers view it as a strength of the study. Conducting observations in a natural setting makes it possible to capture people's natural and unfiltered responses.

As an example, consider two studies of the ways people respond to traumatic events. In a 1993 paper, psychologists James Pennebaker and Kent Harber took a quantitative approach to examining the community-wide impact of the 1989 Loma Prieta earthquake (near San Francisco). These researchers conducted phone surveys of 789 area residents, asking people to indicate, using a 10-point scale, how often they “thought about” and “talked about” the earthquake during the three-month period after its occurrence. In analyzing these data, Pennebaker and Harber discovered that people tend to stop talking about traumatic events about two weeks after they occurred but keep thinking about the event for approximately four more weeks. That is, the event is still on people’s minds, but they decide to stop discussing it with other people. In a follow-up study using the 1991 Gulf War, the same researchers found that this conflict leads to an increased risk of illness, measured via an increase in visits to the doctor (Pennebaker & Harber, 1993). The goal of the study was to gather data in a controlled manner and test a set of hypotheses about community responses to trauma.

Contrast Pennebaker and Harber’s approach with the more qualitative one taken by the developmental psychologist Paul Miller and colleagues (2012), who used a qualitative approach to study the ways that parents model coping behavior for their children. These researchers conducted semistructured interviews of 24 parents whose families had been evacuated following the 2007 wildfires in San Diego County and an additional 32 parents whose families had been evacuated following a 2008 series of deadly tornadoes in Tennessee. Because of a lack of prior research on how parents teach their children to cope with trauma, Miller and colleagues approached their interviews with the goal of “documenting and describing” (p. 8) these processes. That is, rather than attempt to impose structure and test a strict hypothesis, the researchers focused on learning from these interviews and letting the interviewees’ perspectives drive the acquisition of knowledge.



Lisafx/iStock/Thinkstock

Paul Miller’s research, which involved a series of semi-structured, qualitative interviews, attempted to document and describe a phenomenon rather than test a theory.

phenomena in their own right. So, for example, Miller and colleagues’ 2012 study of coping focused on “documenting and describing” the ways that parents teach children to cope and learning from a critical evaluation of the interview content. At risk of oversimplifying:

Qualitative and quantitative methods also differ quite strikingly in how they approach analyses of the data. Because all quantitative methods have the goal of discovering findings that can be generalized—that apply across different contexts—all quantitative studies must translate phenomena into numerical values and conduct statistical analyses. So, for example, Pennebaker and Harber’s (1993) study of coping with trauma measured the concrete value of “visits to the doctor,” and then compared changes in this number over time. In contrast, because qualitative methods have the goal of learning and interpreting phenomena from the ground up, qualitative studies focus on discovering the underlying meaning of

Quantitative methods gloss over some of the richness of experience in order to discover knowledge that can be generalized, while qualitative methods sacrifice the ability to generalize in order to capture the richness of experience.

As one final example of this contrast, consider the way that each approach would analyze the content of an interview. Interviewing people can be a very effective way to understand their experiences and can form the basis for many of the descriptive designs we cover in this chapter. A qualitative researcher would likely conduct a smaller number of interviews (perhaps only one, for a case study), due to the time required for analysis. The researcher would read each interview in depth and then start to identify themes that appeared across the entire set. These themes would serve as the basis for understanding people's experiences. (For an excellent deep dive into different theoretical approaches to interview analysis, see Smith [2008].) By comparison, a quantitative researcher would conduct a larger number of interviews, because quantitative text analysis can be very fast. Rather than read each interview, the researcher could input the text of these interviews into a software program, which could count and categorize the overall sentiment of the language people used. These counts and categories would then serve as the basis for quantifying people's experiences on a larger scale.

The following three sections examine three specific examples of descriptive designs—case studies, archival research, and observational research. Because each of these methods has the goal of describing attitudes, feelings, and behaviors, each one can be used from either a quantitative or a qualitative perspective. In other words, qualitative and quantitative researchers use many of the same general methods but do so with different goals. To illustrate this flexibility, each section concludes with a paragraph that contrasts qualitative and quantitative uses of the particular method.

3.2 Case Studies

At the 1996 meeting of the American Psychological Association, James Pennebaker—now chair of the psychology department at the University of Texas—delivered an invited address, describing his research on the benefits of therapeutic writing. Rather than follow the normal approach to an academic conference presentation, showing graphs and statistical tests to support his arguments, Pennebaker told a story. In the mid-1980s, when Pennebaker's lab was starting to study the effects of structured writing on physical and psychological health, one study participant was an American soldier who had served in the Vietnam War. Like many others, this soldier experienced difficulty adjusting to what had happened during the war and consequent trouble reintegrating into “normal” life. In Pennebaker's study, he was asked to simply spend 15 minutes per day, over the course of a week, writing about a traumatic experience—in this case, his tour of duty in Vietnam. At the end of this week, as might have been expected, this veteran felt awful, revisiting unpleasant memories over a decade old. But over the next few weeks, amazing things started to happen: The soldier slept better; he made fewer visits to his doctor; he even reconnected with his wife after a long separation.

Pennebaker's presentation was a **case study**, which provides a detailed, in-depth analysis of one person over a period of time. Although this case study was collected as part of a larger quantitative experiment, case studies are usually conducted in a therapeutic setting and

involve a series of interviews. An interviewer will typically study the subject in detail, recording everything from direct quotes and observations to his or her own interpretations. We encountered this technique briefly in Chapter 2 (2.1), in discussing Oliver Sacks's case studies of individuals learning to live with neurological impairments.

Pros and Cons of Case Studies

Case studies in psychology are a form of qualitative research and represent the lowest point on our continuum of control. Because they involve one person at a time, without a control group, case studies are often unsystematic. That is, the participants are chosen, rather than selected randomly, because they tell a compelling story or because they represent an unusual set of circumstances. Studying these individuals allows for a great deal of exploration, which can often inspire future research. However, it is nearly impossible to generalize from one case study to the larger population. In addition, because the case study includes both direct observation and the researcher's interpretation, a researcher's biases run the risk of influencing the interpretations. For example, Pennebaker's personal investment in demonstrating that writing has health benefits could have led to more positive interpretations of the Vietnam veteran's outcomes. However, in this particular case study, the benefits of writing mirror those seen in hundreds of controlled experimental studies involving thousands of people, so we can feel confident in the conclusions from the single case.

Case studies have two distinct advantages over other forms of research. First is the simple fact that anecdotes are persuasive. Despite Pennebaker's nontraditional approach to a scientific talk, the audience came away utterly convinced of the benefits of therapeutic writing. And although Oliver Sacks studied one neurological patient at a time, the collection of stories in his books sheds very convincing light on the ability of humans to adapt to their circumstances. Second, case studies provide a useful way to study rare populations and individuals with rare conditions. For example, from a scientific point of view, the ideal might be to gather a random sample of individuals living with severe memory impairment due to alcohol abuse and conduct some sort of controlled study in a laboratory environment. This approach could allow us to make causal statements about the results, as Chapter 5 (5.4) will discuss. But from a practical point of view, such a study would be nearly impossible to conduct, making case studies such as Sacks's interviews with William Thompson the best strategy for understanding this condition in depth.

Examples of Case Studies

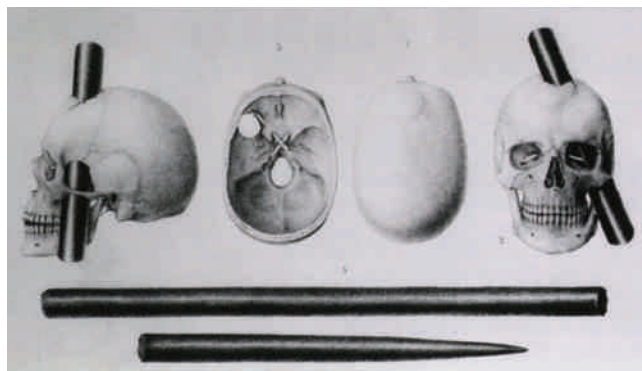
Throughout the history of psychology, case studies have been used to address a number of important questions and to provide a starting point for controlled quantitative studies. For example, in developing his theories of cognitive development, the Swiss psychologist Jean Piaget first studied the way that his own children developed and changed their thinking styles. Piaget proposed that children would progress through a series of four stages in the way that they approached the world—sensorimotor, preoperational, concrete operational, and formal operational—with each stage involving more sophisticated cognitive skills than the previous

stage. By observing his own children, Piaget noticed preliminary support for this theory and later was able to conduct more controlled research with larger populations.

Perhaps one of the most famous case studies in psychology is that of Phineas Gage, a 19th-century railroad worker who suffered severe brain damage. In September of 1848, Gage was working with a team to blast large sections of rock to make way for new rail lines. After a large hole was drilled into a section of rock, Gage's job was to pack the hole with gunpowder, sand, and a fuse and then tamp it down with a long cylindrical iron rod (known as a "tamping rod"). On this particular occasion, it seems Gage forgot to pack in the sand. So, when the iron rod struck gunpowder, the powder exploded, sending the 3-foot long iron rod through his face, behind his left eye, and out the top of his head. Against all odds, Gage survived this incident with relatively few physical side effects. However, everyone around him noticed that his personality had changed—Gage became more impulsive, violent, and argumentative. Gage's physician, John Harlow, reported the details of this case in an 1868 article. The following passage offers a strong example of the rich detail that is often characteristic of case studies:

He is fitful, irreverent, indulging at times in the grossest profanity (which was not previously his custom), manifesting but little deference for his fellows, impatient of restraint or advice when it conflicts with his desires. A child in his intellectual capacity and manifestations, he has the animal passions of a strong man. Previous to his injury, although untrained in the schools, he possessed a well-balanced mind, and was looked upon by those who knew him as a shrewd, smart businessman, very energetic and persistent in executing all his plans of operation. In this regard his mind was radically changed, so decidedly that his friends and acquaintances said he was "no longer Gage." (pp. 339–342)

Gage's transformation ultimately inspired a large body of work in psychology and neuroscience that attempts to understand the connections between brain areas and personality. The area of his brain destroyed by the tamping rod is known as the frontal lobe, now understood to play a critical role in impulse control, planning, and other high-level thought processes. Gage's story is a perfect illustration of the pros and cons of case studies: On the one hand, it is difficult to determine exactly how much the brain injury affected his behavior because he is only one person. On the other hand, Gage's tragedy inspired researchers to think about the connections among mind, brain, and personality. As a result, we now have a vast—and still growing—understanding of the brain. The story illustrates a key point about case studies: Although individual cases provide only limited knowledge about people in general, these cases often lead researchers to conduct additional work that does lead to generalizable knowledge.



Everett Collection

Various views show an iron rod embedded in Phineas Gage's (1823–1860) skull.

Qualitative versus Quantitative Approaches

Case studies tend to be qualitative more often than not: The goal of this method is to study a particular case in depth, as a way to learn more about a rare phenomenon. In both Pennebaker's study of the Vietnam veteran and Harlow's study of Phineas Gage, the researcher approached the interview process as a way to gather information and learn from the bottom up about the interviewee's experience. However, a case study can certainly represent quantitative research. This is often the case when researchers conduct a series of case studies, learning from the first one or the first few and then developing hypotheses to test on future cases. For example, a researcher could use the case of Phineas Gage as a starting point for hypotheses about frontal lobe injury, perhaps predicting that other cases would show similar changes in personality. Another way in which case studies can add a quantitative element is for researchers to conduct analyses within a single subject. For example, a researcher could study a patient with brain damage for several years following an injury, tracking the association between deterioration of brain regions with changes in personality and emotional responses. At the end of the day, though, these examples would still suffer the primary drawback of case studies: Because they examine a single individual, researchers find it difficult to generalize findings.

Research: Thinking Critically

Analyzing Acupuncture

Follow the link below to a press release from the Peninsula College of Medicine and Dentistry. This short article reviews recent research from the college, suggesting that acupuncture treatment might be of benefit to patients suffering from "unexplained" symptoms. As you read the article, consider what you have learned so far about the research process, and then respond to the questions below.

<http://www.sciencedaily.com/releases/2011/05/110530080513.htm>

Think about it:

1. In this study, researchers interviewed acupuncture patients using open-ended questions and recorded their verbal responses, which is a common qualitative research technique. What advantages does this approach have over administering a quantitative questionnaire with multiple-choice items?
2. What are some advantages of adding a qualitative element to a controlled medical trial like this?
3. What would be some disadvantages of relying exclusively on this approach?

3.3 Archival Research

Slightly further along the continuum of control is **archival research**, which involves drawing conclusions by analyzing existing sources of data, including both public and private records. Sociologist David Phillips (1977) hypothesized that media coverage of suicides would lead to "copycat" suicides. He tested this hypothesis by gathering archival data from two sources: front-page newspaper articles devoted to high-profile suicides and the number of fatalities

in the 11-day period following coverage of the suicide. By examining these patterns of data, Phillips found support for his hypothesis. Specifically, fatalities appeared to peak three days after coverage of a suicide, and a greater degree of publicity was associated with a greater peak in fatalities.

Pros and Cons of Archival Research

It is difficult to imagine a better way to test Phillips's hypothesis about copycat suicides. A researcher could never randomly assign people to learn about suicides and then wait to see whether they killed themselves. Nor could someone interview people right before they commit suicide to determine whether they were inspired by media coverage. Archival research provides a test of the hypothesis by examining data that already exist and, thereby, avoids most of the ethical and practical problems of other research designs. One key element of archival research is that it neatly sidesteps issues of **participant reactivity**, or the tendency of people to behave differently when they are aware of being observed. Any time research is conducted in a laboratory, participants know they are part of a study and may not behave in a completely natural manner. In contrast, archival data involves making use of records of people's natural behaviors. The subjects of Phillips's study of copycat suicides were individuals who decided to kill themselves, who had no awareness that they would be part of a research study.

Archival research is also an excellent strategy for examining trends and changes over time. For example, much of the evidence for global warming comes from observing upward trends in recorded temperatures around the globe. To gather this evidence, researchers dig into existing archives of weather patterns and conduct statistical tests of the changes over time. Psychologists and other social scientists also make use of this approach to examine population-level changes in everything from suicide rates to voting patterns over time. These comparisons can sometimes involve a blend of archival and current data. For example, a great deal of social-psychology research has been dedicated to understanding people's stereotypes about other groups. In a classic series of studies known as the "Princeton Trilogy," researchers documented the stereotypes held by Princeton students for 25 years (1933 to 1969). Social psychologist Stephanie Madon and her colleagues (2001) collected a new round of data but also conducted a new analysis of the previous archival data. These new analyses suggested that, over time, people have become more willing stereotype other groups, even as the stereotypes themselves have become less negative.



AP Photo

Copycat suicides often peak 3 days after media coverage of a high profile suicide, such as when Nirvana's Kurt Cobain killed himself in 1994.

One final advantage of archival research is that once a researcher gains access to the relevant archives, it requires relatively few resources. The typical laboratory experiment involves one participant at a time, sometimes requiring the dedicated attention of more than one research assistant for an hour or more. After researchers assemble data from the archives, though, conducting statistical analyses is a relatively simple matter. In a 2001 article, the psychologists Shannon Stirman and James Pennebaker used a text-analysis computer program to compare the language of poets who committed suicide (e.g., Sylvia Plath) with the language of similar poets who had not committed suicide (e.g., Denise Levertov). In total, these researchers examined 300 poems from 20 poets, half of whom had committed suicide. Consistent with Durkheim's theory of suicide as a form of "social disengagement," Stirman and Pennebaker (2001) found that suicidal poets used more self-references and fewer references to other people in their poems. The impressive part of the study is this: Once they had assembled their archive of poems, their computer program took only seconds to analyze the language and generate a statistical profile of each poet.

Overall, however, archival research is still relatively low on the continuum of control. Researchers have to accept the archival data in whatever form they exist, with no control over the way they were collected. For instance, in Stephanie Madon's (2001) reanalysis of the "Princeton Trilogy" data, she had to trust that the original researchers had collected the data in a reasonable and unbiased way. In addition, because archival data often represent natural behavior, it can be difficult to categorize and organize responses in a meaningful and quantitative way. The upshot is that archival research often requires some creativity on the researcher's part—such as analyzing poetry using a text-analysis program. In many cases, as we discuss next, the process of analyzing archives involves developing a coding strategy for extracting the most relevant information.

Content Analysis—Analyzing Archives

In most examples so far, the data come in a straightforward, ready-to-analyze form. That is, it is relatively simple to count the number of suicides, track the average temperature, or compare responses to questionnaires about stereotyping over time. In other cases, the data can come as a sloppy, disorganized mass of information. How does someone who wants to analyze literature, media images, or changes in race relations on television accomplish the analysis? These types of data can yield incredibly useful information, provided the researcher can develop a strategy for extracting it.

Mark Frank and Tom Gilovich—both psychologists at Cornell University—were interested in whether cultural associations with the color black affected behavior. In virtually all cultures, the term "black" is associated with evil—the bad guys wear black hats; people have a "black day" when things turn sour; and some are excluded from social groups by being "blacklisted" or "blackballed." These associations appear to be independent of any culture-specific prejudices regarding race or skin color. Frank and Gilovich (1988) wondered whether "a cue as subtle as the color of a person's clothing" (p. 74) would influence aggressive behavior. To test this hypothesis, they examined aggressive behaviors in professional football and hockey games, comparing teams whose uniforms were black to teams who wore other colors. Imagine for a moment being a researcher for this study. Professional sporting events contain a

wealth of behaviors and events. How would information about the relationship between uniform color and aggressive behavior be extracted?

Frank and Gilovich (1988) solved this problem by examining public records of penalty yards (football) and penalty minutes (hockey) because these represent instances of punishment for excessively aggressive behavior, as recognized by the referees. In addition, in both sports, the size of the penalty increases according to the degree of aggression. These penalty records were obtained from the central offices of both leagues, covering the period from 1970 to 1986. Consistent with the researchers' hypothesis, teams with black uniforms were "uncommonly aggressive" (p. 76). Most strikingly, two NHL hockey teams changed their uniforms to black during the period under study and showed a marked increase in penalty minutes with the new uniforms. One equally compelling alternative explanation is that, rather than the teams *acting* more aggressive in black uniforms, referees *perceived* them to be more aggressive while wearing black uniforms. Both explanations are consistent with the idea that cultural associations can affect behavior.

Even this analysis, however, is relatively straightforward because it involved data that were already in quantitative form (penalty yards and minutes). In many cases, the starting point is a jumbled mess of human behavior. In a pair of journal articles, psychologist Russell Weigel and colleagues (1980; 1995) examined the portrayal of race relations on prime-time television. To do so, they had to make several critical decisions about what to analyze and how to quantify it. The process of systematically extracting and analyzing the contents of a collection of information is known as **content analysis**. In essence, content analysis involves developing a plan to code and record specific behaviors and events in a consistent way. We can break this plan down into a three-step process.

Step 1—Identify Relevant Archives

Before we develop our coding scheme, we have to start by finding the most appropriate source of data. Sometimes the choice is fairly obvious: To compare temperature trends, the most relevant archives will be weather records. To track changes in stereotyping over time, the most relevant archive is questionnaire data assessing people's attitudes. In other cases, this decision involves careful consideration of both the research question and practical concerns. Frank and Gilovich decided to study penalties in professional sports because these data were both readily available (from the central league offices) and highly relevant to their hypothesis about aggression and uniform color.

Because these penalty records were publicly available, the researchers were able to access them easily. But if the research question involved sensitive or personal information—such as hospital records or personal correspondence—researchers would need to obtain permission from a responsible party. Say we wanted to analyze the love letters written by soldiers serving overseas and then try to predict relationship stability. Given the personal, even intimate nature of these letters, we would need permission from each person involved before proceeding with the study. However researchers manage to obtain access to private records, protecting the privacy and anonymity of the people involved is paramount. This would mean, for example, using pseudonyms and/or removing names and other identifiers from published excerpts of personal letters.

Step 2—Sample From the Archives

In Weigel's research on race relations, the most obvious choice of archives comprised snippets of both television programming and commercials. Yet this decision was only the first step of the process. Should the researchers examine every second of every program ever aired on television? Naturally not; instead, their approach was to take a smaller sample of television programming. Chapter 4 (4.3) will discuss sampling in more detail, but the basic process involves taking a smaller, representative collection of the broader population to conserve resources. Weigel and colleagues (1980) decided to sample one week's worth of prime-time programming from 1978, assembling videotapes of everything broadcast by the three major networks at the time (CBS, NBC, and ABC). The research team narrowed its sample by eliminating news, sports, and documentary programming because the hypotheses centered on portrayals of fictional characters of different races.

Step 3—Code and Analyze the Archives

Content analysis' third and most involved step is to develop a system for coding and analyzing the archival data. Even a sample of one week's worth of prime-time programming contains a near-infinite amount of information. In the race-relations studies, Weigel et al. elected to code four key variables: (1) the "total human appearance time," or time during which people were onscreen; (2) the "Black appearance time," in which Black characters appeared onscreen; (3) the "cross-racial appearance time," in which characters of two races were onscreen at the same time; and (4) the "cross-racial interaction time," in which cross-racial characters interacted. In the original (1980) paper, these authors reported that Black characters were shown only 9% of the time, and cross-racial interactions only 2% of the time. Fortunately, by the time of their 1995 follow-up study, the rate of Black appearances had doubled, and the rate of cross-racial interactions had more than tripled. However, depressingly little change occurred in some of the qualitative dimensions that they measured, including the degree of emotional connection between characters of different races.

This study also highlights the variety of options for coding complex behaviors. The four key ratings of "appearance time" consist of simply recording the amount of time that each person or group is onscreen. In addition, the researchers assessed several abstract qualities of interaction using judges' ratings. The degree of emotional connection, for instance, was measured by having judges rate the "extent to which cross-racial interactions were characterized by conditions promoting mutual respect and understanding" (Weigel et al., 1980, p. 888). As Chapter 2 (2.2) explained, any time researchers use judges' ratings, it is important to collect ratings from more than one rater and to make sure they agree in their assessments.

A researcher's goal is to find a systematic way to record the observations most relevant to the hypothesis. This is particularly true for quantitative research, where the key is to start with clear operational definitions that capture the variables of interest. This involves both deciding the most appropriate variables and the best way to measure these variables. For example, if someone who analyzes written communication might decide to compare words, sentences, characters, or themes across the sample. A study of newspaper coverage might code the amount of space or number of stories dedicated to a topic, while a study of television news might code the amount of airtime given to different positions. The best strategy in each case will be the one that best represents the variables of interest.

Qualitative versus Quantitative Approaches

Archival research can represent either qualitative or quantitative research, depending on the researcher's approach to the archives. Most of the examples in this section represent the quantitative approach: Frank and Gilovich (1988) counted penalties to test their hypothesis about aggression, and Stirman and Pennebaker (2001) counted words to test their hypothesis about suicide. However, the race-relations work by Weigel and colleagues (1980; 1995) represents a nice mix of qualitative and quantitative research. In the initial 1980 study, their primary goal was to document the portrayal of race relations on prime-time television, learning from the ground up (i.e., qualitative). In the 1995 follow-up study, though, the researchers primarily wanted to determine whether these portrayals had changed over a 15-year period. That is, they tested the hypothesis that race relations were portrayed in a more positive light (i.e., quantitative). Another way in which archival research can be qualitative is to study open-ended narratives, without attempting to impose structure upon them. This approach is commonly used to study free-flowing text, such as personal correspondence or letters to the editor in a newspaper. A researcher approaching these from a qualitative perspective would attempt to learn from these narratives, without attempting to impose structure via the use of content analyses.

3.4 Observational Research

Moving further along the continuum of control, we come to the descriptive design with the greatest amount of researcher control. **Observational research** involves studies that directly observe behavior and record these observations in an objective and systematic way. Your previous psychology courses may have explored the concept of attachment theory, which argues that an infant's bond with his or her primary caregiver has implications for later social and emotional development. Mary Ainsworth, a Canadian developmental psychologist, and John Bowlby, a British psychologist and psychiatrist, articulated this theory in the 1960s. They argued that children can form either "secure" or a variety of "insecure" attachments with their caregivers (Ainsworth & Bell, 1970; Bowlby, 1963).

To assess these classifications, Ainsworth and Bell developed an observational technique called the "strange situation." Mothers would arrive at their laboratory with their children for a series of structured interactions, including having the mother play with the infant, leave him alone with a stranger, and then return to the room after a brief absence. The researchers were most interested in coding the ways in which the infant responded to these various episodes (eight in total). One group of infants, for example, was curious when the mother left but then returned to playing with toys, trusting that she would return. Another group showed immediate distress when the mother left and clung to her nervously upon her return. Based on these and other behavioral observations, Ainsworth and colleagues classified these groups of infants as "securely" and "insecurely" attached to their mothers, respectively.



Rayes/Photodisc/Thinkstock

Observational research can be used to measure an infant's attachment to a caregiver.

Research: Making an Impact

Harry Harlow

In the 1950s, U.S. psychologist Harry Harlow conducted a landmark series of studies on the mother–infant bond using rhesus monkeys. Although contemporary standards would consider his research unethical, the results of his work revealed the importance of affection, attachment, and love on healthy childhood development.

Prior to Harlow’s findings, it was believed that infants attached to their mothers as a part of a drive to fulfill exclusively biological needs, in this case obtaining food and water and avoiding pain (Herman, 2007; van der Horst & van der Veer, 2008). In an effort to clarify the reasons that infants so clearly need maternal care, Harlow removed rhesus monkeys from their natural mothers several hours after birth, giving the young monkeys a choice between two surrogate “mothers.” Both mothers were made of wire, but one was bare and one was covered in terry cloth. Although the wire mother provided food via an attached bottle, the monkeys preferred the softer, terry-cloth mother, even though the latter provided no food (Harlow & Zimmerman, 1958; Herman, 2007).

Further research with the terry-cloth mothers contributed to the understanding of healthy attachment and childhood development (van der Horst & van der Veer, 2008). When the young monkeys were given the option to explore a room with their terry-cloth mothers and had the cloth mothers in the room with them, they used the mothers as a safe base. Similarly, when exposed to novel stimuli such as a loud noise, the monkeys would seek comfort from the cloth-covered surrogate (Harlow & Zimmerman, 1958). However, when the monkeys were left in the room without their cloth mothers, they reacted poorly—freezing up, crouching, crying, and screaming.

A control group of monkeys who were never exposed to either their real mothers or one of the surrogates revealed stunted forms of attachment and affection. They were left incapable of forming lasting emotional attachments with other monkeys (Herman, 2007). Based on this research, Harlow discovered the importance of proper emotional attachment, stressing the importance of physical and emotional bonding between infants and mothers (Harlow & Zimmerman, 1958; Herman, 2007).

Harlow’s influential research led to improved understanding of maternal bonding and child development (Herman, 2007). His research paved the way for improvements in infant and child care and in helping children cope with separation from their mothers (Bretherton, 1992; Du Plessis, 2009). In addition, Harlow’s work contributed to the improved treatment of children in orphanages, hospitals, day care centers, and schools (Herman, 2007; van der Horst & van der Veer, 2008).

Pros and Cons of Observational Research

Observational designs are well suited to a wide range of research questions, provided the questions can be addressed through *directly observable* behaviors and events. For example, researchers can observe parent–child interactions, or nonverbal cues to emotion, or even crowd behavior. However, if they are interested in studying thought processes—such as how close mothers feel to their children—then observation will not suffice. This point harkens back to the discussion of behavioral measures in Chapter 2 (2.2): In exchange for giving up access to internal processes, researchers gain access to unfiltered behavioral responses.

To capture these unfiltered behaviors, it is vital for the researcher to be as unobtrusive as possible. As we have already discussed, people have a tendency to change their behavior when they are being observed. In the bullying study by Craig and Pepler (1997) discussed at the beginning of this chapter, the researchers used video cameras to record children's behavior unobtrusively. Imagine how (artificially) low the occurrence of bullying might be if the playground had been surrounded by researchers with clipboards!

If researchers conduct an observational study in a laboratory setting, they have no way to hide the fact that people are being observed, but the use of one-way mirrors and video recordings can help people to become comfortable with the setting. Researchers who conduct an observational study out in the real world have even more possibilities for blending into the background, including using observers who are literally hidden. For example, someone hypothesizes that people are more likely to pick up garbage when the weather is nicer. Rather than station an observer with a clipboard by the trash can, the researcher could place someone out of sight behind a tree, or perhaps sitting on a park bench pretending to read a magazine. In both cases, people would be less conscious of being observed and therefore more likely to behave naturally.

One extremely clever strategy for blending in comes from a study by the social psychologist Muzafer Sherif et al. (1954), involving observations of cooperative and competitive behaviors among boys at a summer camp. For Sherif, it was particularly important to make observations in this context without the boys realizing they were part of a research study. Sherif took on the role of camp janitor, which allowed him to be a presence in nearly all of the camp activities. The boys never paid enough attention to the “janitor” to realize his omnipresence—or his discreet note-taking. The brilliance of this idea is that it takes advantage of the fact that people tend to blend into the background once we become used to their presence.

Types of Observational Research

Several variations of observational research exist, according to the amount of control that a researcher has over the data collection process. **Structured observation** involves creating a standard situation in a controlled setting and then observing participants' responses to a predetermined set of events. The “strange situation” studies of parent–child attachment (discussed above) are a good example of structured observation—mothers and infants are subjected to a series of eight structured episodes, and researchers systematically observe and record the infants' reactions. Even though these types of studies are conducted in a laboratory, they differ from experimental studies in an important way: Rather than systematically manipulate a variable to make comparisons, researchers present the same set of conditions to all participants.

Another example of structured observation comes from the research of John Gottman, a psychologist at the University of Washington. For nearly three decades, Gottman and his colleagues have conducted research on the interaction styles of married couples. Couples who take part in this research are invited for a three-hour session in a laboratory that closely resembles a living room. Gottman's goal is to make couples feel reasonably comfortable and natural in the setting to get them talking as they might do at home. After allowing them to settle in, Gottman adds the structured element by asking the couple to discuss an “ongoing issue or problem” in their marriage. The researchers then sit back to watch the sparks fly, recording

everything from verbal and nonverbal communication to measures of heart rate and blood pressure. Gottman has observed and tracked so many couples over the decades that he is able to predict, with remarkable accuracy, which couples will divorce in the 18 months following the lab visit (Gottman & Levenson, 1992).

Naturalistic observation, meanwhile, involves observing and systematically recording behavior in the real world. This can be conducted in two broad ways—with or without intervention on the part of the researcher. *Intervention* in this context means that the researcher manipulates some aspect of the environment and then observes people's responses. For example, a researcher might leave a shopping cart just a few feet away from the cart-return area and track whether people move the cart. (Given the number of carts that are abandoned just inches away from their proper destination, someone must be doing this research all the time.) Recall an example from Chapter 1 (the discussion of ethical dilemmas in section 1.5) in which Harari et al. (1995) used naturalistic observation to study whether people would help in emergency situations. In brief, these researchers staged what appeared to be an attempted rape in a public park and then observed whether groups or individual males were more likely to rush to the victim's aid.

The ABC network has developed a hit reality show that mimics this type of research. The show, *What Would You Do?*, sets up provocative situations in public settings and videotapes people's reactions (full episodes are available online at <http://abcnews.go.com/WhatWouldYouDo/>). An unwitting participant in one of these episodes might witness a customer stealing tips from a restaurant table, or a son berating his father for being gay, or a man proposing to his girlfriend who minutes earlier had been kissing another man at the bar. Of course, these observation "studies" are more interested in shock value than data collection (or Institutional Review Board [IRB] approval; see Section 1.5), but the overall approach can be a useful strategy to assess people's reactions to various situations. In fact, some of the scenarios on the show are based on classic studies in social psychology, such as the well-documented phenomenon that people are reluctant to take responsibility for helping in emergencies.

Alternatively, naturalistic studies can involve simply recording ongoing behavior without any attempt by the researchers to intervene or influence the situation. In these cases, the goal is to observe and record behavior in a completely natural setting. For example, researchers might station themselves at a liquor store and observe the numbers of men and women who buy beer versus wine. Or, they might observe the numbers of people who give money to the Salvation Army bell-ringers during the holiday season. A researcher can use this approach to compare different conditions, provided the differences occur naturally. That is, researchers could observe whether people donate more money to the Salvation Army on sunny or snowy days, or compare donation rates when the bell ringers are different genders or races. Do people give more money when the bell-ringer is an attractive female? Or do they give more to someone who looks needier? These are all research questions that could be addressed using a well-designed naturalistic observation study.

Finally, **participant observation** involves having the researcher(s) conduct observations while engaging in the same activities as the participants. The goal is to interact with these participants to gain better access and insight into their behaviors. In one famous example, the psychologist David Rosenhan (1973) was interested in the experience of people hospitalized for mental illness. To study these experiences, he had eight perfectly sane people gain admission to different mental hospitals. These fake patients were instructed to give accurate

life histories to a doctor but lie about one diagnostic symptom. They all claimed to hear an occasional voice saying the words “empty,” “hollow,” and “thud.” Such auditory hallucinations are a symptom of schizophrenia, and Rosenhan chose these words to vaguely suggest an existential crisis.

Once admitted, these “patients” behaved in a normal and cooperative manner, with instructions to convince hospital staff that they were healthy enough to be released. In the meantime, they observed life in the hospital and took notes on their experiences—a behavior that many doctors interpreted as “paranoid note-taking.” The main finding of this study was that hospital staff tended to view all patient behaviors through the lens of their initial diagnoses. Despite immediately acting “normally,” these fake patients were hospitalized an average of 19 days (with a range from 7 to 52) before being released. All but one was diagnosed with “schizophrenia in remission” upon release. Rosenhan’s other striking finding was that treatment was generally depersonalized, with staff spending little time with individual patients.

In another example of participant observation, Festinger, Riecken, and Schachter (1956) decided to join a doomsday cult to test their new theory of cognitive dissonance. Briefly, this theory argues that people are motivated to maintain a sense of consistency among their various thoughts and behaviors. So, for example, a person who smokes a cigarette despite being aware of the health risks might rationalize smoking by convincing herself that lung-cancer risk is really just genetic. In this case, Festinger and colleagues stumbled upon the case of a woman named Mrs. Keach, who was predicting the end of the world, via alien invasion, at 11 p.m. on a specific date six months in the future. What would happen, they wondered, when this prophecy failed to come true? (One can only imagine how shocked they would have been had the prophecy turned out to be correct.)



RENARD/BSIP/Superstock

Psychologists David Rosenhan’s study of staff and patients in a mental hospital found that patients tended to be treated based on their diagnosis, not on their actual behavior.

To answer this question, the researchers pretended to be new converts and joined the cult, living among the members and observing them as they made their preparations for doomsday. Sure enough, the day came, and 11 p.m. came and went without the world ending. Mrs. Keach first declared that she had forgotten to account for a time-zone difference, but as sunrise started to approach, the group members became restless. Finally, after a short absence to communicate with the aliens, Mrs. Keach returned with some good news: The aliens were so impressed with the devotion of the group that they decided to postpone their invasion. The group members rejoiced, rallying around this brilliant piece of rationalizing, and quickly began a new campaign to recruit new members.

As these examples illustrate, participant observation can provide access to amazing and one-of-a-kind data, including insights into group members’ thoughts and feelings. This approach

also provides access to groups that might be reluctant to allow outside observers. However, the participant approach has two clear disadvantages over other types of observation. The first problem is ethical; data are collected from individuals who do not have the opportunity to give informed consent. Indeed, the whole point of the technique is to observe people without their knowledge. Before an IRB can approve this kind of study, researchers must show an extremely compelling reason to ignore informed consent, as well as extremely rigorous measures to protect identities. The second problem is methodological; the approach provides ample opportunity for the objectivity of observations to be compromised by the close contact between researcher and participant. Because the researchers are a part of the group, they can change the dynamics in subtle ways, possibly leading the group to confirm their hypothesis. In addition, the group can shape the researchers' interpretations in subtle ways, leading them to miss important details.

Another spin on participant observation is called **ethnography**, or the scientific study of the customs of people and cultures. This is very much a qualitative method that focuses on observing people in the real world and learning about a culture from the perspective of the person being studied—that is, learning from the ground up rather than testing hypotheses. Ethnography is used primarily in other social-science fields, such as anthropology. In one famous example, the cultural anthropologist Margaret Mead (1928) used this approach to shed light on differences in social norms around adolescence between American and Samoan societies. Mead's conclusions were based on interviews she conducted over a six-month period, observing and living alongside a group of 68 young women. Mead concluded from these interviews that Samoan children and adolescents are largely ignored until they reach the age of 16 and become full members of society. Among her more provocative claims was the idea that Samoan adolescents were much more liberal in their sexual attitudes and behaviors than American adolescents.

Mead's work has been the subject of criticism by a handful of other anthropologists, one of whom has even suggested that Mead was taken in by an elaborate joke played by the group of young girls. Still others have come to Mead's rescue and challenged the critics' interpretations. The nature of this debate between Mead's critics and her supporters highlights a distinctive characteristic of qualitative methods: "Winning" the argument is based on challenging interpretations of the original interviews and observations. In contrast, disagreements around quantitative methods are generally based on examining statistical results from hypothesis testing. While quantitative methods may lose much of the richness of people's experiences, they do offer an arguably more objective way of settling theoretical disputes.

Steps in Observational Research

One of the major strengths of observational research is its high degree of **ecological validity**; that is, the research can be conducted in situations that closely resemble the real world. Think of the chapter examples so far—married couples observed in a living-room-like laboratory; doomsday cults observed from within; bullying behaviors on the school playground. In every case, people's behaviors are observed in the natural environment or something very close to it. However, this ecological validity comes at a price; the real world is a jumble of information, some relevant, some not so much. The challenge for researchers, then, is to decide on a system that provides the best test of their hypothesis, one that can sort out the signal from the

noise. This section discusses a three-step process for conducting observational research. The key point to note right away is that most of this process involves making decisions ahead of time so that the process of data collection is smooth, simple, and systematic.

Step 1—Develop a Hypothesis

For research to be systematic, it is important to impose structure by having a clear research question, and, in the case of quantitative research, a clear hypothesis as well. Other chapters have covered hypotheses in detail, but the main points bear repeating: A hypothesis must be testable and falsifiable, meaning that it must be framed in such a way that it can be addressed through empirical data and might be disconfirmed by these data. In the example involving Salvation Army donations, we predicted that people might donate more money to an attractive bell-ringer. This hypothesis could easily be tested empirically and could just as easily be disconfirmed by the right set of data—say, if attractive bell-ringers brought in the *fewest* donations.

This particular example also highlights an additional important feature of observational hypotheses; namely, they must be based on observable behaviors. That is, we can safely make predictions about the amount of money people will donate because we can directly observe it. We are, nonetheless, unable to make predictions in this context about the *reasons* for donations. We would have no way to observe, say, that people donate more to attractive bell-ringers because they are trying to impress them. In sum, one limitation of observing behavior in the real world is that it prevents researchers from delving into the cognitive and motivational reasons behind the behaviors.

Step 2—Decide What and How to Sample

Once a researcher has developed a hypothesis that is testable, falsifiable, and observable, the next step is to decide what kind of information to gather from the environment to test this hypothesis. The simple fact is that the world is too complex to sample everything. Imagine that someone wanted to observe the dinner rush at a restaurant. A nearly infinite list of possibilities for observation presents itself: What time does the restaurant get crowded? How often do people send their food back to the kitchen? What are the most popular dishes? How often do people get in arguments with the wait staff? To simplify the process of observing behavior, the researcher will need to take a **sample**, or a smaller portion of the population, that is relevant to the hypothesis. That is, rather than observing “dinner at the restaurant,” the researcher’s goal is to narrow his or her focus to something as specific as “the number of people waiting in line for a table at 6 p.m. versus 9 p.m.”

The choice of what and how to sample will ultimately depend on the best fit for the hypothesis.



Steve Mason/Photodisc/Thinkstock

The dinner scene at a busy restaurant offers a wide variety of behaviors to observe. In order to simplify the observation process, researchers should narrow the focus by taking a sample.

The context of observational research offers three strategies for sampling behaviors and events. The first strategy, **time sampling**, involves comparing behaviors during different time intervals. For example, to test the hypothesis that football teams make more mistakes when they start to get tired, researchers could count the number of penalties in the first five minutes and the last five minutes of the game. This data would allow researchers to compare mistakes at one time interval with mistakes at another time interval. In the case of Festinger's (1956) study of a doomsday cult, time sampling was used to compare how the group members behaved before and after their prophecy failed to come true.

The second strategy, **individual sampling**, involves collecting data by observing one person at a time to test hypotheses about individual behaviors. Many of the examples already discussed involve individual sampling: Ainsworth and colleagues (1970) tested their hypotheses about attachment behaviors by observing individual infants, while Gottman (1992) tests his hypotheses about romantic relationships by observing one married couple at a time. These types of data allow researchers to examine behavior at the individual level and test hypotheses about the kinds of things people do—from the way they argue with their spouses to whether they wear team colors to a football game.

The third strategy, **event sampling**, involves observing and recording behaviors that occur throughout an event. For example, we could track the number of fights that break out during an event such as a football game, or the number of times people leave the restaurant without paying the check. This strategy allows for testing hypotheses about the types of behaviors that occur in a particular environment or setting. For instance, a researcher might compare the number of fights that break out in a professional football versus a professional hockey game. Or, the next time we host a party, we could count the number of wine bottles versus beer bottles that end up in the recycling bin. The distinguishing feature of this strategy is its focus on *occurrence* of behaviors more than on the *individuals* performing these behaviors.

Step 3—Record and Code Behavior

Having formulated a hypothesis and decided on the best sampling strategy, researchers must perform one final and critical step before beginning data collection. Namely, they have to develop good operational definitions of the variables by translating the underlying concepts into *measurable* variables. Gottman's research turns the concept of marital interactions into a range of measurable variables, such as the number of dismissive comments and passive-aggressive sighing—all things that can be observed and counted objectively. Rosenhan's 1973 study involving fake schizophrenic patients turned the concept of patient experience into measurable variables such as the amount of time staff members spent with each patient—again, something very straightforward to observe.

It is vital that researchers decide up front what kinds and categories of behavior they will be observing and recording. In the last section, we narrowed down our observation of dinner at the restaurant to the number of people in line at 6 p.m. versus the number of people in line at 9 p.m. But how can we be sure of an accurate count? What if two people are waiting by the door while the other two members of the group are sitting at the bar? Are those at the bar waiting for a table or simply having drinks? One possibility might be to count the number of individuals who walk through the door in different time periods, although our count could be inflated by those who give up on waiting or who only enter to sneak in and out of the restroom.

In short, observing behavior in the real world can be messy. The best way to deal with this mess is to develop a clear and consistent categorization scheme and stick with it. That is, in testing a hypothesis about the most crowded time at a restaurant, researchers would choose one method of counting people and use it for the duration of the study. In part, this choice of a method is a judgment call, but researchers' judgment should be informed by three criteria. First, they should consider *practical issues*, such as whether their categories can be directly observed. A researcher can observe the number of people who leave the restaurant but cannot observe whether they got impatient. Second, they should consider *theoretical issues*, such as how well the categories represent the underlying theory. Why did researchers decide to study the most crowded time at the restaurant? Perhaps this particular restaurant is in a new, up-and-coming neighborhood, and they expect the restaurant to become crowded over the course of the evening. The time would also lead researchers to include people sitting both at tables and at the bar—because this crowd may come to the restaurant with the sole intention of staying at the bar. Finally, researchers should consider *previous research* in choosing their categories. Have other researchers studied dining patterns in restaurants? What kinds of behaviors did they observe? If these categories make sense for the project, researchers may feel free to re-use them—no need to reinvent the wheel.

Last but not least, a researcher should take a step back and evaluate both the validity and the reliability of the coding system. (See Section 2.2 for a review of these terms.) Validity in this case means making sure the categories capture the underlying variables in the hypothesis (i.e., construct validity; see Section 2.2). For example, in Gottman's studies of marital interactions, some of the most important variables are the emotions expressed by both partners. One way to observe emotions would be to count the number of times a person smiles. However, we would have to think carefully about the validity of this measure, because smiling could indicate either genuine happiness or condescension. As a general rule, the better and more specific researchers' operational definitions, the more valid their measures will be (Chapter 2).

Reliability in this context means making sure data are collected in a consistent way. If research involves more than one observer using the same system, their data should look roughly the same (i.e., interrater reliability). This reliability is accomplished in part by making the observation task simple and straightforward—for example, having trained assistants use a checklist to record behaviors rather than depending on open-ended notes. The other key to improving reliability is careful training of the observers, giving them detailed instructions and ample opportunities to practice the rating system.

Observation Examples

To explain how all of this comes together, we will explore a pair of examples, from research question to data collection.

Example 1—Theater Restroom Usage

First, imagine, for the sake of this example, that someone is interested in whether people are more likely to use the restroom before or after watching a movie. Such a research question could provide valuable information for theater owners in planning employee schedules (i.e., when are bathrooms most likely to need cleaning). Thus, studying patterns of human behavior results in valuable applied knowledge.

The first step is to develop a specific, testable, and observable hypothesis. In this case, we might predict that people are more likely to use the restroom after the movie, as a result of consuming those 64-ounce sodas during the movie. Just for fun, we will also compare the restroom usage of men and women. Perhaps men are more likely to wait until after the movie, whereas women are just as likely to go before as after? This pattern of data might look something like the percentages in Table 3.1. That is, men make 80% of their restroom visits after the movie and 20% before the movie, while women make about 50% of their restroom visits at each time.

Table 3.1: Hypothesized restroom visits

Gender	Men	Women
Before movie	20%	50%
After movie	80%	50%
Total	100%	100%

The next step is to decide on the best sampling strategy to test this hypothesis. Of the three sampling strategies discussed—individual, event, and time—which one seems most relevant here? The best option would probably be time sampling because the hypothesis involves comparing the number of restroom visitors in

two time periods (before versus after the movie). So, in this case, we would need to define a time interval for collecting data. We could limit our observations to the 10 minutes before the previews begin and the 10 minutes after the credits end. The potential problem here, of course, is that some people might use either the previews or the end credits as a chance to use the restroom. Another complication arises in trying to determine which movie people are watching; in a giant multiplex theater, movies start just as others are finishing. One possible solution, then, would be to narrow the sample to movie theaters that show only one movie at a time and to define the sampling times based on the actual movie start- and end-times.

Having determined a sampling strategy, the next step is to identify the types of behaviors we want to record. This particular hypothesis poses a challenge because it deals with a rather private behavior. To faithfully record people “using the restroom,” we would need to station researchers in both men’s and women’s restrooms to verify that people actually, well, “use” the restroom while they are in it. However, this strategy poses the potential downside that the researcher’s presence (standing in the corner of the restroom) will affect people’s behavior. Another, less intrusive option would be to stand outside the restroom and simply count “the number of people who enter.” The downside to *that*, of course, is that we technically do not know why people are going into the restroom. But sometimes research involves making these sorts of compromises—in this case, we chose to sacrifice a bit of precision in favor of a less-intrusive measurement. This compromise would also serve to reduce ethical issues with observing people in the restroom.

So, in sum, we started with the hypothesis that men are more likely to use the restroom after a movie, while women use the restroom equally before and after. We then decided that the best sampling strategy would be to identify a movie theater showing only one movie and to sample from the 10-minute periods before and after the actual movie’s running time.

Finally, we decided that the best strategy for recording behavior would be to station observers outside the restrooms and count the number of people who enter. Now, say we conduct these observations every evening for one week and collect the data in Table 3.2.

Table 3.2: Findings from observing restroom visits

Gender	Men	Women
Before movie	75 (25%)	300 (60%)
After movie	225 (75%)	200 (40%)
Total	300 (100%)	500 (100%)

Notice that more women ($N = 500$) than men ($N = 300$) attended the movie theater during our week of sampling. The real test of our hypothesis, however, comes from examining the percentages within gender groups. That is, of the 300 men who went into the restroom, what percentage of them did so before the movie and what percentage of them did so after the movie? In this dataset, women used the restroom with relatively equal frequency before (60%) and after (40%) the movie. Men, in contrast, were three times as likely to use the restroom after (75%) than before (25%) the movie. In other words, our hypothesis appears to be confirmed by examining these percentages.

Example 2—Cell Phone Usage While Driving

Imagine that we are interested in patterns of cell phone usage among drivers. Several recent studies have reported that drivers using cell phones are as impaired as drunk drivers, making this an important public safety issue. Thus, if we could understand the contexts in which people are most likely to use cell phones, it would provide valuable information for developing guidelines for safe and legal use of these devices. So, this study might count the number of drivers using cell phones in two settings: while navigating rush-hour traffic and while moving on the freeway.

The first step is to develop a specific, testable, and observable hypothesis. In this case, we might predict that people are more likely to use cell phones when they are bored in the car. So, we hypothesize that we will see more drivers using cell phones while stuck in rush-hour traffic than while moving on the freeway.

The next step is to decide on the best sampling strategy to test this hypothesis. Of the three sampling strategies discussed—individual, event, and time—which one seems most relevant here? The best option would probably be individual sampling because we are interested in the cell phone usage of individual drivers. That is, for each individual car we see during the observation period, we want to know whether the driver is using a cell phone. One strategy for collecting these observations would be to station observers along a fast-moving stretch of freeway, as well as along a stretch of road that is clogged during rush hour. These observers would keep a record of each passing car and note whether the driver is on the phone.

After selecting a sampling strategy, we next must decide the types of behaviors to record. One challenge this study presents is how broadly to define cell phone usage. Should we include both talking and text messaging? Given our interest in distraction and public safety, we probably want to include text messaging. Several states have recently banned this practice while driving, often in response to tragic accidents. Because we will be observing moving vehicles, the most reliable approach might be to simply note whether drivers have a cell phone in their hand. As with the restroom study, we sacrifice a little bit of precision (i.e., knowing what the driver is using the cell phone for) to capture behaviors that are easier to record.

To sum up, we started with the hypothesis that drivers would be more likely to use cell phones when stuck in traffic. We then decided that the best sampling strategy would be to station observers along two stretches of road who would note whether drivers were using cell phones. Finally, we decided that the cell phone usage would be defined as each driver holding a cell phone. Now, suppose we conducted these observations over a 24-hour period and collected the data in Table 3.3.

Table 3.3: Findings from observing cell phone usage

	Rush Hour	Highway
Cell Phone	30 (30%)	200 (67%)
No Cell Phone	70 (70%)	100 (33%)
Total	100	300

phones? In this data set, 30% of those in rush hour were using cell phones, compared with 67% of those on the highway. In other words, the data did not confirm our hypothesis. Drivers in rush hour were less than half as likely to be using cell phones. The next step in this research program would be to speculate on the reasons the data contradicted the hypothesis.

The results show that more cars passed by on the highway ($N = 300$) than on the street during the rush-hour stretch ($N = 100$). The real test of our hypothesis, though, comes from examining the percentages within each stretch. That is, of the 100 people observed during rush hour and the 300 observed on the highway, what percentage was using cell

Qualitative versus Quantitative Approaches

The general method of observation lends itself equally well to qualitative and quantitative approaches, although some types of observation fit one approach better than the other. For example, structured observation tends to focus on hypothesis testing and quantification of responses. In Mary Ainsworth's (1970) "strange situation" research (described previously), the primary goal was to expose children to a predetermined script of events and to test hypotheses about how children with secure and insecure attachments would respond to these events. In contrast, naturalistic observation—and, to a greater extent, participant observation—tends to focus on learning from events as they occur naturally. In Leon Festinger's "doomsday cult" study, the researchers joined the group to observe the ways members reacted when their prophecy failed to come true. Margaret Mead (1928) spent several months living with Samoan adolescents to understand social norms around coming of age.

Research: Thinking Critically

"Irritable Heart" Syndrome in Civil War Veterans

Follow the link below to an article by science writer and editor K. Kris Hirst. In this article, Hirst reviews compelling research from health psychologist Roxanne Cohen Silver and her colleagues at the University of California, Irvine. Cohen Silver and her colleagues reviewed the service records of 15,027 Civil War veterans, finding an astounding rate of mental illness—long before post-traumatic stress disorder was recognized. As you read the article, consider what you have learned so far about the research process, and then respond to the questions below.

<http://psychology.about.com/od/ptsd/a/irritableheart.htm>

Think about it:

1. What hypotheses are the researchers testing in this study?
2. How did the researchers quantify trauma experienced by Civil War soldiers? Do you think this is a valid way to operationalize trauma? Explain why or why not.
3. Would this research be best described as case studies, archival research, or natural observation? Does the study involve elements of more than one type? Explain.

3.5 Describing Your Data

Before we move on from descriptive research designs, this last section discusses the process of presenting descriptive data in both graphical and numeric form. No matter how the researcher presents data, a good description is accurate, concise, and easy to understand. In other words, researchers have to represent the data accurately and in the most efficient way possible so that their audience can understand it. Another, more eloquent way to think of these principles is to take the advice of Edward Tufte, a statistician and expert in the display of visual information. Tufte (2001) suggests that when people view visual displays, they should spend time on *content-reasoning* rather than *design-decoding*. The sole purpose of designing visual presentations is to communicate information. So, the audience should spend time thinking about the information being presented, not trying to puzzle through the display itself. The following sections explain guidelines for accomplishing this goal in both numeric and visual form.

Table 3.4 presents hypothetical data from a sample of 20 participants. In this example, we have asked people to report their gender and ethnicity, as well as answer questions about their overall life satisfaction and daily stress. Each row in this table represents one participant

Table 3.4: Raw data from a sample of 20 individuals

Subject ID	Gender	Ethnicity	Life satisfaction	Daily stress
1	Male	White	40	10
2	Male	White	47	9
3	Female	Asian	29	8
4	Male	White	32	9
5	Female	Hispanic	25	3
6	Female	Hispanic	35	3
7	Female	White	28	8
8	Male	Hispanic	40	9
9	Male	Asian	37	10
10	Female	African-American	30	10
11	Male	White	43	8
12	Male	Asian	40	4
13	Male	White	48	7
14	Female	African-American	30	4
15	Female	White	37	7
16	Male	Hispanic	40	1
17	Female	White	36	1
18	Male	African-American	45	8
19	Female	White	42	8
20	Female	African-American	38	7

in the study, and each column represents one of the variables for which data were collected. This chapter focuses on ways to describe the sample characteristics. Later chapters will return to these principles in discussing graphs that display the relationship between two or more variables.

Numeric Descriptions

Because psychology is a scientific discipline, it often expresses preference for presenting data in number form. These numbers provide a metric that can be used to compare findings from one study to another, to evaluate the overall consistency of whatever phenomenon is being studied. Following is a brief overview of some common numeric descriptors for data.

Frequency Tables

Often, a good first step in approaching a data set is to obtain a sense of the frequencies for demographic variables—in this example, gender and ethnicity. The **frequency tables** shown in Table 3.5 are designed to present the number and percentage of the sample that fall into each of a set of categories. As this pair of tables shows, the sample consisted of an equal number of men and women (i.e., 50% for each gender). The majority of participants were White (45%), with the remainder divided almost equally between African-American (20%), Asian (15%), and Hispanic (20%) ethnicities.

Researchers can gain a lot of information from numerical summaries of data. In fact, numeric descriptors form the starting point for doing inferential statistics and testing hypotheses. A statistics course explores these statistics in detail, but for now it is important to understand that two numeric descriptors can provide a wealth of information about a data set: measures of central tendency and measures of dispersion.

Table 3.5: Frequency table summarizing ethnicity and sex distribution

Gender	Frequency	Percentage
Female	10	50.0
Male	10	50.0
Total	20	100.0
Ethnicity	Frequency	Percentage
African-American	4	20.0
Asian	3	15.0
Hispanic	4	20.0
White	9	45.0
Total	20	100.0

Measures of Central Tendency

The first number we need to describe our data is a measure of **central tendency**, which represents the most typical case in our data set. Central tendency is a single number that provides an overall sense of all the numbers. Think of what happens when colors are mixed: Adding yellow to blue creates green, so green gives us an overall sense of the combination of the two colors. In the same way, think of a household where one parent has a high salary, another has a moderate salary, and a teenager makes minimum wage. Taking the average of all three gives us an overall sense of the income for the entire household.

Central tendency can be represented by these three indices:

The **mean** is the mathematical average of a data set, calculated by adding up all the scores in the data set and then dividing this total by the number of scores in the data set. Because we are adding and dividing our scores, the mean can only be calculated using interval or ratio data (see Chapter 2 for a review of the four scales of measurement).

The **median**, another measure of central tendency, represents the number in the middle of a dataset, with 50% of scores both above and below it. The median is identified by placing the list of values in ascending numeric order, then selecting the number in the middle. This measure of central tendency can be used for ordinal, interval, or ratio data because it does not require mathematical manipulation to obtain.

The final measure of central tendency, the **mode**, represents the most frequent score in a data set and is obtained either by visual inspection of the values or by consulting a frequency table like in the one in Table 3.5. Because the mode represents a simple frequency count, it can be used with any of the four scales of measurement. In addition, it is the only measure of central tendency that is valid for use with nominal data—that is, those that do not have a numerical value—since the numbers assigned to these data are arbitrary.

One important takeaway is that the scale of measurement largely dictates the choice between measures of central tendency—nominal scales can only use the mode, and only interval or ratio scales can use the mean. (For a review of these scales of measurement, see Chapter 2, Section 2.3.) The other piece of the puzzle is to consider which measure best represents the data. Remember that the central tendency is a way to represent the “typical” case using a single number, so the goal is to settle on the most representative number. The examples in Table 3.6 illustrate this process.

Table 3.6: Comparing the mean, median, and mode

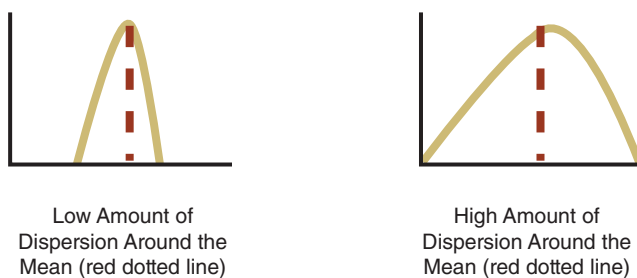
Data	Mean	Median	Mode	Discussion
1,2,3,4,5,11,11	5.29	4	11	- Both the mean and the median seem to represent the data fairly well. - The mean is a slightly better choice because it hints at the higher scores. - The mode is not representative—two people seem to have higher scores than everyone else.
1,1,1,5,10,10,100	18.29	5	1	- The mean is inflated by the atypical score of 100 and therefore does not represent the data accurately. - The mode is also not representative because it ignores the higher values. - In this case, the median is the most representative value to describe this dataset.

Measures of Dispersion

The second measure used to describe a dataset is a measure of **dispersion**, or the spread of scores around the central tendency—also referred to as measures of “variability.” Measures of dispersion tell us just how typical the typical score is. If the dispersion is low, then scores are

clustered tightly around the central tendency; if dispersion is higher, then the scores stretch out farther from the central tendency. Figure 3.2 presents a conceptual illustration of dispersion. The graph on the left has a low amount of dispersion because the scores (i.e., the yellow curve) cluster tightly around the average value (i.e., the red dotted line). The graph on the right shows a high amount of dispersion because the scores (yellow curve) spread out widely from the average value (red dotted line). The graph on the right might represent the earlier example of household income: The average income represents all three family members, but between the high-earning parent and the minimum-wage-earning teenager is a fairly large spread.

Figure 3.2: Two distributions with a low versus high amount of dispersion



One of the most straightforward measures of dispersion is the **range**, which is the difference between the highest and lowest scores. In Table 3.6, the range of the first dataset would be found by simply subtracting the lowest value (1) from the highest value (11), to get a range of 10. The range is useful in giving a general idea of the spread of scores, although it does not say much about how tightly these scores cluster around the mean.

The most common measures of dispersion are the **variance** and standard deviation, both of which represent the average difference between the mean and each individual score. The variance is calculated by subtracting each score from the mean to obtain a **deviation score**, squaring and summing these individual deviation scores, and then dividing by the sample size. The more scores are spread out around the mean, the higher the sum of these deviation scores will be, and therefore the higher the variance will be. Another common measure, the standard deviation (*SD*), is calculated as the square root of the variance.

Once we know the central tendency and the dispersion of variables, we have a good sense of what the sample looks like. These numbers also provide a valuable part in calculating the inferential statistics that we ultimately use to test our hypotheses.

Standard Scores

So far we have discussed ways to describe a particular sample in numeric terms. What do we do when we want to compare results from different samples—or from studies using different scales? Say we want to compare the anxiety levels of two people; unfortunately, in this example, these people were measured using different anxiety scales:

Joe scored 25 on the ABC Anxiety Scale, which has a mean of 15 and a standard deviation of 2.

Deb scored 40 on the XYZ Anxiety Scale, which has a mean of 30 and a standard deviation of 10.

At first glance, Deb's anxiety score appears higher, but note that the scales have different properties: The ABC scale has an average score of 15, while the XYZ scale has a higher average score of 30. The dispersion of these scales is also different; scores on the ABC scale cluster more tightly around the mean (i.e., the standard deviation is 2 compared to 10 on the XYZ scale).

The solution for comparing these scores is to convert both of them to **standard scores** (often expressed as **z scores**), which represent the distance of each score from the sample mean, expressed in standard deviation units. Standard scores let researchers translate raw scores into distributions with a predefined mean and standard deviation for easier interpretation. For example, scores on IQ tests are converted (i.e., *standardized*) onto a scale that has a mean of 100 and a standard deviation of 15. This tells us that a person with an IQ score of 100 is right at the average for the population, while someone with a score of 130 is two standard deviations above average.

The formula for a z score is worth examining in greater detail, as a way to understand the broader concept. Memorizing or using the formula in this research methods course is not required. The formula for a z score is:

$$z = (x - M)/SD$$

This formula subtracts the mean (M) from the individual score (x) and then divides this difference by the standard deviation of the sample (SD). To compare Joe's score with Deb's score, we simply substitute the appropriate numbers, using the mean and standard deviation from the scale that each one completed. This enables us to place scores from very different distributions on the same scale, making them easier to compare with one another. So, in this case:

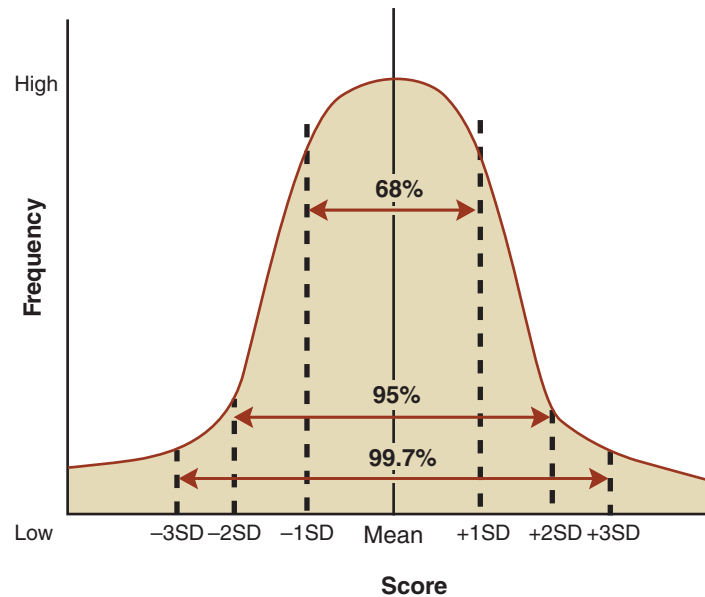
$$\text{Joe: } z = (x - M)/SD = (25 - 15)/2 = 10/2 = 5$$

$$\text{Deb: } z = (x - M)/SD = (40 - 30)/10 = 10/10 = 1$$

The resulting scores represent each person's score in standard deviation terms: Joe is 5 standard deviations above the mean of the ABC scale, while Deb is only 1 standard deviation above the mean of the XYZ scale. Or, in plain English, Joe is considerably more anxious than Deb.

To understand just *how* anxious Joe is, it is helpful to know a bit about why this technique works. Anyone who has taken a statistics class will have encountered the concept of the **normal distribution** (or "bell curve"), a symmetric distribution with an equal number of scores on either side of the mean, as Figure 3.3 illustrates.

It turns out that many variables in the social and behavioral sciences fit this normal distribution, provided the sample sizes are large enough. A normal distribution is useful because it has a consistent set of properties, such as having the same value for mean, median, and mode. In addition, if the distribution is normal, each standard deviation cuts off a known percentage of the curve, as illustrated in Figure 3.3. That is, 68% of scores will fall within ± 1 standard deviation of the mean; 95% of scores will fall within $\pm$ two standard deviations; and 99.7% of scores will fall within $\pm$ three standard deviations.

Figure 3.3: Standard deviations and the normal distribution

These percentages allow us to understand individual data points in even more useful ways, because we can easily move back and forth between z scores, percentages, and standard deviations. Take the example of Joe and Deb's anxiety scores: Deb has a z score of 1, which means her anxiety is 1 standard deviation above the mean. Furthermore, as we can see by consulting the normal distribution (Figure 3.3), her anxiety level is higher than 84% of the population. Joe has a z score of 5, which means his anxiety is 5 standard deviations above the mean. This also means that his anxiety is higher than 99.999% of the population. (For a handy online calculator that converts between z scores and percentages, see: <http://www.measuringusability.com/pcalcz.php>.)

Discussions of intelligence test scores also commonly use the relationship between z scores and percentiles. Tests that purport to measure IQ are converted to a scale that has a mean of 100 and a standard deviation of 15. Because IQ is normally distributed, we can move easily back and forth between z scores and percentages. For example, someone who has an IQ test score of 130 falls 2 standard deviations above the mean and falls in the upper 2.5% of the population. A person with an IQ test score of 70 is 2 standard deviations below the mean and thus falls in the bottom 2.5% of the population.

Ultimately, the use of standard scores allows us to take data that have been collected on different scales—perhaps in different laboratories and different countries—and place them on the same metric for comparison. As we have discussed in several contexts, science is all about the accumulation of knowledge one study at a time. The best support for an idea comes when data from different researchers, using different measures to capture the same concept, back the idea. The ability to convert these different measures back to the same metric is an invaluable tool for researchers who want to compare research results.

Visual Descriptions

Displaying data in visual form is often one of the most effective ways to communicate findings—as the cliché goes, a picture is worth a thousand words. What sort of visual should a researcher use? The choice of graphs is guided by two criteria: the scale of measurement and the best fit for the results. This section introduces some of the most common visual displays, based on hypothetical data used in Table 3.4.

Displaying Frequencies

One common type of graph is the **bar graph**, which also summarizes the frequency of data by category. Figure 3.4a depicts a bar graph, showing four categories of ethnicity along the horizontal axis and the number of people falling into each category indicated by the height of the bars. So, for example, this sample contains nine White participants and four Hispanic participants. Notice that these bar graphs contain exactly the same information as the frequency table in Table 3.5. When reporting results in a paper, a researcher would, of course, use only one of these methods. More often than not, graphical displays are the most effective way to communicate information.

Figure 3.4a: Bar graph displaying frequency by ethnicity

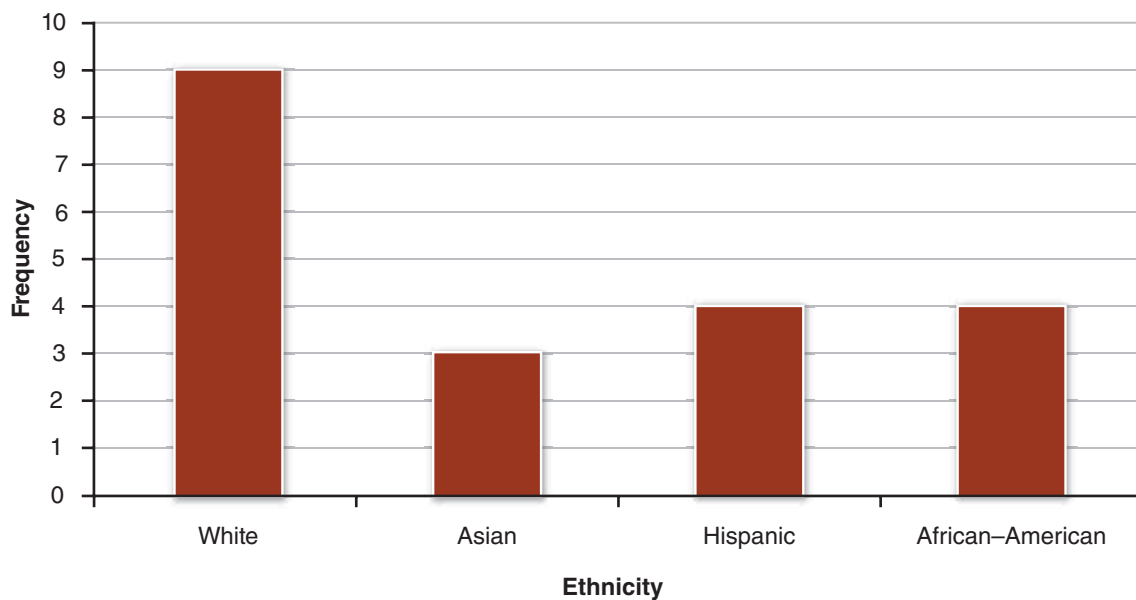
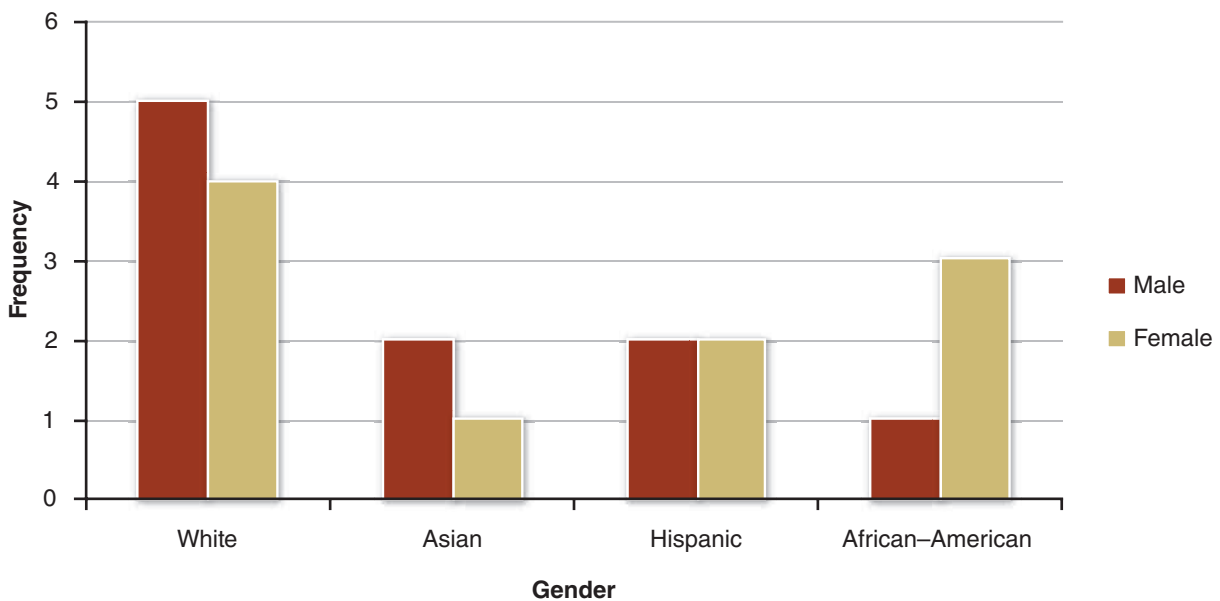


Figure 3.4b shows another variation on the bar graph, the **clustered bar graph**, which summarizes the frequency by two categories at one time. In this case, the bar graph displays information about both gender and ethnicity. As in the previous graph, categories of ethnicity are displayed along the horizontal axis. But this time, we have divided the total number of each ethnicity by the gender of respondents—indicated using different colored bars. For example, notice that the nine White participants are divided into five males and four females. Similarly, the four African-American participants are divided into one male and three females.

Figure 3.4b: Clustered bar graph displaying frequency by ethnicity and gender



Keep in mind that bar graphs are used for qualitative, or nominal, categories. We could just as easily have listed Caucasian participants second, third, or fourth along the axis because ethnicity is measured on a nominal scale.

When we want to present quantitative data—that is, those values measured on an ordinal, interval, or ratio scale—we use a different kind of graph called a **histogram**. As Figure 3.5a shows, histograms are drawn with the bars touching one another to indicate that the categories are quantitative and on a continuous scale. This figure has broken down the “life-satisfaction” values into three categories (less than 30, 31–40, and 41–50) and displayed the frequencies for each category in numerical order. For example, six people had life satisfaction scores falling between 31 and 40.

Finally, all of our bar graphs and histograms so far have displayed data that have been split into categories. However, as Figure 3.5b illustrates, histograms can also present data on a continuous scale. Figure 3.5b also has an additional new feature—a curved line overlaid on the graph. This curve represents a normal distribution and allows us to gauge visually how close our sample data are to being normally distributed.

Figure 3.5a: Histogram showing frequencies by life satisfaction (quantitative) categories

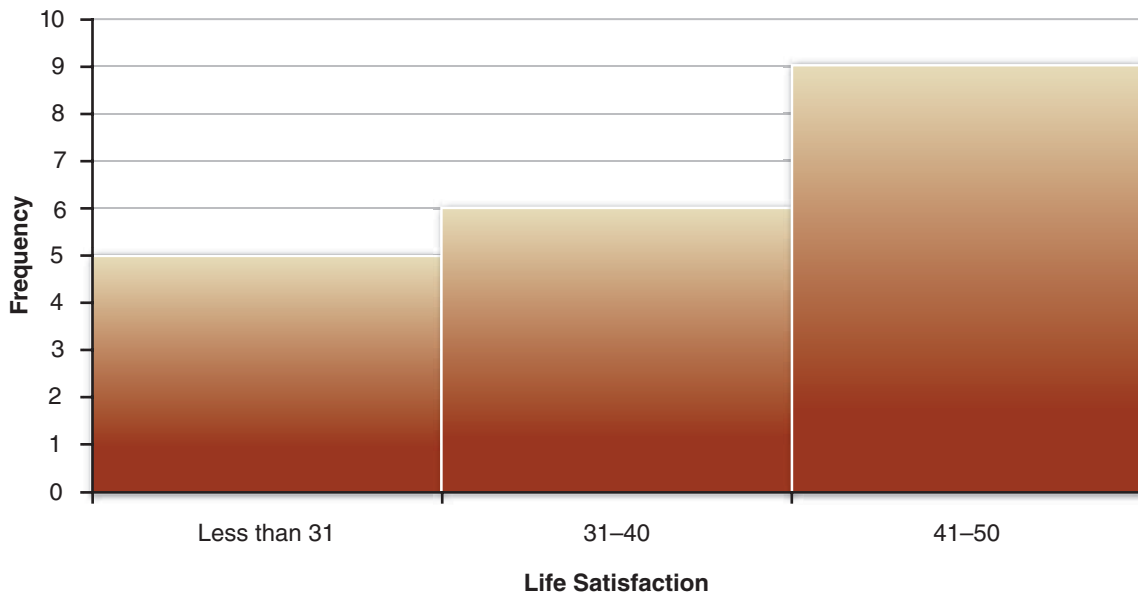
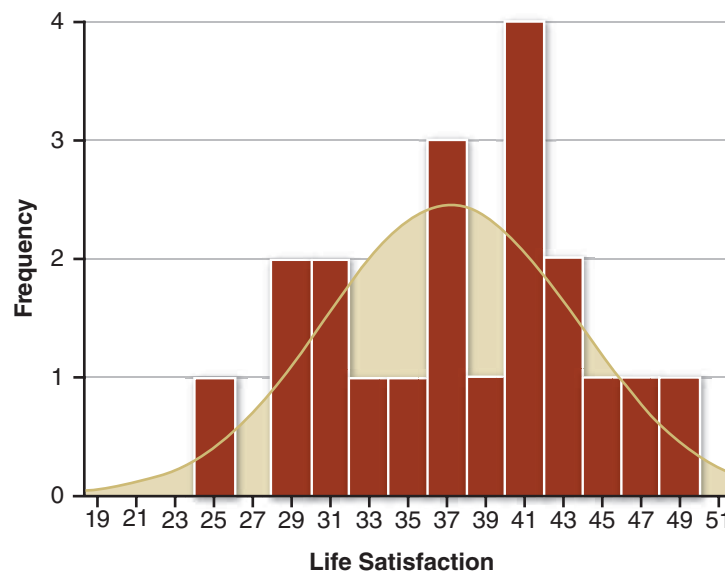


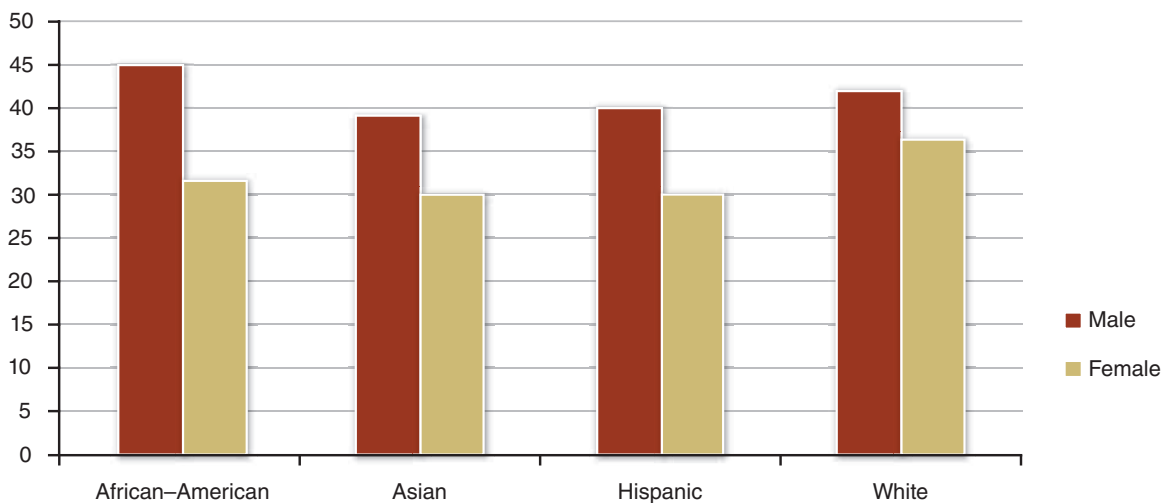
Figure 3.5b: Histogram showing life satisfaction scores on a continuous scale



Displaying Central Tendency

Graphs are also commonly used to display numeric descriptors in an easy-to-understand visual format. Referring back to the sample data in Table 3.4 provides information about ethnicity and gender but also about reports of daily stress and life satisfaction. Thus, a natural question is whether there are gender or ethnic differences in these two variables. Figure 3.6 displays a clustered bar graph showing the mean level of life satisfaction in each group of participants. Of note is that males appear to report more life satisfaction than females, as revealed by the fact that the red bars are always higher than the gold bars. We can also see some variation in satisfaction levels by ethnicity: African-American males (45) appear to report slightly more satisfaction than White males (42).

Figure 3.6: Clustered bar graph displaying life satisfaction scores by gender and ethnicity

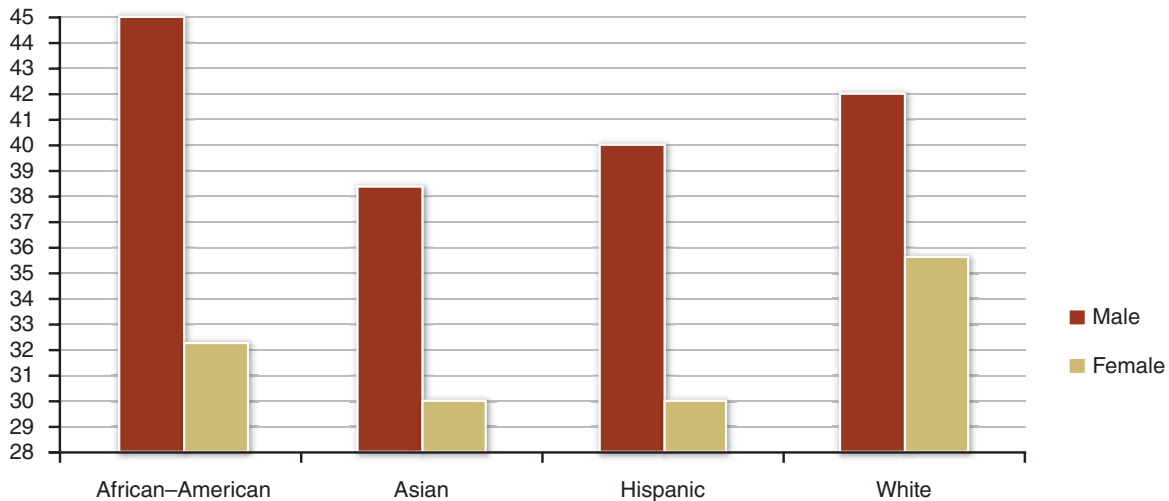


These particular data are fictional, of course, but even if our graph depicted real data, we would want to be cautious in interpreting them. One reason for caution is that the data represent a descriptive study. We might be able to state which demographic groups report more life satisfaction, but we would be unable to determine the reasons for the difference. Another, more important, reason for caution is that visual presentations can be misleading, and we would need to conduct statistical analyses to discover the real patterns of differences.

The best way to appreciate this latter point is to notice what happens when we tweak the graph a little bit. The original graph in Figure 3.6 is a fair representation of the data: The scale starts at zero, and the y-axis on the left side increases by reasonable intervals. However, if we were trying to win an argument about gender differences in happiness, we could always alter the scale, as Figure 3.7 shows. These bars represent the same set of means, but we have compressed the y-axis to show only a small part of the range of the scale. That is, rather than ranging from 0 to 50, this misleading graph ranges from 28 to 45, in increments of 1. To the uncritical eye, the graph appears to show an enormous gender difference in life satisfaction; to the trained eye, it shows an obvious attempt to make the findings seem more interesting. Anytime we encounter a bar graph used to support a particular argument, we must always pay close

attention to the scale of the results: Does it represent the actual range of the data, or is it compressed to exaggerate the difference? Likewise, any time researchers create a graph to display results, they have a responsibility to ensure that the graph is an accurate representation of the data.

Figure 3.7: Clustered bar graph altered to exaggerate the differences



Summary and Resources

Chapter Summary

This chapter has focused on descriptive designs, the first of three specific research designs the text will discuss. As the name implies, the primary goal of descriptive designs is to describe attitudes and behavior, without any pretense of making causal claims. One common feature of all descriptive designs is that they are able to assess behaviors in their natural environment, or at least in something very close to it. The chapter covered three types of descriptive research: case studies, archival research, and observational research. Because each of these methods has the goal of describing attitudes, feelings, and behaviors, each one can be used from either a quantitative or a qualitative perspective.

In a case study, the researcher studies one person in great detail over a period of time. This approach is often used to study special populations and to gather detailed information about rare phenomena. On the one hand, case studies represent the lowest point on the continuum of control because of the lack of a comparison group and the difficulty of generalizing from a single case. On the other hand, case studies are a valuable tool for beginning to study a phenomenon in depth. We discussed the example of Phineas Gage, who suffered severe brain damage and showed drastic changes in his personality and cognitive skills. Although it is difficult to generalize from the specifics of Gage's experience, this case has helped to inspire more than a century's worth of research into the connections among mind, brain, and behavior.

Archival research involves drawing new conclusions by analyzing existing sources of data. This approach is often used to track changes over time or to study things that would be impossible to measure in a laboratory setting. For example, we discussed Phillips's study of copy-cat suicides, which he conducted by matching newspaper coverage of suicides to subsequent spikes in fatality rates. There would be no practical or ethical way to study these connections other than examining the patterns as they occurred naturally. Archival studies are still relatively low on the continuum of control, primarily because the researcher does not have much control over how the data are collected. In many cases, analyzing archives involves a process known as content analysis, or developing a coding strategy to extract relevant information from a broader collection of content. Content analysis involves a three-step process: identifying the most relevant archives, sampling from these archives, and finally coding and recording behaviors. For example, Weigel and colleagues studied race relations on television by sampling a week's worth of prime-time programming and recording the screen time dedicated to portraying interactions between characters of different races.

Lastly, observational research involves directly observing behavior and recording observations in a systematic way. This approach is well suited to a wide variety of research questions, provided the variables can be directly observed. That is, researchers can observe *what* people do but not *why* they do it. In exchange for giving up access to internal processes, researchers gain access to unfiltered behavioral responses—especially when they find ways to observe people unobtrusively. We discussed three main types of observational research. Structured observation involves creating a standardized situation, often in a laboratory setting, and tracking people's responses. Naturalistic observation involves observing behavior as it occurs naturally, often in its natural context. Participant observation involves having the researcher take part in the same activities as the participants in order to gain greater insight into their private behaviors. All three of these variations go through a similar three-step process as archival research: choose a hypothesis, choose a sampling strategy, and then code and record behaviors.

Finally, this chapter discussed principles for describing data in both visual and numeric form. To move toward conducting statistical analyses, summarizing data in numeric form is also useful. We discussed two categories of numeric summaries, central tendency and dispersion. Measures of central tendency (i.e., mean, median, and mode) provide information about the "typical" score in a dataset, while measures of dispersion (i.e., range, variance, and standard deviation) provide information about the distribution of scores around the central tendency—that is, they tell us *how typical* the typical score is. Finally, the chapter described the process of translating scores into standard scores (aka, *z* scores), which express individual scores in terms of standard deviations. This technique is useful for comparing results from different studies and using different measures. The chapter also discussed guidelines for visual presentation. Remember that the sole purpose of visual information is to communicate research findings to an audience. Thus, a researcher's descriptions should always be accurate, concise, and easy to understand. The most common visual displays for summarizing data are bar graphs (for nominal data) and histograms (for quantitative data). Regardless of the choice of visual display, it should represent the data accurately; it is especially important to make sure that the *y*-axis accurately represents the range of data.

Key Terms

archival research A descriptive design that involves drawing conclusions by analyzing existing sources of data, including both public and private records.

bar graph A visual display that summarizes the frequency of data by category; used to display nominal data.

case study A descriptive design that provides a detailed, in-depth analysis of one person over a period of time.

central tendency A numeric descriptor that represents the most typical case in a data set.

clustered bar graph A visual display that summarizes the frequency by two categories at one time; used to display nominal data.

content analysis The process of systematically extracting and analyzing the contents of a collection of information.

deviation score The difference between an individual score and the sample mean, obtained by subtracting each score from the mean.

dispersion A numeric descriptor that represents the spread of scores around the central tendency.

ecological validity The extent to which the research setting resembles conditions in the real world.

ethnography A qualitative methodology involving the scientific study of the customs of people and cultures.

event sampling In observational research, a technique that involves observing and recording behaviors that occur throughout an entire event.

frequency tables Summary tables that present the number and percentage of the sample that fall into each of a set of categories.

histogram A variation of a bar graph used to display ordinal, interval, or ratio data; histograms are drawn with the bars touching one another to indicate that the categories are quantitative.

individual sampling In observational research, a technique that involves collecting data by observing one person at a time in order to test hypotheses about individual behaviors.

mean A measure of central tendency that represents the mathematical average of a data set; calculated by adding all the scores together and then dividing by the number of scores.

median A measure of central tendency that represents the number in the middle of a data set, with 50% of scores both above and below it.

mode A measure of central tendency that represents the most frequent score in a data set, obtained either by visual inspection of the values or by consulting a frequency table.

naturalistic observation A type of observational study that involves observing and systematically recording behavior in the real world; can be done with or without intervention by the researcher.

normal distribution (or “bell curve”) A symmetric distribution with an equal number of scores on either side of the mean; has the same value for mean, median, and mode.

observational research A descriptive design that involves directly observing behavior and recording these observations in an objective and systematic way.

participant observation A type of observational study that involves having the researcher(s) conduct observations while engaging in the same activities as the participants; the goal is to interact with participants to gain access and insight into their behaviors.

participant reactivity The tendency of people to behave differently when they are aware of being observed.

range A measure of dispersion that represents the difference between the highest and lowest scores.

sample A subset, or smaller portion, of the population, members of which are representative of the larger population.

standard scores (or z scores) Scores that represent the distance of each score from the sample mean, expressed in standard deviation units; calculated by subtracting a score from the mean, then dividing by the standard deviation.

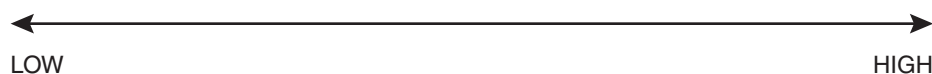
structured observation A type of observational study that involves creating a standard situation in a controlled setting and then observing participants' responses.

time sampling In observational research, a technique that involves comparing behaviors during different time intervals.

variance A measure of dispersion that represents the average difference between the mean and each individual score; calculated by subtracting each score from the mean to obtain a deviation score, squaring and summing these individual deviation scores, and dividing by the sample size.

Apply Your Knowledge

1. Compare and contrast the sets of the following terms. Your answers should demonstrate that you understand each term.
 - a. individual sampling versus event sampling
 - b. participant observation versus naturalistic observation
 - c. mean versus median versus mode
 - d. variance versus standard deviation
 - e. bar graph versus histogram
2. Place each of the three research methods we have discussed in this chapter (listed below) on the continuum of control.



archival research
case study
naturalistic observation

3. For each of the following research methods, list one advantage and one disadvantage.
 - a. archival research

advantage:

disadvantage:
 - b. case studies

advantage:

disadvantage:
 - c. observation studies

advantage:

disadvantage:
4. For each of the following data sets, compute the mean, median, mode, and standard deviation. Once you have determined all three measures of central tendency, decide which one best represents the data.
 - a. 2, 2, 4, 5
 - b. 10, 13, 15, 100
5. Mike scores an 80 on a math test that has a mean of 100 and a standard deviation of 20. Convert Mike's test score into a z score.
6. For each of the following relationships, state the best way to present it graphically (bar graph, clustered bar graph, or histogram).
 - a. average income by years of school completed (ratio scale)
 - b. average income based on category of school completed (high school, some college, college degree, master's degree, and doctoral degree)
 - c. average income based on gender and category of school completed
7. For each of the following questions, state how you would test them using an observational design.
 - a. Are people who own red cars more likely to drive recklessly?
 - (1) What would your hypothesis be?
 - (2) Where would you acquire your sample and how (i.e., which type)?
 - (3) What categories of behavior would you record? How would you define them?
 - b. Are men more likely than women to "lose control" at a party?
 - (1) What would your hypothesis be?
 - (2) Where would you acquire your sample and how (i.e., which type)?
 - (3) What categories of behavior would you record? How would you define them?
 - c. How many fights break out in an average NHL (hockey) game?
 - (1) What would your hypothesis be?
 - (2) Where would you acquire your sample and how (i.e., which type)?
 - (3) What categories of behavior would you record? How would you define them?

Critical Thinking Questions

1. Explain the tradeoffs involved in taking a qualitative versus a quantitative approach to a research question. What are the pros and cons of each one?
2. What are the advantages and disadvantages of conducting participant observation?

Research Scenarios: Try It

What makes a teacher great? This a question that has dogged U.S. educators, parents, and politicians for decades. Common sense suggests the importance of specialized training, experience, and personal characteristics such as energy and passion. But some researchers, such as Bob Pianta, the dean of the University of Virginia's Curry School of Education, are taking a more empirical approach to the question by observing teachers—"the good and the not so good"—in action. (You can read more about this research here: <http://curry.virginia.edu/academics/directory/robert-c-pianta>.)

Imagine that you are an educational psychologist interested in understanding what makes an effective high school science teacher. In past research you have found that certain teachers—whom you call "super-teachers"—consistently inspire high numbers of students to go on to pursue technological careers. You set out to answer the question "What makes these super-teachers so great?"

1. Which of the following would be the best method to begin exploring what qualities make super-teachers so effective?
 - a. A case study of one or two super-teachers, based on classroom observation and qualitative interviews
 - b. An archival analysis of school records
 - c. A structured observation of super-teacher interactions with students
 - d. A controlled experiment evaluating whether a specific trait accounts for the effectiveness of super-teachers
2. Suppose that after completing your initial study, you are struck by the frequency and quality of interactions between super-teachers and their students. These super-teachers are constantly engaging students by giving them individualized feedback during class. You decide to design an observational study around this idea. The first step of observational research is to develop a testable hypothesis. With respect to your initial observations about super-teacher feedback, which of the following would be an appropriate hypothesis to test using observational methods?
 - a. Students perceive super-teachers as being more supportive than other teachers.
 - b. Super-teachers believe in the importance of feedback more than other teachers.
 - c. Super-teachers give more individual feedback in class than other teachers.
 - d. Super-teachers give more feedback in class than other teachers because they care more about students.

3. The second step of observational research involves deciding what and how to sample. You arrange to video record several months of high-school science classes taught by both super-teachers and non-super-teachers. Suppose you decide to code the recordings with regard to whether and how teachers interact directly with individual students. What type of sampling is this?
 - a. Event sampling
 - b. Individual sampling
 - c. Time sampling
 - d. Random sampling

4. The next step is to operationalize your variable. What might be a good way of operationalizing the frequency of teacher feedback in class?
 - a. Number of times the teacher answers a student's question
 - b. Rating the quality of each teacher-student interaction
 - c. Number of times the teacher speaks directly to a particular student
 - d. Number of times the teacher responds to a particular student about class material

5. The approach of coding recorded classes is a type of naturalistic observation. Identify one advantage and one disadvantage of using such methodology to test your hypothesis:
 - a. *advantage*: high ecological validity; *disadvantage*: cannot infer causality
 - b. *advantage*: high construct validity; *disadvantage*: cannot infer causality
 - c. *advantage*: high ecological validity; *disadvantage*: not systematic
 - d. *advantage*: high construct validity; *disadvantage*: not systematic

See Appendix A for answers to Research Scenarios: Try It questions.

4 Survey Designs— Predicting Behavior



Duncan Smith/Photodisc/Thinkstock

Learning Objectives

By the end of this chapter, you should be able to:

- Describe the distinguishing features of survey research.
- Outline best practices for designing questionnaires to ensure quality responses.
- Explain the reasons for sampling for the population.
- Distinguish the different types of sampling strategies.
- Explain the logic behind common approaches to analyzing survey data.

In a highly influential book published in the 1960s, the sociologist Erving Goffman (1963) defined *stigma* as an unusual characteristic that triggers a negative evaluation. In his words, “the stigmatized person is one who is reduced in our minds from a whole and usual person to a tainted, discounted one” (p. 3). People’s beliefs about stigmatized characteristics exist largely in the eye of the beholder, but have substantial influence on social interactions with the stigmatized (see Snyder, Tanke, & Berscheid, 1977). A large research tradition in psychology has been devoted to understanding both the origins of stigma and the consequences of being stigmatized. According to Goffman and others, the characteristics associated with the greatest degree of stigma have three features in common: they are highly visible, they are perceived as controllable, and they are misunderstood by the public.

Recently, researchers have taken considerable interest in people’s attitudes toward members of the gay and lesbian community. Although these attitudes have become more positive over time, this group still encounters harassment and other forms of discrimination on a regular basis (see Almeida, Johnson, Corliss, Molnar, & Azrael, 2009; Berrill, 1990). One of the top recognized experts on this subject is Gregory Herek, professor of psychology at the University of California at Davis (<http://psychology.ucdavis.edu/herек/>). In a 1988 article, Herek conducted a survey of heterosexuals’ attitudes toward both lesbians and gay men, with the goal of understanding the predictors of negative attitudes. Herek approached this research question by constructing a questionnaire to measure people’s attitudes toward these groups. In three studies, participants were asked to complete this attitude measure, along with other existing scales assessing attitudes about gender roles, religion, and traditional ideologies.

Herek’s (1988) research revealed that, as hypothesized, heterosexual males tended to hold more negative attitudes about gay men and lesbians than did heterosexual females. However, the same psychological mechanisms seemed to explain the prejudice in both genders. That is, negative attitudes toward gays and lesbians were associated with increased religiosity, more traditional beliefs about family and gender, and fewer experiences actually interacting with gay men and lesbians. These associations meant that Herek could predict people’s attitudes toward gay men and lesbians based on knowing their views about family, gender, and religion, as well as their past interactions with the stigmatized group. In this paper, Herek’s primary contribution to the literature was the insight that reducing stigma toward gay men and lesbians “may require confronting deeply held, socially reinforced values” (1988, p. 473). This insight was only possible because people were asked to report these values directly.

This chapter continues along the continuum of control, moving on to survey research, in which the primary goal is either describing or predicting attitudes and behavior. For our purposes, **survey research** refers to any method that relies on people’s direct reports of their own attitudes, feelings, and behaviors. So, for example, in Herek’s (1988) study, the participants reported their attitudes toward lesbians and gay men, rather than these attitudes being somehow directly observed by the researchers. Compared to the descriptive designs we discussed in Chapter 3, survey designs tend to have more control over both data collection and question content. Thus, survey research falls somewhere between purely descriptive research (Chapter 3) and the explanatory power of experimental designs (Chapter 5). This chapter provides an overview of survey research from conceptualization through analysis. It will discuss the types of research questions that are best suited to survey research and provide an overview of the decisions to consider in designing and conducting a survey study. We will then cover the process of data collection, with a focus on selecting the people who will complete surveys. Finally, the chapter will describe the three most common approaches for analyzing survey data.

Research: Making an Impact

Kinsey Reports

Alfred Kinsey's research on human sexuality is an example of social research that changed the way society thought about a complex issue—in this case, ideas about “normal” sexual behavior. Kinsey's research, particularly two books on male and female sexuality known together as the Kinsey Reports, illuminated the discrepancies between the assumptions made by a “moral public” and the actual behavior of individuals. His shift in the approach to studying sex—applying scientific methods and reasoning rather than basing conclusions on medical speculation and dogmatic opinions—changed the nature of sex research and the general public's view of sex for decades to come.

Kinsey's major contribution was in challenging the prevailing assumptions about sexual activity in the United States and obtaining descriptive data from both men and women that described their own sexual practices (Bullough, 1998). By collecting actual data instead of relying on speculation, Kinsey made the study of sexuality more scientifically based. The results of his surveys revealed a variety of sexual behaviors that shocked many members of society and redefined the sexual morality of modern America.

Until Kinsey's research, the general, Victorian viewpoint was that women should not show any interest in sex and should submit to their husband without any sign of pleasure (Davis, 1929). Kinsey's data challenged the prevailing assumption that women were asexual. His studies revealed that 25% of the women studied had experienced an orgasm by the age of 15 and more than half by the age of 20 (Kinsey, Pomeroy, Martin, & Gebhard, 1953). Eventually, these results were bundled into the various elements that fueled the women's movement of the 1960s and encouraged further examination of female sexuality (Bullough, 1998).

Kinsey's data also contributed to the budding gay and lesbian liberation movement. Until the Kinsey Reports, studies of human sexuality were based on the assumption that homosexuals were mentally ill (Bullough, 1998). When Kinsey's data revealed that many males and females practiced homosexuality to some degree, he suggested that sexuality was more of a continuum than a series of categories into which people fit. In addition, the Kinsey Reports revealed that the number of extramarital relationships people were having was higher than most expected. Forty percent of married American males reported having an extramarital relationship (Kinsey, et al., 1953).

These ideas, though controversial, prompted society to take a realistic look at the actual sexual practices of its members. The topic of sexuality became less dogmatic as society became more open about sexual activities and preferences.

Kinsey's data not only encouraged social change but also revolutionized the way in which scientists study sexuality. By examining data and studying sex from an unbiased standpoint, Kinsey successfully transformed the study of human sexuality into a science. His research not only changed our way of studying sexual behavior but also allowed society to become less restrictive in its expectations of “normal” sexual behavior.

Think About It

1. What type of data formed the basis of Kinsey's reports? What are the pros and cons of this type?
2. How did applying the scientific method change the national conversation about sexuality?

4.1 Introduction to Survey Research

Whether you are aware of it or not, most people encounter survey research throughout most of their lives. Every time we decide to answer that call from an unknown number, and the person on the other end of the line insists on knowing the call recipient's household income and favorite brand of laundry detergent, we are helping to conduct survey research. When news programs try to predict the winner of an election two weeks early, these reports are based on survey research of eligible voters. In both cases, the researcher is trying to make predictions about the products people buy or the candidates they will elect based on what people say about their own attitudes, feelings, and behaviors.



shironosov/iStock/Thinkstock

Surveys are used to describe or predict attitudes and behavior.

Surveys can be used in a variety of contexts and are most appropriate for questions that involve people describing their attitudes, their behaviors, or a combination of the two. For example, if we want to examine the predictors of attitudes toward the death penalty, we could ask people their opinions on this topic and also ask them about their political party affiliation. Based on these responses, we could test whether political affiliation predicted attitudes toward the death penalty. Or, imagine we want to know whether students who spend more time studying are more likely to do well on their exams.

This question could be answered using

a survey that asked students about their study habits and then tracked their exam grades. We will return to this example near the end of the chapter, as we discuss the process of analyzing survey data to test our hypotheses about predictions.

The common thread of these two examples is that they require people to report either their thoughts (e.g., opinions about the death penalty) or their behaviors (e.g., the hours they spend studying). Contrast these with an example that might be a poor fit for survey research: If a researcher wanted to test whether a new drug led to increased risk of developing blood clots, it would be much safer to test for these clots using medical technology, rather than asking people for their beliefs ("on a scale from 1 to 5, how many clots have you developed this week?"). Thus, when deciding whether a survey is the best fit for a research question, a researcher must consider whether people will be both able and willing to report the opinions or behaviors accurately. The next section expands on both of these issues.

Distinguishing Features of Surveys

Survey research designs have three distinguishing features that set them apart from other designs. First, all survey research relies on either written or verbal **self-reports** of people's attitudes, feelings, and behaviors. This self-reporting means that researchers will ask

participants a series of questions and record their responses. The approach has several advantages, including being relatively straightforward and allowing a degree of access to psychological processes (e.g., “Why do you support candidate X?”). However, researchers should also be also cautious in their interpretation of self-report data because participants’ responses often reflect a combination of their true attitude and concern over how this attitude will be perceived. Scientists refer to this concern as **social desirability**, which means that people may be reluctant to report unpopular attitudes. For example, if we were to ask people their attitudes about different racial groups, their answers might reflect both their true attitude and their desire not to appear racist. We return to the issue of social desirability later in this chapter and discuss some tactics for designing questions that can help to sidestep these concerns and capture respondents’ true attitudes.

The second distinguishing feature of survey research is its ability to access internal states that cannot be measured through direct observation. The discussion of observational designs in Chapter 3 explained that one limitation of these designs was a lack of insight into *why* people behave the way they do. Survey research can address this limitation directly: By asking people what they think, how they feel, and why they behave in certain ways, researchers come closer to capturing the underlying psychological processes. However, people’s reports of their internal states should be taken with a grain of salt, for two reasons. First, as mentioned, these reports may be biased by social-desirability concerns, particularly when unpopular attitudes are involved. Second, a large body of literature in social psychology suggests that people may not understand the true reasons for their behavior. In an influential review paper, psychologists Richard Nisbett and Tim Wilson (1977) argued that we make poor guesses after the fact about why we do things, based more on our assumptions than on any real introspection. Thus, survey questions can provide access to internal states, but researchers should always interpret responses with caution.

Third, on a more practical note, survey research allows us to collect large amounts of data with relatively little effort and few resources. Many of the descriptive designs Chapter 3 discussed require observing one person at a time, and the same will hold true when Chapter 5 explores experimental designs. Survey-research designs stand out as the most efficient, because surveys can be distributed to large groups of people simultaneously. Still, their actual efficiency depends on the decisions researchers make during the design process. In reality, efficiency is often in a delicate balance with the accuracy and completeness of the data.

Broadly speaking, survey research can be conducted using either verbal or written self-reports (or a combination of the two). Before diving into the details of writing and formatting a survey, we need to understand the pros and cons of administering a survey as an interview (i.e., a verbal survey) or a questionnaire (i.e., a written survey).

Interviews

An **interview** involves a verbal question-and-answer exchange between the researcher and the participant. This verbal exchange can take place either face-to-face or over the phone. So, our earlier telemarketer example represents an interview because the questions are asked verbally via phone. Likewise, if we are approached in a shopping mall and asked to answer questions about our favorite products, we experience a survey in interview form because

the questions are administered verbally face-to-face. And, if a person has ever participated in a focus group, during which a group of people gives their reactions to a new product, the researchers are essentially conducting an interview with the group.

Interview Schedules

Regardless of how the interview is administered, the interviewer (i.e., the researcher) has a predetermined plan, or script, for how the interview should go. This plan, or script, for the progress of the interview is known as an **interview schedule**. When conducting an interview—including those telemarketing calls—the researcher/interviewer has a detailed plan for the order of questions to be asked, along with follow-up questions that depend on the participant's responses.

Broadly speaking, researchers employ two types of interview schedules. A **linear** (also called “structured”) **schedule** will ask the same questions, in the same order, for all participants. In contrast, a **branching schedule** unfolds more like a flowchart, with the next question dependent on participants' answers. Interviewers typically use a branching schedule in cases with follow-up questions that only make sense for some of the participants. For example, a researcher might first ask people whether they have children; if they answer “yes,” the interviewer might then follow up by asking how many.

One danger in using a branching schedule is that it is based partly on the researcher's assumptions about the relationships between variables. Granted, to ask only people with children to indicate how many they have is fairly uncontroversial. Imagine the following scenario, however. Say we first ask participants for their household income, and then ask about their political donations:

- “How much money do you make? \$18,000? OK, how likely are you to donate to the Democratic Party?”
- “How much money do you make? \$250,000? OK, how likely are you to donate money to the Republican Party?”

The way these questions branch implicitly assumes that wealthier people are more likely to be Republicans, and less wealthy people are more likely to be Democrats. The data might support this assumption or they might not. By planning the follow-up questions in this way, though, we are unable to capture cases that do not fit our stereotypes (i.e., the wealthy Democrats and the poor Republicans). Researchers must therefore be careful about letting their biases shape the data-collection process.

Advantages and Disadvantages of Interviews

Interviews offer a number of advantages over written surveys. For one, people are often more motivated to talk than they are to write. Consider the example of an actual undergraduate research assistant who was dispatched to a local shopping mall to interview people about their experiences in romantic relationships. He had no trouble at all recruiting participants, many of whom would go on and on (and on, and on) about recent relationships—one woman

even confided to him that she had just left an abusive spouse earlier that week. For better or for worse, these experiences would have been more difficult to capture in writing.

Related to this benefit, people's oral responses are typically richer and more detailed than their written responses. Think of the difference between asking someone to "describe your views on gun control" and asking someone to "indicate on a scale of 1 to 7 the degree to which you support gun control." The former is more likely to capture the richness and subtlety involved in people's attitudes about guns. On a practical note, an interview format also allows the researcher to ensure that respondents understand the questions. Poorly worded written-questionnaire items force survey participants to guess at the researcher's meaning, and these guesses introduce a large source of error variance. On the other hand, if an interview question is poorly asked, people can easily ask the interviewer to clarify. Finally, using an interview format allows researchers to reach a broader cross-section of people and to include those who are unable to read and write—or, perhaps, unable to read and write the language of the survey.

Interviews also have two clear disadvantages compared to written surveys. First, interviews cost more in terms of both time and money. It took more time for the graduate assistant to go to a shopping mall than it would have taken to mail out packets of surveys (but no more money—research-assistant positions tend to be unpaid). Second, the interview format allows many opportunities for interviewers to pass on their personal biases. These biases are unlikely to be deliberate, but participants can often pick up on body language and subtle facial expressions when the interviewer disagrees with their answers. Such cues may influence them to shape their responses to make the interviewer happier. The best way to understand the pros and cons of interviewing is to recognize that both are a consequence of personal interaction. The interaction between interviewer and interviewee allows for richer responses but also the potential for these responses to be biased. Researchers must weigh these pros and cons and decide which method is the best fit for their survey. The next section turns to the process of administering surveys in writing.

One additional problem with interviews is the increasing difficulty of obtaining representative samples for interviews over the telephone due to low or declining use of landline phones, coupled with the use of unlisted numbers and call-screening devices. In the United States, the Pew Research Center (2012) reports that overall response rate—a ratio of completed interviews to the number of phone numbers dialed—was just 9% in 2012, one-fourth of the 36% level from 1997. Thus, significant differences may exist between people who elect to respond to phone surveys and those who do not.

Questionnaires

A **questionnaire** is a survey that involves a written question-and-answer exchange between the researcher and the participant. The exchange is a bit different from interview formats—in this case, the questions are designed ahead of time, then distributed to participants, who write their responses and return the questionnaire to the researcher. The next section discusses details for designing these questions. First, however, we will take a quick look at the process of administering written surveys.

Distribution Methods

Questionnaires can be distributed in three primary ways, each with its own pattern of advantages and disadvantages:

Distributing by mail: Until recently, researchers commonly distributed surveys by sending paper copies through the mail to a group of participants (see the section on “Sampling” for more discussion on how this group is selected). Mailing surveys is relatively cheap and relatively easy to do, but it is unfortunately one of the worst methods in terms of response rates. People tend to ignore questionnaires that they receive in the mail, dismissing them as one more piece of junk. Researchers have a few methods available for increasing response rates, including providing incentives, making the survey interesting, and making it as easy as possible to return the results (e.g., with a postage-paid envelope). However, even using all of these tactics, researchers consider themselves extremely lucky to obtain a 30% response rate from a mail survey. That means a researcher who mails 1,000 surveys will be doing well to receive 300 back. More typical response rates for mail surveys can be in the single digits. Because of this low return on investment, researchers have begun relying on other methods for their written surveys.

Distributing in person: Another option for researchers is to distribute a written survey in person, simply handing out copies and asking participants to fill them out on the spot. This method is certainly more time-consuming; a researcher has to be stationed for long periods of time to collect data. In addition, people are less likely to answer the questions honestly because the presence of a researcher makes them worry about social desirability. Last, the sample for this method is limited to people who are in the physical area at the time that questionnaires are being distributed. As the chapter discusses later, this limitation might lead to problems in the composition of the sample. On the plus side, however, this method tends to result in higher compliance rates because people find it harder to say no to someone face-to-face than to ignore a piece of mail.

Distributing online: During the last two decades, online surveys have become the dominant method of data collection, for both market research and academic research. Online distribution involves posting a questionnaire on a web page, and then directing participants to this web page to complete the questionnaire. Online surveys offer many benefits over other forms of data collection, including the ability to: present audio and visual stimuli, randomize the order of questions, and implement complex branching logic (e.g., asking people to evaluate local grocery stores depending on where they live).

Most recently, researchers have begun exploring the best ways to design surveys for mobile devices. According to a report from the International Telecommunications Union, in 2013, 6.8 billion mobile phones were in use, compared to a world population of 7.2 billion. In 2012, 44% of Americans slept next to



AndreyPopov/iStock/Thinkstock

Approximately 20–30% of online surveys are completed on a mobile device.

their phones (Pew Research Center, 2012). Not surprisingly, consensus in the market research industry is that approximately 20–30% of *online* surveys are actually completed on a *mobile* device (Poynter, Williams, & York, 2014). Why does this matter? People take surveys on their smartphones because it is convenient (or, in some cases, because it is their only Internet device). However, despite recent exponential advancement, mobile phones still have smaller screens, less functional keyboards, and less predictability in displaying images and videos. (Imagine someone being asked to view a set of two-minute-long advertisements on an iPhone while trying to complete a survey before a doctor's appointment.) Researchers do have ways to make this experience more pleasant for respondents and consequently to increase the quality of data obtained. For example, mobile surveys work best when they are shorter overall, when the question text is short and straightforward, and when response scales (discussed below) are kept at five points (see Poynter et al., 2014, for a review). The latter point is a direct result of small screen size: Longer response scales require respondents to scroll back and forth on their screens to see the entire scale. Unfortunately, but understandably, some applied research suggests that people tend to ignore the scale points that they cannot see—perhaps using only four points out of a ten-point scale.

Because these methods are relatively new, the jury is still out on whether online and mobile distribution results in biased samples or biased responses. However, worth keeping in mind is that approximately 13% of the U.S. population does not have Internet access (Internet Users by Country, 2014). This group is disproportionately older (65+) and represents the lowest income and least educated segments of the population. Thus, if research questions involve reaching these groups, it is necessary to supplement online surveys with other distribution methods. For readers interested in more information on designing and conducting Internet research, Sam Gosling and John Johnson's (2010) recent book provides an excellent resource. In addition, several groups of psychological researchers have been attempting to understand the psychology of Internet users (read about recent studies on this website: <http://www.spring.org.uk/2010/10/internet-psychology.php>).

Advantages and Disadvantages of Questionnaires

Just as interview methods do, written questionnaires claim their own set of advantages and disadvantages. Written surveys allow researchers to collect large amounts of data with little cost or effort, and they can offer a greater degree of anonymity than interviews. Anonymity can be a particular advantage in dealing with sensitive or potentially embarrassing topics. That is, people may be more willing to answer a questionnaire about their alcohol use or their sexual history than they would be to discuss these things face-to-face with an interviewer. On the downside, written surveys miss out on one advantage of interviews because no one is available to clarify confusing questions. Fortunately, researchers have one relatively easy way of minimizing this problem: make survey questions as clear as possible. The next section explains the process of questionnaire design.

4.2 Questionnaire Design

One of the most important steps in conducting survey research is deciding how to construct and assemble the questionnaire items. In some cases, a researcher will be able to answer research questions using questionnaires that other researchers have already developed. For

example, quite a bit of psychology research uses standard scales that measure self-esteem, prejudice, depression, or stress levels. The advantage of these ready-made measures is that other people have already gone to the trouble of making sure they are valid and reliable. So, someone interested in the relationship between stress and depression could distribute the Perceived Stress Scale (Cohen, Kamarck, & Mermelstein, 1983) and the Beck Depression Inventory (Beck, Steer, Ball, & Ranieri, 1996) to a group of participants and more quickly move along on to the fun part of data analyses.

However, in many cases, no perfect measure exists for a research question—either because no one has studied the topic before or because the current measures are all flawed in some way. When this happens, researchers need to go through the process of designing their own questions. This section discusses strategies for writing questions and choosing the most appropriate response format.

Five Rules for Better Questions

Each of the rules listed below is designed to make research questions as clear and easy to understand as possible so as to minimize the potential for error variance. We discuss each rule below and illustrate it with contrasting pairs of items: “bad” items that do not follow the rule and “better” items that do.

1. **Use simple language.** One of the simplest and most important rules to keep in mind is that people have to be able to understand the survey questions. This means avoiding jargon and specialized language whenever possible.

BAD: “Have you ever had an STD?”

BETTER: “Have you ever had a sexually transmitted disease?”

BAD: “What is your opinion of the S-CHIP program?”

BETTER: “What is your opinion of the State Children’s Health Insurance Program?”

It is also a good idea to simplify the language as much as possible, so that people spend time answering the question rather than trying to decode its meaning. For example, words like *assist* and *consider* can be replaced with simpler words like *help* and *think*. This may seem odd—or perhaps even condescending to participants—but it is always better to err on the side of simplicity. Remember, when people are forced to guess at the meaning of questions, these guesses add error variance to their answers.

2. **Be precise.** Another way to ensure that people understand the question is to be as precise as possible with wording. Ambiguously (or vaguely) worded questions will introduce an extra source of error variance into the data because people may interpret these questions in varying ways.

BAD: “What drugs do you take?” (Legal drugs? Illegal drugs? Now? In college?)

BETTER: “What prescription drugs are you currently taking?”

BAD: “Do you like sports?” (Playing? Watching? Which sports??)

BETTER: “How much do you enjoy watching basketball on television?”

3. **Use neutral language.** Questions should be designed to *measure* participants’ attitudes, feelings, or behaviors rather than to *manipulate* these things. That is, avoid **leading questions** that are written in such a way that they suggest an answer.

BAD: “Do you beat your children?” (Who would say yes?)

BETTER: “Is it acceptable to use physical forms of discipline?”

BAD: “Do you agree that the president is an idiot?”

BETTER: “How would you rate the president’s job performance?”

This guideline can be used to sidestep social desirability concerns. If the researcher suspects that people may be reluctant to report holding an attitude—for example, using corporal punishment with their children—it helps to phrase the question in a nonthreatening way: “using physical forms of discipline” versus “beating your children.” Many current measures of prejudice adopt this technique. For example, McConahay’s (1986) “modern racism” scale contains items such as “Discrimination against Blacks is no longer a problem in the United States.” People who hold prejudicial attitudes are more likely to confess agreement with statements like this one than with blunter ones, like “I hate people from Group X.”

4. **Ask one question at a time.** One remarkably common error that people make in designing questions is to include a **double-barreled question** (one which asks more than one question at a time). A new-patient questionnaire at a doctor’s office often asks whether the patient suffer from “headaches and nausea.” What if an individual only suffers from one of these or has a lot of nausea and an occasional headache? The better approach is to ask about each of these symptoms separately.

BAD: “Do you suffer from pain and numbness?”

BETTER: “How often do you suffer from pain?” “How often do you suffer from numbness?”

BAD: “Do you like watching football and boxing?”

BETTER: “How much do you enjoy watching football?” “How much do you enjoy watching boxing?”

5. **Avoid negations.** One final and simple way to clarify questions is to avoid questions with negative statements because these can often be difficult to understand. The first example below may be a little silly, but the second comes from a real survey of voter opinion.

BAD: “Do you never not cheat on your exams?” (Wait, what? Do I cheat? Do I not cheat? What is this asking?)

BETTER: “Have you ever cheated on an exam?”

BAD: “Are you against rejecting the ban on pesticides?” (Wait, so, am I for the ban? Against the ban? What is this asking?)

BETTER: “Do you support the current ban on pesticides?”

Participant-Response Options

This section discusses the issue of deciding how participants should respond to survey questions. The decisions researchers make at this stage will affect the type of data they ultimately collect, so it is important to choose carefully. This section reviews the primary decisions a researcher will need to make about response options, as well as the pros and cons of each one.

One of the first choices to make is whether to collect open-ended or fixed-format responses. As the names imply, **fixed-format responses** require participants to choose from a list of options (e.g., “Choose your favorite color”), while **open-ended responses** ask participants to provide unstructured responses to a question or statement (e.g., “How do you feel about legalizing marijuana?”). Open-ended responses tend to be richer and more flexible but harder to translate into quantifiable data—analogue to the tradeoff we discussed in comparing written versus oral survey methods. To put it another way, some concepts are difficult to reduce to a seven-point fixed-format scale, but number ratings on these scales are easier to analyze than a paragraph of free-flowing text.

Another reason to think carefully about this decision is that fixed-format responses will, by definition, restrict people’s options in answering the question. In some cases, these restrictions can even act as leading questions. In a study of people’s perceptions of history, Dario Páez Rovira and his colleagues (Rovira, Deschamps, & Pennebaker, 2006) asked respondents to indicate the “most significant event over the last 50 years.” When this was asked in an open-ended way (i.e., “list the most significant event”), 2% of participants listed the invention of computers. Another version of the survey asked the question using a fixed-format way (i.e., “choose



Eduard Lysenko/iStock/Thinkstock

Thirty percent of participants selected the invention of computers as the most significant event of the past 50 years when presented with fixed-format responses, but when a different group was asked the same question in an open-ended format, only 20% listed the invention of computers.

the most significant event”). When asked to select from a list of four options (World War II, invention of computers, Tiananmen Square, or man on the moon), 30% chose the invention of computers. In exchange for having easily coded data, the researchers accidentally forced participants into a smaller number of options. The result, in this case, was a distorted sense of the importance of computers in people’s perceptions of history.

Fixed-Format Options

Although fixed-format responses can sometimes constrain or skew participants’ answers, researchers tend to use them more often than not. This decision is largely practical; fixed-format responses allow for more efficient data collection from a much larger sample. (Imagine the chore of having to hand-code 2,000 essays.) But once researchers have decided on this option for the questionnaire, the decision process is far from over. In this section, we discuss three possibilities for constructing a fixed-format response scale.

True/false. One fixed-format option asks questions using a true/false format, which asks participants to indicate whether they endorse a statement. For example:

“I attended church last Sunday.”	True	False
“I am a U.S. citizen.”	True	False
“I am in favor of abortion.”	True	False

This last example may strike you as odd, and in fact it illustrates an important limitation in the use of true/false formats: They are best used for statements of facts rather than attitudes. It is relatively straightforward to answer whether we attended church or are a U.S. citizen. However, people’s attitudes toward abortion are often complicated—one might be “pro-choice” but still support some restrictions, or “pro-life” but support exceptions (e.g., in cases of rape). For most people, a true/false question cannot even come close to capturing the complexity of these beliefs. However, for survey items that involve simple statements of fact, the true/false format can be a good option.

Multiple choice. A second option uses a multiple-choice format, which asks participants to select from a set of predetermined responses.

“Which of the following is your favorite fast-food restaurant?”

- a) McDonald’s
- b) Burger King
- c) Wendy’s
- d) Taco Bell

“Whom did you vote for in the 2012 presidential election?”

- a) Mitt Romney
- b) Barack Obama

“How do you travel to work most days? (Select all that apply.)”

- a) drive alone
- b) carpool
- c) public transportation

As these examples show, multiple-choice questions offer quite a bit of freedom in both the content and the response-scaling of questions. A researcher can ask participants either to select one answer or, as in the last example, to select all applicable answers. A survey can cover everything from preferences (e.g., favorite fast-food restaurant) to behaviors (e.g., how people travel to work).

Multiple-choice formats do have a downside. Whenever the survey provides a set of responses, it restricts participants' responses to that set. This is the problem that Rovira and colleagues (2006) encountered in asking people about the most significant events of the last century. In each of the examples above, the categories fail to capture all possible responses. What if someone's favorite restaurant is In-and-Out Burger? What if a respondent voted for Ralph Nader? What if a person telecommutes or bicycles to work? Researchers have two relatively easy ways to avoid (or at least minimize) this problem. First, when choosing the response options, plan carefully. During the design process, it helps to brainstorm with other people to ensure the survey is capturing the most likely range of responses. However, it is often impossible to provide every option that people might conceive. The second solution is to provide an “other” response to a multiple-choice question, which allows people to write in an option that the survey neglected to include. For example, our last question about traveling to work could be rewritten as:

“How do you travel to work on most days? (Select all that apply.)”

- a) drive alone
- b) carpool
- c) public transportation
- d) other (please specify): _____

This way, people who telecommute, or bicycle, or even ride their trained pony to work will have a way to respond rather than skipping the question. And, if researchers start to notice a pattern in these write-in responses (e.g., 20% of people added “bicycle”), then they have valuable knowledge to improve the next incarnation of the survey.

Rating scales. Last, but certainly not least, another option uses a rating-scale format, which asks participants to respond on a scale representing a continuum.

“Sometimes it is necessary to sacrifice liberty in the name of security.”

1	2	3	4	5
not at all necessary			very necessary	

“I would vote for a candidate who supported the death penalty.”

1	2	3	4	5
not at all likely			very likely	

“The political party in power right now has really messed things up.”

1	2	3	4	5
strongly disagree			strongly agree	

This format is well suited to capturing attitudes and opinions, and, indeed, is one of the most common approaches to attitude research. Rating scales are easy to score, and they give participants some flexibility in indicating their agreement with or endorsement of the questions. Researchers have two critical decisions to make about the construction of rating-scale items; both have implications for how they analyze and interpret results.

First, a researcher needs to decide the **anchors**, or labels, for the response scale. Rating scales offer a good deal of flexibility in these anchors, as the examples above demonstrate. A survey can frame questions in terms of “agreement” with a statement or “likelihood” of a behavior, or researchers can customize the anchors to match their questions (e.g., “not at all necessary”). Scales that use anchors of “strongly agree” and “strongly disagree” are also referred to as **Likert scales**. At a fairly simple level, the choice of labels affects the interpretation of the results. For example, if we asked the “political party in power” question above, we have to be aware that the anchors are phrased in terms of agreement with the statement. In discussing these results, we would be able to discuss how much people agreed with the statement, on average, and whether agreement correlated with other things. If this seems like an obvious point, readers would be amazed how often researchers (or the media) will take an item like this and spin the results to talk about the “likelihood of voting” for the party in power—confusing an attitude with a behavior. So, in short, researchers must make sure they are being honest when presenting and interpreting research data.

At a more conceptual level, a researcher needs to decide whether the anchors for the rating scale make use of a **bipolar scale**, which has polar opposites at its endpoints, or a **unipolar scale**, which assesses a single construct. The difference between these options is best illustrated by an example:

Bipolar: How would you rate your current mood?

Sad—————Happy

Unipolar: How would you rate your current mood?

1	2	3	4	5	6	7
not at all sad					very sad	

1	2	3	4	5	6	7
not at all happy					very happy	

The bipolar option requires participants to place themselves on a continuous scale somewhere between “sad” and “happy,” which are polar opposites. The bipolar scale assumes that the endpoints represent the only two options; participants can be sad, happy, or somewhere in between. In contrast, the unipolar option asks participants to rate themselves on two scales, indicating their level of both “sadness” *and* “happiness.” A pair of unipolar scales

assumes that it is possible to experience varying degrees of each item—participants can be moderately happy, but also a little bit sad, for example. The decision to use a bipolar or a unipolar scale comes down to the context. What is the most logical way to think about these constructs? What have previous researchers done?

In the 1970s, Sandra Lipsitz Bem revolutionized the way researchers thought about gender roles by arguing against a bipolar approach. Previously, gender role identification had been measured on a bipolar scale from “masculine” to “feminine”; the scale assumed that a person could be one or the other. Bem (1974) argued instead that people could easily have varying degrees of masculine and feminine traits. Her scale, the Bem Sex Role Inventory, asks respondents to rate themselves on a set of 60 unipolar traits. Someone with mostly feminine and hardly any masculine traits would be described as “feminine.” Someone with high ratings on both masculine and feminine traits would be described as “androgynous.” And, someone with low ratings on both masculine and feminine traits would be described as “undifferentiated.” View and complete Bem’s scale online at: <http://garote.bdmonkeys.net/bsri.html>.



iulianvalentin/iStock/Thinkstock

Sandra Lipsitz Bem insisted that people have varying degrees of masculine and feminine traits.

After settling on the best way to anchor the scale, the researcher’s second critical decision is to decide on the number of points in the response scale. Notice that all of the examples in this section have an odd number of points (i.e., five or seven). Odd numbers are usually preferable for rating-scale items because the middle of the scale (i.e., “3” or “4”) allows respondents to give a neutral, middle-of-the-road answer. That is, on a scale from “strongly disagree” to “strongly agree,” the midpoint can be used to indicate “neither” or “I’m not sure.” However, in some cases, a researcher may not want to allow a neutral option in a scale. Using an even number of points (e.g., four or six) essentially compels people either to agree or disagree with the statement; this type of scaling is referred to as **forced choice**.

So, how many points should the scale have? As a general rule, more points will translate into more variability in responses—the more choice people have (up to a point), the more likely they are to distribute their responses among those choices. From a researcher’s perspective, the big question is whether this variability is meaningful. For example, if we assess college students’ attitudes about a student-fee increase, student opinions will likely vary depending on the size of the fee and the ways in which it will be used. Thus, we might prefer a five- or seven-point scale to a two-point (yes or no) scale. However, past a certain point, increases in the scale range cease to connect to meaningful variation in attitudes. In other words, the difference between a 5 and a 6 on a seven-point scale is fairly intuitive for participants to grasp. What is the real difference, though, between an 80 and an 81 on a 100-point scale? When scales become too large, researchers risk introducing another source of error variance as participants impose their own interpretations on the scaling. In sum, more points do not always translate to a better scale.

Back to the question: How many points should the scale have? The ideal compromise supported by most statisticians is to use a seven-point scale whenever possible because of the differences between scales of measurement. As the discussion in Chapter 2 explained, the way variables are measured has implications for data analyses. For the most popular statistical tests to be legitimate, variables need to be on an interval scale (i.e., with equal intervals between points) or a ratio scale (i.e., with a true zero point). Based on mathematical modeling research, statisticians have concluded that the variability generated by a seven-point scale is most likely to mimic an interval scale (e.g., Nunnally, 1978). So, from a statistical perspective, a seven-point scale is ideal because it allows us the most flexibility in data analyses.

Finalizing the Questionnaire

After constructing the questionnaire items, researchers face one last important step before beginning data collection. This section discusses a few guidelines for assembling the items into a coherent questionnaire. One main goal at this stage is to think carefully about the order of the individual items.

First, keep in mind that the first few questions will set the tone for the rest of the questionnaire. It is best to start with questions that are both interesting and nonthreatening to help ensure that respondents complete the questionnaire with open minds. For example:

BAD OPENING: “Do you agree that your child’s teacher is an idiot?” (threatening, and also a leading question)

BETTER OPENING: “How would you rate the performance of your child’s teacher?”

BAD OPENING: “Would you support a 1% sales tax increase?” (boring)

BETTER OPENING: “How do you feel about raising taxes to help fund education?”

Second, strive whenever possible to have continuity in the different sections of the questionnaire. Imagine constructing a survey to give to college freshmen. It might include questions on family background, stress levels, future plans, campus engagement, and so on. The survey will be most effective if it groups questions by topic. So, for instance, students respond to a set of questions about future plans on one page and then answer a set of questions about campus engagement on another page. This approach makes it easier for participants to progress through the questions without having to switch mentally between topics.

Third, remember that individual questions are always read in context. This means that if the college-student survey begins with a question about plans for the future and then asks about stress, respondents will likely have their future plans in mind when they think about their stress level. Consider again the example of the graduate assistant. His department used to administer a gigantic survey packet (on paper) to the 2,000 students enrolled in Introductory Psychology each semester. One year, a faculty member included a measure of identity, asking participants to complete the statements “I am _____” and “I am not _____.” As researchers started to analyze data from this survey, they discovered an astonishing 60% of students had filled in the blank with “I am not a homosexual!” This response seemed downright strange, until the surveyors realized that the questionnaire immediately preceding the identity one measured prejudice toward gay and lesbian individuals. So, as these students completed the

identity measure, they had homosexuality on their minds and felt compelled to point out that they were not homosexual. In other words, responses are all about context.

Finally, after assembling a draft version of the questionnaire, perform a test run. This test run, called **pilot testing**, involves giving the questionnaire to a small sample of people, getting their feedback, and making any necessary changes. One of the best ways to pilot test is to find a patient group of friends to complete the questionnaire who will provide extensive feedback. In soliciting their feedback, ask questions like the following:

Was anything confusing or unclear?

Was anything offensive or threatening?

How long did the questionnaire take you to complete?

Did it seem repetitive or boring? Did it seem too long?

Were there particular questions that you liked or disliked? Why?

The answers to these questions will supply valuable information to revise and clarify the questionnaire before devoting resources to a full round of data collection. The next section turns to the question of how to find and select participants for this stage of the research.

Research: Thinking Critically

Beauty and First Impressions

Follow the link below to a press release from the University of British Columbia, describing a recent publication by researchers in the psychology department. This study suggests that physical beauty may play a role in how easily we form first impressions of other people. As you read the article, consider what you have learned so far about the research process, and then respond to the questions below.

<http://news.ubc.ca/2010/12/21/beautiful-people-convey-personality-traits-better-during-first-impressions/>

Think About It:

1. Suppose the following questions were part of the questionnaire given after the three-minute one-on-one conversations in this study. Based on the goals of the study and the rules discussed in this chapter, identify the problem with each of the following questions and suggest a better item.

a) Jane is very neat.

1 2 3 4 5

strongly disagree

strongly agree

main problem:

better item:

b) Jane is generous and organized.

1 2 3 4 5
strongly disagree strongly agree

main problem:

better item:

c) Jane is extremely attractive. TRUE FALSE

main problem:

better item:

2. What are the strengths and weaknesses of using a fixed-format questionnaire in this study versus open-ended responses?
3. The researchers state that they took steps to control for the “positive bias that can occur in self-reporting.” How might social desirability influence the outcome of this particular study? What might the researchers have done to reduce the effect of social desirability?

4.3 Sampling From the Population

At this point, the chapter should have conveyed an understanding of how to construct survey items. The next step is to find a group of people to fill out the survey. But where does a researcher find this group? And how many people are needed? On the one hand, researchers want as many people as possible to capture the full range of attitudes and experiences. On the other hand, they have to conserve time and other resources, which often means choosing a smaller sample of people. This section examines the strategies researchers can use to select samples for their studies.

Researchers refer to the entire collection of people who could possibly be relevant to a study as the **population**. For example, if we were interested in the effects of prison overcrowding, our population would consist of prisoners in the United States. If we wanted to study voting behavior in the next presidential election, the population would be U.S. residents eligible to vote. And if we wanted to know how well college students cope with the transition from high school, our population would include every college student enrolled in every college in the country.

These populations suggest an obvious practical complication. How can we get every college student—much less every prisoner—in the country to fill out our questionnaire? We cannot; instead, researchers will collect data from a sample, a subset of the population. Instead of trying to reach all prisoners, we might sample inmates from a handful of state prisons. Rather than attempt to survey all college students in the country, researchers often restrict their studies to a collection of students at one university.

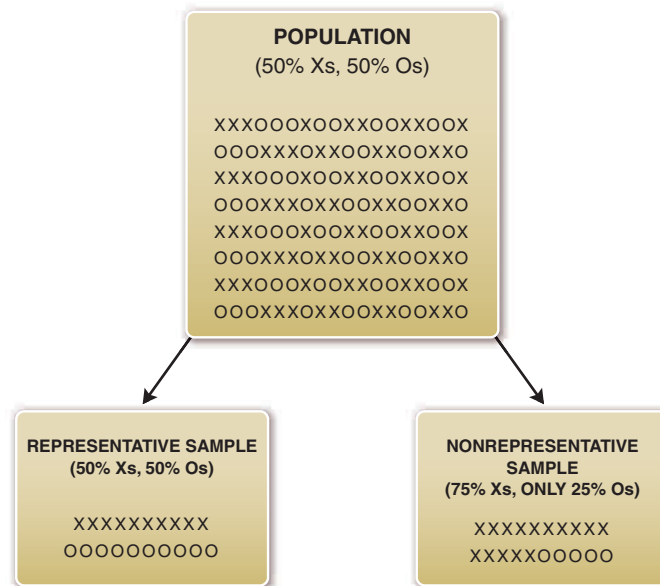
The goal in choosing a sample is to make it as representative as possible of the larger population. That is, if researchers choose students at one university, they need to be reasonably similar to college students elsewhere in the country. If the phrase “reasonably similar” sounds vague, this is because the basis for evaluating a sample varies depending on the hypothesis

and the key variables. For example, if we wanted to study the relationship between family income and stress levels, we would need to make sure that our sample mirrored the population in the distribution of income levels. Thus, a sample of students from a state university might be a better choice than students from, say, Harvard (which costs about \$60,000 per year including room and board). On the other hand, if the research question deals with the pressures faced by students in selective private schools, then Harvard students could be a representative sample for the study.

Figure 4.1 shows a conceptual illustration of both a representative and nonrepresentative sample, drawn from a larger population. The population in this case consists of 144 individuals, split evenly between Xs and Os. Thus, we would want our sample to come as close as possible to capturing this 50/50 split. The sample of 20 individuals on the left is representative of the sample because it is split evenly between Xs and Os. But the sample of 20 individuals on the right is nonrepresentative because it contains 75% Xs. Because the population has far fewer Os than we might expect, this sample does not accurately represent the population. This failure of the sample to represent the population is also referred to as **sampling bias**.

From where do these samples come? Broadly speaking, researchers have two broad categories of sampling strategies at their disposal: probability sampling and nonprobability sampling.

Figure 4.1: Representative and nonrepresentative samples of a population



Probability Sampling

Researchers use **probability sampling** when each person in the population has a known chance of being in the sample. This is possible *only* in cases where researchers know the exact size of the population. For instance, the current population of the United States is 322.1 million people (www.census.gov/popclock/). If we were to select a U.S. resident at random, each

resident would have a one in 322.1 million chance of being selected. Whenever researchers have this information, probability-sampling strategies are the most powerful approach because they greatly increase the odds of getting a representative sample. Within this broad category of probability sampling are three specific strategies: simple random sampling, stratified random sampling, and cluster sampling.

Simple random sampling, the most straightforward approach, involves randomly picking study participants from a list of everyone in the population. The term for this list is a **sampling frame** (e.g., imagine a list of every resident of the United States). To have a truly representative random sample, researchers must have a sampling frame; they must choose from it randomly; and they must have a 100% response rate from those selected. (As Chapter 2 discussed, if people drop out of a study, it can threaten the validity of the hypothesis test.)

Researchers use **stratified random sampling**, a variation of simple random sampling, when subgroups of the population might be left out of a purely random sampling process. Imagine a city with a population that is 80% Caucasian, 10% Hispanic, 5% African American, and 5% Asian. If we were to pick 100 residents at random, the chances are very good that our entire sample would consist of Caucasian residents and ignore the perspective of all ethnic minority residents. To prevent this problem, researchers use stratified random sampling—breaking the sampling frame into subgroups and then sampling a random number from each subgroup. In this example, we could divide the list of residents into four ethnic groups and then pick a random 25 from each of these groups. The end result would be a sample of 100 people that captured opinions from each ethnic group in the population. Notice that this approach results in a sample that does not exactly represent the underlying population—that is, Hispanics constitute 25% of the sample, rather than 10%. One way to correct for this issue is to use a statistical technique known as “weighting” the data. Although the full details are beyond the scope of this book, weighting involves trying to correct for problems in representation by assigning each participant a weighting coefficient for analyses. In essence, people from groups that are underrepresented would have a weight greater than 1, while those from groups that are overrepresented would have a weight less than 1. For more information on weighting and its uses, see <http://www.applied-survey-methods.com/weight.html>.



bowdenimages/iStock/Thinkstock

In a neighborhood with a majority of Caucasian residents, stratified random sampling is needed to capture the perspective of all ethnic groups in the community.

Finally, researchers employ **cluster sampling**, another variation of random sampling, when they do not have access to a full sampling frame (i.e., a full list of everyone in the population). Imagine that we want to study how cancer patients in the United States cope with their illness. Because no list exists of every cancer patient in the country, we have to get a little creative with our sampling. The best way to think about cluster sampling is as

“samples within samples.” Just as with stratified sampling, we divide the overall population into groups, but cluster sampling differs in that we are dividing into groups based on more than one level of analysis. In our cancer example, we could start by dividing the country into regions, then randomly selecting cities from within each region, and then randomly selecting hospitals from within each city, and finally randomly selecting cancer patients from each hospital. The end result would be a random sample of cancer patients from, say, Phoenix, Miami, Dallas, Cleveland, Albany, and Seattle; taken together, these patients would provide a fairly representative sample of cancer patients around the country.

Nonprobability Sampling

The other broad category of sampling strategies is known as **nonprobability sampling**. These strategies are used in the (remarkably common) case in which researchers do not know the odds of any given individual’s being in the sample. This uncertainty represents an obvious shortcoming—if we do not know the exact size of the population and do not have a list of everyone in it, we have no way to know that our sample is representative. Despite this limitation, researchers use nonprobability sampling on a regular basis. We will discuss two of the most common nonprobability strategies here.

In many cases, it is not possible to obtain a sampling frame. When researchers study rare or hard-to-reach populations or study potentially stigmatizing conditions, they often recruit by word-of-mouth. The term for this is **snowball sampling**—imagine a snowball rolling down a hill, picking up more snow (or participants) as it goes. If we wanted to study how often homeless people took advantage of social services, we would be hard pressed to find a sampling frame that listed the homeless population. Instead, we could recruit a small group of homeless people and ask each of them to pass the word along to others, and so on. If we wanted to study changes in people’s identities following sex-reassignment surgery, we would find it difficult to track down this population via public records. Instead, we could recruit one or two patients and ask for referrals to others. The resulting sample in both cases is unlikely to be representative, but researchers often have to compromise for the sake of obtaining access to a population. Snowball sampling is most often used in qualitative research, where the advantages of gaining a rich narrative from these individuals outweigh the loss of representativeness.

One of the most popular nonprobability strategies is known as **convenience sampling**, or simply including people who show up for the study. Any time a 24-hour news station announces the results of a viewer poll, they are likely based on a convenience sample. CNN and Fox News do not randomly select from a list of their viewers; they post a question onscreen or online, and people who are motivated (or bored) enough to respond will do so. As a matter of fact, the vast majority of psychology research studies are based on convenience samples of undergraduate college students. Research in psychology departments often works like this: Experimenters advertise their studies on a website, and students enroll in these studies, either to earn extra cash or to fulfill a research requirement for a course. Students often pick a particular study based on whether it fits their busy schedules or whether the advertisement sounds interesting. These decisions are hardly random and, consequently, neither is the sample. The goal here is not to disparage all psychology research—that would be self-defeating—but to emphasize that all of the decisions researchers make have both pros and cons.

Choosing a Sampling Strategy

Although researchers always strive for a representative sample, no such thing as a perfectly representative one exists. Some degree of **sampling error**, defined as the degree to which the characteristics of the sample differ from the characteristics of the population, is always present. Instead of aiming for perfection, then, researchers aim for an estimate of how far from perfection their samples are. These estimates are known as the **margin of error**, or the degree to which the results from a particular sample are expected to deviate from the population as a whole.

One of the main advantages of a probability sample is that we are able to calculate these errors, as long as we know our sample size and desired level of confidence. In fact, most of us encounter margins of error every time we see the results of an opinion poll. For example, CNN may report that “Candidate A is leading the race with 60% of the vote, $\pm 3\%$.” This means Candidate A’s approval percentage in the sample is 60%, but based on statistical calculations, her real percentage is between 57% and 63%. The smaller the error (3% in this example), the more closely the results from the sample match the population. Naturally, researchers conducting these opinion polls want the error of estimation to be as small as possible. How persuaded would anyone be to learn that “Candidate A has a 10-point lead, plus or minus 20 points?” This margin of error ought to trigger our skepticism, because the real difference is between 30 points and -10 points—i.e., a 10-point lead for *the other candidate*.

Researchers’ most direct means of controlling the margin of error is by changing the sample size. Most survey research aims for a margin of error of less than five percentage points. Based on standard calculations, this requires a sample size of 400 people per group. That is, if we want to draw conclusions about the entire sample (e.g., “30% of registered voters said X”), then we would need at least 400 respondents to say this with some confidence. If we want to draw conclusions about subgroups (e.g., “30% of women compared to 50% of men”), then we would actually need at least 400 respondents of each gender to draw conclusions with confidence.

The magic number of 400 represents a compromise—a researcher is willing to accept 5% error for the sake of keeping time and costs down. It is worth noting, however, that some types of research have more stringent standards: For political polls to be reported by the media, they must have at least 1,000 respondents, which brings the margin of error down to three percentage points. In contrast, some areas of applied research may have more relaxed standards. In marketing research, for example, budget considerations sometimes lead to smaller samples, which means drawing conclusions at lower levels of confidence. For example, with a sample size of 100 people per group, researchers have to contend with 8–10% margin of error—almost double the error, but at a fraction of the costs.

If probability sampling is so powerful, why are nonprobability strategies so popular? One reason is that convenience samples are more practical; they are cheaper, easier, and almost always possible to conduct with relatively few resources because researchers can avoid the costs of large-scale sampling. A second reason is that convenience is often a good-enough starting point for a new line of research. For example, if we wanted to study the predictors of relationship satisfaction, we could start by testing our hypotheses in a controlled setting using college student participants and then extend the research to the study of adult married couples. Finally, and relatedly, in many cases it is acceptable to have a nonrepresentative sample because researchers do not need to generalize results. If we want to study the prevalence

of alcohol use in college students, it may be perfectly acceptable to use a convenience sample of college students. Although, even in this case, researchers would have to keep in mind that they are studying drinking behaviors among students who volunteered to complete a study on drinking behaviors.

In some cases, however, it is critical to use probability sampling, despite the extra effort required. Specifically, researchers use probability samples any time it is important to generalize and any time it is important to predict behavior of a population. The best way of understanding these criteria is to think of political polls. In the lead-up to an election, each campaign is invested in knowing exactly what the voting public thinks of its candidate. In contrast to a CNN poll, which is based on a convenience sample of viewers, polls conducted by a campaign will be based on randomly selected households from a list of registered voters. The resulting sample is much more likely to be representative, much more likely to tell the campaign how the entire population views its candidate, and therefore much more likely to be useful.

4.4 Analyzing Survey Data

Now comes the fun part. Once researchers have designed a survey, chosen an appropriate sample, and collected some data, it is time for analyses. As with the descriptive designs Chapter 3 explained, the goal of these analyses is to subject hypotheses to a statistical test. Surveys can be used both to describe and predict thoughts, feelings, and behaviors. Since Chapter 3 already covered the basics of descriptive analysis, this section will focus on predictive analyses, which are designed to assess the associations between and among variables. Researchers typically use three approaches to test predictive hypotheses: correlational analyses, chi-square analyses, and regression analyses. Each has its advantages and disadvantages, and each is most appropriate for a different kind of data. This section will walk through the basics of each analysis. Because the statistics course discusses these approaches in more detail, the goal here is to acquire a more conceptual overview of each technique and its usefulness in answering research questions.

Correlational Analysis

The beginning of this chapter described an example of a survey research question: What is the relationship between the number of hours that students spend studying and their grades in the class? In this case, the hypothesis claims that we can predict something about students' grades by knowing how many hours they spend studying.

Imagine we collected a small amount of data (shown in Table 4.1) to test this hypothesis. (Of course, a true test of this hypothesis would require more than 10 people in the sample, but these data will do as an illustration.)

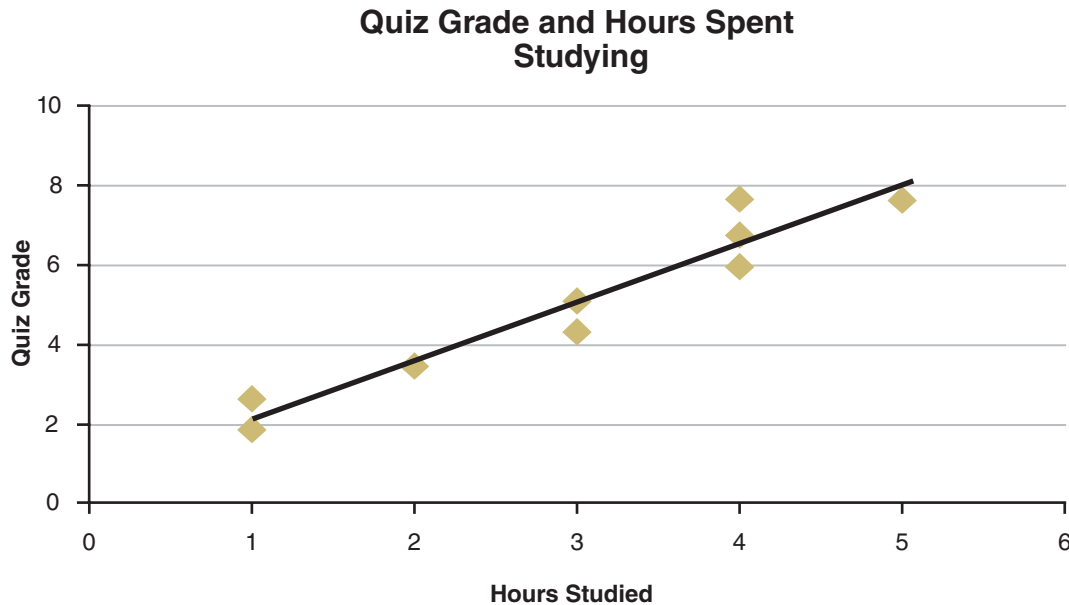
Table 4.1: Data for quiz grade/hours studied example

Participant	Hours Studied	Quiz Grade
1	1	2
2	1	3
3	2	4
4	3	5
5	3	6
6	3	6
7	4	7
8	4	8
9	4	9
10	5	9

The Logic of Correlation

The important question here is whether and to what extent we can predict grades based on study time. One common statistic for testing these kinds of hypotheses is a **correlation**, which gives an assessment of the linear relationship between two variables. A stronger correlation between two variables indicates a stronger association between them. In the case of the current example, the stronger the correlation between study time and quiz grade, the more accurately we can predict grades based on knowing how long the student spends studying.

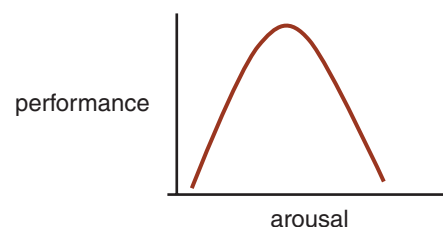
Before we calculate the correlation between these variables, it is always a good idea to visualize the data on a graph. Chapter 3 discussed a type of graph, called a scatterplot, that displays points of data on two variables at a time. The scatterplot in Figure 4.2 shows our sample data from the studying/quiz grade study.

Figure 4.2: Scatterplot for quiz grade/hours studied example

Each point on the graph represents one participant. For example, the point in the top right corner represents a student who studied for five hours and earned a 9 on the quiz. The two points in the bottom right represent students who studied for only one hour and earned a 2 and a 3 on the quiz.

Researchers have three reasons to graph data before conducting statistical tests. First, a graph allows us to get a general sense of the pattern—in this case, students who study less appear to do worse on the quiz. As a result, we will be better informed going into our statistical calculations. Second, the graph lets us examine the raw data for any outliers, or points that stand out as clear exceptions to the overall pattern. These outlier points may indicate that a respondent misunderstood the question and should be dropped from analyses. On the other hand, a cluster of outlier points could indicate the presence of subgroups within our data. Perhaps most students do worse if they study less, but a group of students is able to ace the quizzes without any preparation. Examining this cluster of people in more detail might suggest either a refinement of our hypothesis or an interesting direction for future research.

Third, the graph assures researchers that there is a linear relationship between the variables. This is a very important point about correlations: The math of the standard correlation formula is based on how well the data points fit a straight line, which means nonlinear relationships might be overlooked. Figure 4.3 demonstrates a robust nonlinear finding in psychology regarding the relationship between task performance and physiological arousal. As this graph shows, people tend to perform their best on just about any task when they have a moderate level of arousal.

Figure 4.3: Curvilinear relationship between arousal and performance

When arousal is too high, people find it difficult to calm down and concentrate; when arousal is too low, people find it difficult to care about the task at all. If we simply ran a standard correlation with data on performance and arousal, the correlation would be zero because the points do not fit a straight line. Thus, it is critical to visualize the data before jumping ahead to the statistics. Otherwise, researchers risk overlooking an important finding in the data. (It is important to note that non-linear relationships like this one *can* still be analyzed, but the calculations quickly become complex. In fact, these analyses even require specialized knowledge to use statistical software.)

Interpreting Coefficients

Once we are satisfied that our data look linear, it is time to calculate our statistics. Researchers typically calculate using a computer software program, such as SPSS, SAS, or Microsoft Excel. The number used to quantify the correlation is called the **correlation coefficient**. This number ranges from -1 to $+1$ and contains two important pieces of information:

- The *direction* of the relationship is based on the sign of the correlation coefficient. A $+0.8$ would indicate a positive correlation, meaning that as one variable increases, so does the other variable. A -0.8 would indicate a negative correlation, meaning that as one variable increases, the other variable decreases. (Refer back to Section 2.1 for a review of these two terms.)
- The *size* of the relationship is based on the absolute value of the correlation coefficient. The farther the coefficient is from zero in either direction, the stronger the relationship between variables. For example, both a $+0.8$ and a -0.8 indicate strong relationships.

So, for example, a $+0.2$ represents a weak positive relationship and a -0.7 represents a strong negative relationship.

Calculating the correlation for our quiz-grade study produces a coefficient of 0.962 , indicating a strong positive relationship between studying and quiz grade. What does this mean in plain English? Students who spend more hours studying tend to score higher on the quiz.

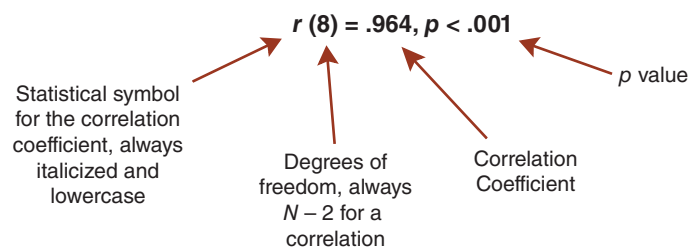
How do we know whether to get excited about a correlation of 0.962 ? As with all of our statistical analyses, we look this value up in a critical value table, or, more commonly, let the computer software do this for us. The critical value table provides a *p* value representing the odds that our correlation is due to random chance. In this case, the *p* value is less than 0.001 . This means that the chance of our correlation being a random fluke is less than 1 in 1,000; we can feel pretty confident in our results.

When interpreting correlation results, realize that statistical significance is closely tied to the sample size. In a small sample, it is possible to see moderate to strong relationships that do not meet the threshold for statistical significance. One good option in these cases is to collect additional data. If the correlation maintains its size and also attains statistical significance, researchers can have some confidence in the results. It is also possible to have the opposite problem: Large sample sizes can make even the smallest relationships show high levels of statistical significance. In a 2008 journal article, Newman, Groom, Handelsman and Pennebaker analyzed differences in language use between men and women. Because the authors had a sample of over 14,000 text samples, even the tiniest differences in language were statistically

significant. For example, men used words related to anger about 4% more than women; with such a large sample, this trivial difference was significant at $p < 0.05$. To deal with this issue, the authors chose to use a more conservative threshold of $p < 0.001$, considering all other results to be too trivial.

Returning to our quiz-grade study, we now have all the information we need to report this correlation in a research paper. The standard way of reporting a correlation coefficient includes information about the sample size (N) and p value, as well as the coefficient itself. Our quiz-grade study would be reported as Figure 4.4 depicts.

Figure 4.4: Correlation coefficient diagram



Where, then, does this leave our hypothesis? We started by predicting that students who spent more time studying would perform better on their quizzes than those who spent less time studying. We then designed a study to test this hypothesis by collecting data on study habits and quiz grades. Finally, we analyzed these data and found a significant, strong, positive correlation between hours studied and quiz grade. Based on this study, our hypothesis has been confirmed—students who study more have higher quiz grades. Of course, because this is a correlational study, we are unable to make causal statements. It could be that studying more for an exam helps students to learn more. Or, it could be the case that previous low quiz grades make students give up and study less. A third variable of motivation could cause students both to study more and perform better on the quizzes. To tease these explanations apart and determine causality calls for a different type of research design, which Chapter 5 will discuss.

Multiple Regression Analysis

Correlations are the best tool to test the linear relationship between pairs of quantitative variables. However, in many cases, researchers are interested in comparing the influence of several variables at once. Imagine we want to expand the study about hours studying and quiz grade by looking at other variables that might predict students' quiz grades. We have already learned that the hours students spend studying correlate positively with their grades. But what about SAT scores? Will students with higher standardized-test scores do better in all of their college classes? What about the number of classes that students have previously taken in the subject area? Will increased familiarity with the subject be associated with higher scores? To compare the influence of all three variables, we can use a slightly different analytic approach. **Multiple regression** is a variation on correlational analysis in which more than one predictor variable is used to predict a single outcome variable. In this example, we would

attempt to predict the outcome variable of quiz scores based on three predictor variables: SAT scores, number of previous classes, and hours studied.

Multiple regression requires an extensive set of calculations; consequently, it is always performed by computer software. A detailed look at these calculations is beyond the scope of this book, but a conceptual overview will help convey the unique advantages of this analysis. Essentially, the calculations for multiple regression are based on the correlation coefficients between each of our predictor variables, as well as between each of these variables and the outcome variable. Table 4.2 shows these correlations for our revised quiz-grade study. If we scan the top row, we see the correlations between quiz grade and the three predictor variables: SAT ($r = 0.14$), previous classes ($r = 0.24$), and hours studied ($r = 0.25$). The remainder of the table shows correlations between the various predictor variables; for example, hours studied and previous classes correlate at $r = 0.24$. When researchers conduct multiple regression analysis using computer software, the software will use all of these correlations in performing its calculations.

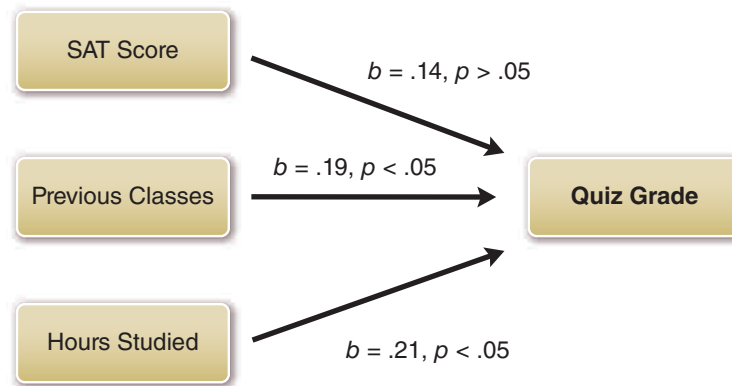
Table 4.2: Correlations for a multiple regression analysis

	Quiz Grade	SAT Score	Previous Classes	Hours Studied
Quiz Grade	—	0.14	0.24*	0.25*
SAT Score		—	.02	-.02
Previous Classes			—	0.24*
Hours Studied				—

Note. Correlations marked with an asterisk (*) are statistically significant at the 95% confidence level. This notation in results tables is common and allows researchers to quickly spot the most interesting findings.

The advantage of multiple regression is that it considers both the individual and the combined influence of the predictor variables. Figure 4.5 shows a visual diagram of the individual predictors of quiz grades. The numbers along each line are known as **regression coefficients**, or **beta weights**. These values are very similar to correlation coefficients but differ in an important way: They represent the effects of each predictor variable while controlling for the effects of all the other predictors. That is, the value of $b = 0.21$ linking hours studied with quiz grades is the independent contribution of hours studied, controlling for SAT scores and previous classes. If we compare the size of these regression coefficients, we see that, in fact, hours spent studying is still the largest predictor of quiz grades ($b = 0.21$), compared to both SAT scores ($b = 0.14$) and previous classes ($b = 0.19$).

Even if individual variables only have a small influence, they can add up to a larger combined influence. So, if we were to analyze the predictors of quiz grades in this study, we would find a combined **multiple correlation coefficient** of $r = 0.34$. The multiple correlation coefficient represents the combined association between the outcome variable and the full set of predictor variables. Note that in this case, the combined r of 0.34 is larger than any of the individual correlations in Table 4.2, which ranged from 0.14 to 0.25. These numbers mean that we are better able to predict quiz grades from examining all three variables than we are from examining any single variable. Or, as the saying goes, the whole is greater than the sum of its parts.

Figure 4.5: Predictors of quiz grades

Multiple regression is an incredibly useful and powerful analytic approach, but it can also be a difficult concept to grasp. Before moving on, we will revisit the concept in the form of an analogy. Imagine someone has just eaten the most delicious hamburger of his life and is determined to understand what makes it so good. Many things contribute to the taste of the hamburger: the quality of the meat, the type and amount of cheese, the freshness of the bun, perhaps the smoked chili peppers layered on top. If the diner were to approach this investigation using multiple regression, he would be able to distinguish the influence of each variable (how important is the cheese compared to the smoked peppers?) as well as take into account the full set of ingredients (does the freshness of the bun really matter when the other elements taste so good?). Ultimately, the individual would be armed with the knowledge of which elements are most important in crafting the perfect hamburger and would understand more about the perfect hamburger than if he had examined each ingredient in isolation.

Chi-Square Analyses

Both correlations and regressions are well suited to testing hypotheses about prediction, as long as we can demonstrate a linear relationship between two variables. Linear relationships, however, require that variables be measured on one of the quantitative scales, that is, ordinal, interval, or ratio scales (see Section 2.3 for a review). What if we want to test an association between nominal, or categorical, variables? In these cases, we need an alternative statistic called the **chi-square statistic**, which determines whether two nominal variables are independent from or related to one another. Chi-square is often abbreviated with the symbol χ^2 , which shows the Greek letter chi with the superscript 2 for squared. (This statistic is also referred to as the chi-square test for independence—a slightly longer but more descriptive synonym.)

The idea behind this test is similar to that of the correlation coefficient. If two variables are independent, then knowing the value of one variable does not tell us anything about the value of the other variable. As we will see in the example below, a larger chi-square reflects a larger deviation from what we would expect by chance and is thus an index of statistical significance.

Imagine that we want to know whether people in rural or urban areas are more likely to support a sales-tax increase. We can easily speculate why either group might be more likely to do so—perhaps people living in cities are more politically liberal or perhaps people living in small towns are better able to see benefits of higher local taxes. So, we might survey a sample of 100 people, asking them to indicate both their location (rural or urban) and their support for a sales-tax proposal. The survey produces the following results (in Table 4.3), presented in a **contingency table**, which displays the number of individuals in each combination of our nominal variables. Notice that we have more urban than rural residents, reflecting the higher population density in cities.

But, as it turns out, the raw numbers are less important than the ratios within each group. The chi-square calculation works by first considering what each cell in the table would look like *if there were no relationship at all* (i.e., under the null hypothesis), and then determining how much the data differ from that reference point.

Table 4.3: Chi-square example: support for a sales tax increase

	Rural	Urban	Total
Support	10	45	55
Don't Support	30	15	45
Total	40	60	100

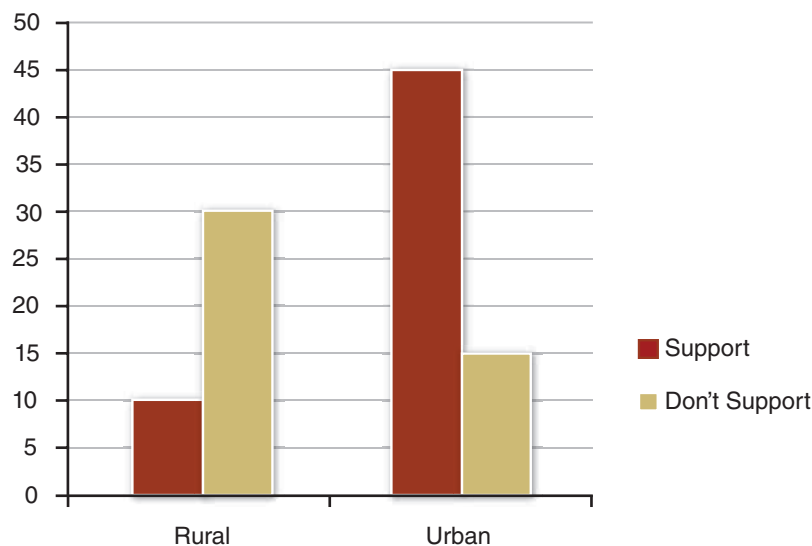
In this example, our final chi-square value is 34.55; this represents the total difference across the table between actual and expected data. The larger this number is, the more our observed data differ from the expected frequencies, and the more our variables relate to one another. In the current example, this means we can predict a person's support for a sales-tax increase based on where he or she lives, which is consistent with our initial hypothesis.

Still, how do we know if our value of 34.55 is meaningful? As with the other statistical tests we have discussed, determining the significance requires looking up the result in a critical-value table to assess whether the calculated value is above threshold. In this case, the critical value for a chi-square with a 2×2 table = 3.84, so we can feel confident in our value of 34.55—almost 10 times higher than the threshold value.

However, unlike correlation and regression coefficients, our chi-square results cannot tell us anything about the direction or magnitude of the relationship. A larger chi-square reflects a larger deviation from what we would expect by chance and is thus an index of statistical significance. To interpret the patterns of our data, we need to visually inspect the numbers in our data table. Better yet, we can create a bar graph like we did in Chapter 3 to display these frequencies visually.

As Figure 4.6 shows, the cell frequencies suggest a fairly clear interpretation: People who live in urban settings are much more likely than people who live in rural settings to support a sales-tax increase. In fact, urban residents *support* the increase by a 3-to-1 margin, while rural residents *oppose* the increase by a 3-to-1 margin.

Figure 4.6: Graph of chi-square results



Research: Thinking Critically

Self-Esteem in Youth and Early Adulthood

Follow the link below to read a press release from the American Psychological Association, describing recent research on self-esteem during adolescence. This study, by a group of Swiss researchers, challenges some of our popular assumptions about gender differences in self-esteem. As you read the article, consider what you have learned so far about the research process, and then respond to the questions below.

<http://www.apa.org/news/press/releases/2011/07/youth-self-esteem.aspx>

Think About It:

1. Why is self-esteem a good topic to study using survey research methods? Does using a survey to study self-esteem present any weaknesses?
2. What type of sampling was used in this study? Was this an appropriate strategy?
3. What type of data analysis discussed in this chapter is appropriate to understanding the influence of multiple variables (mastery, health, income) on self-esteem?

Summary and Resources

Chapter Summary

This chapter has covered the process of survey research from conceptualization through analysis. We first discussed the types of research questions that are best suited to survey research—essentially, those that can be answered based on people’s observations of their own behavior. Survey research can involve either verbal reports (i.e., interviews) or written reports (i.e., questionnaires). In both cases, surveys are distinguished by their reliance on people’s self-reports of their attitudes, feelings, and behaviors.

This chapter covered several key points for writing survey items. The key takeaway of the five rules for better questions is that questions should be written as clearly and unambiguously as possible. This helps to minimize the error variance that might result from participants imposing their own guesses and interpretations on the material. When designing survey items, researchers also have a broad choice between open-ended and fixed-format responses. The former provide richer and more extensive data but are harder to score and code; the latter are easier to code but can constrain people’s responses to a researcher’s choice of categories. If and when researchers settle on a fixed-format response, they have another set of decisions to make regarding the response scaling, labels, and general format.

Once researchers have constructed the scale, it is time to begin data collection. This chapter discussed the concept of sampling, or choosing a portion of the population to use for a study. Broadly speaking, sampling can be either “probability” or “nonprobability,” depending on whether researchers have a known population size from which they sample randomly. Probability sampling is more likely to result in a representative sample, but this approach is not possible in all studies. In fact, a significant proportion of psychology research studies use a form of nonprobability sampling called convenience sampling, meaning that the sample consists of those who show up for the study.

Finally, this chapter covered three approaches to analyzing survey data and testing hypotheses about prediction. The first, correlational analysis, is a very popular way to analyze survey data. The correlation is a statistical test that assesses the linear relationship between two variables. The stronger the correlation between variables, the more we can accurately predict one based on knowing the other. Second, regression analyses allow us to expand our investigations into multiple predictors. Multiple regression offers the advantage of considering both the individual and the combined influence of the predictor variables. However, both correlation and regression require the variables to be quantitative—that is, measured on an ordinal, interval, or ratio scale. In cases where our survey produces nominal or categorical data, we use an alternative called the chi-square statistic, which determines whether two nominal variables are independent or related. The chi-square works by examining the extent to which our observed data deviate from the pattern we would expect if the variables were unrelated.

The common thread in all these analyses is that while they measure the association between variables, they do not tell us anything about the causal relationship between them. To make causal statements, we have to conduct experiments, which the next chapter will discuss.

Key Terms

anchors Labels, or endpoints, for a rating scale.

bipolar scale Rating scale that has polar opposites as its anchors.

branching schedule An interview format in which questions branch in different directions depending on participants' answers.

chi-square statistic A statistical test similar to the correlation coefficient; determines whether two nominal variables are independent or related.

cluster sampling A variation of simple random sampling that involves dividing the sample into groups based on more than one level of analysis.

contingency table A data summary table that shows the number of individuals in each combination of the nominal variables; used as the first step in calculating chi-square.

convenience sampling A nonprobability sampling strategy that involves simply enrolling people who show up for the study.

correlation Statistical test that gives an assessment of the linear relationship between two variables; the stronger the correlation between variables, the more accurately the prediction about one is based on knowing the other.

correlation coefficient The number used to quantify a correlation; this coefficient (r), ranges from -1 to $+1$ and contains information about both the size and direction of a correlation.

double-barreled question A flawed survey item that asks more than one question at a time.

fixed-format response Answer to a limiting question or statement, involving choosing from a list of options; on a survey, fixed-format responses are easier to code but can constrain the data into narrow categories.

forced choice A rating scale that requires respondents to agree or disagree with a statement, usually through the use of an even number of scale points.

interview A survey that is administered verbally.

interview schedule A plan or script for the progress of an interview that describes the list of questions and the order in which they should be asked.

leading question A flawed survey item worded in a way that suggests an answer.

Likert scale Format that uses anchors of "strongly agree" and "strongly disagree" to rate responses to a survey question.

linear schedule An interview format that asks the same questions in the same order for all participants.

margin of error The degree to which the data from the sample are expected to deviate from the population as a whole.

multiple-choice format A type of fixed-format response that asks participants to select from a set of predetermined responses.

multiple correlation coefficient A number that represents the combined association between the outcome variable and the full set of predictor variables.

multiple regression A variation on correlational analysis in which more than one predictor variable is used to predict a single outcome variable.

nonprobability sampling A group of sampling strategies used when the odds of any given individual's being in the sample are unknown.

open-ended response Unstructured answer to a question or statement; on a survey, open-ended responses provide rich data but are hard to code.

pilot testing "Test run" of a survey that involves giving the questionnaire to a small sample of people, getting their feedback, and making any necessary changes.

population The entire collection of people who could possibly be relevant for a study.

probability sampling A group of data-collection strategies used when each person in the population has a known chance of being in the sample.

questionnaire A survey that is administered in writing.

rating scale A fixed-format response that asks participants to place responses on a continuum.

regression coefficients (beta weights) Values that represent the effects of each predictor variable while controlling for the effects of all the other predictors.

sampling bias The failure of the sample to represent the underlying distribution in the population.

sampling error The degree to which the characteristics of the sample differ from the characteristics of the population.

sampling frame A list of all members of a particular population (e.g., a list of every resident of the United States) and a necessary requirement for probability sampling strategies.

self-reports Participants' reports of their own attitudes, feelings, and behaviors.

simple random sampling A probability sampling strategy that involves randomly picking participants from a list of everyone in the population.

snowball sampling A nonprobability sampling strategy that involves recruiting by word-of-mouth referrals.

social desirability Participants' reluctance to give unpopular answers to survey questions.

stratified random sampling A variation of simple random sampling, used when subgroups of the population might be left out of a purely random sampling process; involves breaking the sampling frame into subgroups and then sampling a random number from each subgroup.

survey research Any method that relies on people's observations of their own behavior.

true/false format A fixed-format response that asks participants to indicate whether they endorse a statement.

unipolar scale Rating scale that assess a single construct.

Apply Your Knowledge

1. For each of the following poorly written questionnaire items, identify the major problem and then rewrite it so that the problem is resolved.

- a. How much do you like cats and ponies?

main problem:

better item:

- b. Do you think that John McCain's complete lack of personality proved that he would have been a terrible president?

main problem:

better item:

- c. Do you dislike not playing basketball?

main problem:

better item:

- d. Do you support SB 1070?

main problem:

better item:

- e. How often do you take drugs?

main problem:

better item:

2. Dr. Truxillo is interested in Arizona residents' thoughts and feelings about global warming. For each of the following examples, identify the sampling method used by her research assistants.

- a. Alejandra sets up a table in the mall and hands a survey to people who approach her.
- b. Catherine randomly chooses five cities, then chooses three neighborhoods in each, then randomly samples 5,000 households for a phone survey.
- c. Isaiah starts with a list of the entire population of Arizona and selects participants by dialing random phone numbers.
- d. Anna obtains the master list from Isaiah and divides the population according to education level. She then randomly chooses 500 high school dropouts, 500 college graduates, and 500 people with some postgraduate education.

3. Based on each of the following study descriptions, choose whether the best analysis would be a correlation, a multiple regression, or a chi-square.
 - a. Ahmad is interested in the relationship between annual income and self-reported happiness.
 - b. Sheila is interested in whether some ethnic groups are more likely to use counseling services (a yes-or-no question).
 - c. Angela is interested in knowing the best predictors of recovery from depression, comparing the influence of drugs, therapy, and family resources.
 - d. Kartik is interested in whether high school dropouts or college graduates are more likely to vaccinate their children.
 - e. Nicole is interested in understanding the best predictors of weight loss.
 - f. Isabella is interested in the relationship between self-esteem and prejudice.

Critical Thinking Questions

1. In survey research, explain the trade-off between the “richness” of people’s responses, and the ease of analyzing their responses.
2. When conducting interviews, the researcher has a personal interaction with the subject. Why is this both good and bad?
3. What are some of the new challenges in conducting surveys over the Internet? On mobile devices?
4. Explain the compromises between confidence level and research costs. When might researchers be willing to accept more error in their findings?

Research Scenarios: Try It

Mindfulness can be defined as “awareness of the present moment without judgment.” Recent research has found that a tendency to be mindful in everyday life is associated with reduced risk for emotional disorders such as depression and anxiety. Suppose you are interested in understanding the role that mindfulness may play in reducing negative reactions to stressful events. You decide to investigate this relationship in a sample of college students who undergo a significant stressor at the same time: final exams. You hypothesize that college students who are less mindful will be more stressed during final exams than those who are more mindful.

1. To test this hypothesis, what psychological construct(s) would you have to measure?
 - a. Mindfulness
 - b. Mindfulness and final exams
 - c. Mindfulness and well-being
 - d. Mindfulness and stress
2. Suppose you decide that a questionnaire survey would be the most appropriate method to measure mindfulness and stress. While you can use existing questionnaire measures of stress, you decide to create your own survey measure of mindfulness. What problem might be a particular concern in constructing a questionnaire asking about the tendency to be mindful?
 - a. Social desirability
 - b. Inefficient data collection
 - c. Double-barreled questions
 - d. May be too time-consuming

3. Suppose you want to create a questionnaire item with the following base: “When I have distressing thoughts or images, I just notice them and let them go.” Which of the following response options would be both easy to score and appropriate for this question?
 - a. True/false
 - b. Multiple choice
 - c. Likert-style rating scale
 - d. Open-ended
4. The design of your study requires that participants be students who undergo a final exam. Suppose you run this study at a local college and advertise by posting flyers around school inviting all students to participate in the survey. The study has no eligibility requirements other than being a student and having a final exam at the end of the quarter. What type of sampling does this represent?
 - a. Snowball sampling
 - b. Convenience sampling
 - c. Cluster sampling
 - d. Stratified random sampling
5. Suppose you administer your mindfulness questionnaire at the beginning of the semester and administer your stress questionnaire to the same students during finals. Which of the following correlations between mindfulness and stress would provide support for your original hypothesis?
 - a. $r = 0.65, p < 0.05$
 - b. $r = -0.65, p < 0.05$
 - c. $r = -0.20, p > 0.10$
 - d. $r = 0.001, p > 0.1$

See Appendix A for answers to Research Scenarios: Try It questions.

5 Experimental Designs— Explaining Behavior



Antonio Oquias/Hemera/Thinkstock

Learning Outcomes

By the end of this chapter, you should be able to:

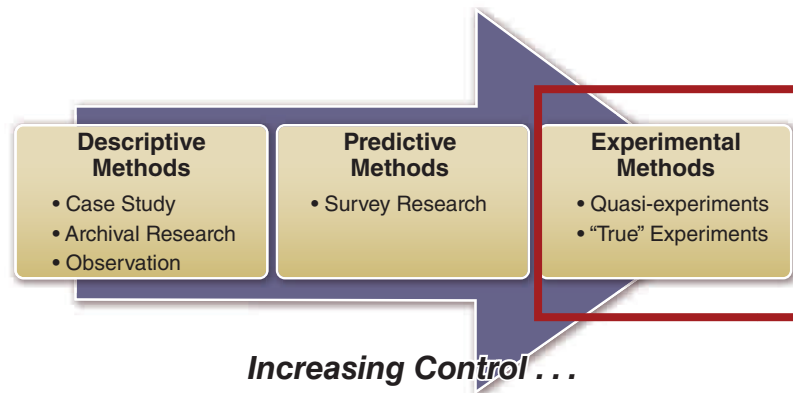
- Use appropriate terminology when discussing experimental designs.
- Identify the key features of experiments for making causal statements.
- Explain the importance of both internal and external validity in experiments.
- Describe the threats to both internal and external validity in experiments.
- Outline the most common types of experimental designs.
- Describe methods for analyzing experimental data.
- Summarize methods for avoiding Type I and Type II error.

One of the oldest debates within psychology concerns the relative contributions of biology and the environment in shaping our thoughts, feelings, and behaviors. Do we become who we are because it is hard-wired into our DNA, or because of our early experiences? Do people share their parents' personality quirks because they carry their parents' genes, or because they grew up in their parents' homes? Researchers can, in fact, address these types of questions in several ways. A consortium of researchers at the University of Minnesota has spent the past three decades comparing pairs of identical and fraternal twins, raised in the same versus different households, to tease apart the contributions of genes and environment. Read more at the research group's website, <http://mctfr.psych.umn.edu/>.

An alternative way to separate genetic and environmental influence is through the use of **experimental designs**, which have the primary goal of explaining the causes of behavior. Recall from the design overview in Chapter 2 (2.1) that experiments can address causal relationships because the experimenter has control over the environment as well as over the manipulation of variables. One particularly ingenious example comes from the laboratory of Michael Meaney, a professor of psychiatry and neurology at McGill University. Meaney used female rats as experimental subjects (Francis, Dioro, Liu, & Meaney, 1999). His earlier research had revealed that the parenting ability of female rats could be reliably classified based on how attentive they were to their rat pups, as well as how much time they spent grooming the pups. The question tackled in the 1999 study was whether these behaviors were learned from the rats' own mothers or transmitted genetically. To answer this question experimentally, Meaney and colleagues had to think very carefully about the comparisons they wanted to make. To simply compare the offspring of good and bad mothers would have been insufficient—this approach could not distinguish between genetic and environmental pathways.

Instead, Meaney decided to use a technique called cross-fostering, or switching rat pups from one mother to another as soon as they were born. The technique resulted in four combinations of rats: (1) those born to inattentive mothers but raised by attentive ones, (2) those born to attentive mothers but raised by inattentive ones, (3) those born and raised by attentive mothers, and (4) those born and raised by inattentive mothers. Meaney then tested the rat pups several months later and observed the way they behaved with their own offspring. Meaney's control over all aspects of how the rat pups were raised was a critical element; he was able to keep everything the same except for the combination of their genetics and rearing environment. The setup of this experiment allowed Meaney to make clear comparisons between the influence of birth mothers and the rearing process. At the end of the study, the conclusion was crystal clear: Maternal behavior is all about the environment. Those rat pups that ultimately grew up to be inattentive mothers were those who had been raised by inattentive mothers.

This final chapter is dedicated to experimental designs, in which the primary goal is to explain behavior. Experimental designs rank highest on the continuum of control (see Figure 5.1) because the experimenter can manipulate variables, minimize extraneous variables, and assign participants to conditions. The chapter begins with an overview of the key features of experiments and then explains the importance of both internal and external validity of experiments. From there, the discussion moves to the process of designing and interpreting experiments and concludes with a summary of strategies for minimizing error in experiments.

Figure 5.1: Experimental designs on the continuum of control

5.1 Experiment Terminology

Before we dive into the details, it is important to cover the terminology that the chapter will use to describe different aspects of experimental designs. Much of this will be familiar from Chapter 2, with a few new additions. First, we will review the basics.

Recall that a variable is any factor that has more than one value. For example, height is a variable because people can be short, tall, or anywhere in between. Depression is a variable because people can experience a wide range of symptoms, from mild to severe. The *independent variable* (IV) is the variable that is manipulated by the experimenter to test hypotheses about cause. The *dependent variable* (DV) is the variable that is measured by the experimenter to assess the effects of the independent variable. For example, in an experiment testing the hypothesis that fear causes prejudice, fear would be the independent variable and prejudice would be the dependent variable. To keep these terms straight, it is helpful to think of the main goal of experimental designs. That is, we test hypotheses about cause by manipulating an independent variable and then looking for changes in a dependent variable. This means that we think the independent variable causes changes in the dependent variable; for example, we hypothesize that fear causes changes in prejudice.

When we manipulate an independent variable, we will always have two or more versions of the variable; this is what distinguishes experiments from, say, structured observational studies. One common way to describe the versions of the IV is in terms of different groups, or **conditions**. The most basic experiments have two conditions: The **experimental condition** receives a treatment designed to test the hypothesis, while the **control condition** does not receive this treatment. In the fear and prejudice example above, the participants who make up the experimental condition would be made to feel afraid, while the participants who make up the control condition would not. This setup allows us to test whether introducing fear to one group of participants leads them to express more prejudice than the other group of participants, who are not made fearful.

Another common way to describe these versions is in terms of **levels** of the independent variable. Levels describe the specific set of circumstances created by manipulating a variable. For example, in the fear and prejudice experiment, the variable of fear would have two levels—afraid and not afraid. We have countless ways to operationalize fear in this experiment. One option would be to adopt the technique used by the Stanford social psychologist Stanley Schachter (1959), who led participants to believe they would be exposed to a series of painful electric shocks. In Schachter’s study, the painful shocks never happened, but they did induce a fearful state as people anticipated them. So, those at the “afraid” level of the independent variable might be told to expect these shocks, while those at the “not afraid” level of the independent variable would not be given this expectation.

At this stage, having two sets of vocabulary terms—“levels” and “conditions”—for the same concept may seem odd. However, with advanced experimental designs using multiple independent variables, there is a subtle difference in how these terms are used. As the designs become more complex, it is often necessary to expand IVs to include several groups and multiple variables. At that point, researchers need different terminology to distinguish between the versions of one variable and the combinations of multiple variables. The chapter will later return to this complexity, in the section “Experimental Design.”

5.2 Key Features of Experiments

The overview of designs in Chapter 2 described the overall process of experiments in the following way: Researchers control the environment as much as possible so that all participants have the same experience. The researchers then manipulate, or change, one key variable, and then measure the outcomes in another key variable. This section examines this process in more detail. Experiments can be distinguished from all other designs by three key features: manipulating variables, controlling the environment, and assigning people to groups.

Manipulating Variables

The most crucial element of an experiment is researcher’s manipulation, or change, of some key variable. To study the effects of hunger, for example, a researcher could manipulate the amount of food given to the participants, or to study the effects of temperature, the experimenter could raise and lower the temperature of the thermostat in the laboratory. In both cases, recall that the researcher needs a way to *operationalize* the concepts (hunger and temperature) into measurable variables. For example, the experimenter could define “hungry” as being deprived of food for eight hours, and define a “hot” room as being 90 degrees Fahrenheit. Because these factors are under the direct control of the experimenters, they can feel more confident that changing them contributes to changes in the dependent variables.

Chapter 2 discussed the main shortcoming of correlational research: These designs do not allow researchers to make causal statements. Recall from that chapter (as well as from Chapter 4) that correlational research is designed to predict one variable from another. One of the examples in Chapter 2 concerned the correlation between income levels and happiness, with the goal of trying to predict happiness levels based on knowing people’s income level. If we

measure these as they occur in the real world, we cannot say for sure which variable causes the other. However, we could settle this question relatively quickly with the right experiment. Suppose we bring two groups into the laboratory and give one group \$100 and a second group nothing. If the first group is happier at the end of the study, it would support the idea that money really does buy happiness. Of course, this experiment is a rather simplistic look at the connection between money and happiness. Even so, because we manipulate levels of money, this study would bring us closer to making causal statements about the effects of money.

To manipulate variables, it is necessary to have at least two versions of the variable. That is, to study the effects of money, we need a comparison group that does not receive money. To study the effects of hunger, we would need both a hungry and a not-hungry group. Having two versions of the variable distinguishes experimental designs from the structured observations discussed in Chapter 3 (3.4), in which all participants receive the same set of conditions in the laboratory. Even the most basic experiment must have two sets of conditions, which are often an experimental group and a control group. However, as this chapter will later explain, experiments can become much more complex. A study might have one experimental group and two control groups, or five degrees of food deprivation, ranging from 0 to 12 hours without food. Decisions about the number and nature of these groups will depend on consideration of both the hypotheses and previous literature.

Researchers have three options for manipulating variables. First, **environmental manipulations** involve changing some aspect of the setting. Environmental manipulations are perhaps the most common in psychology studies, and they include everything from varying the room temperature to varying the amount of money people receive. The key is to change the way that different groups of people experience their time in the laboratory—it is either hot or cold, and they either receive or do not receive \$100.

Second, **instructional manipulations** involve changing the way a task is described to change participants' mindsets. For example, a researcher might give the same math test to all participants but to one group, describe it as an "intelligence test" and to another group, a "problem-solving task." Because an intelligence test is thought to have implications for life success, the experimenter might expect participants in that group to be more nervous about their scores.

Finally, an **invasive manipulation** involves taking measures to change internal, physiological processes; it is usually conducted in medical settings. For example, studies of new drugs involve administering the drug to volunteers to determine whether it has an effect on some physical or psychological symptom. Alternatively, studies of cardiovascular health often involve having participants run on a treadmill to measure how the heart functions under stress.



Monkey Business Images/Monkey Business/Thinkstock

Having a patient run on a treadmill to measure cardiovascular stress is an example of invasive manipulation.

The rule that we *must* manipulate a variable has one qualification. In many experiments, researchers divide participants based on a preexisting difference (e.g., gender) or personality measures (e.g., self-esteem or neuroticism) that capture stable individual differences among people. The idea behind these personality measures is that someone scoring high on a measure of neuroticism (for example) would be expected to be more neurotic across situations than someone scoring lower on the measure. Using this technique allows a researcher to compare how, for example, men and women or people with high and low self-esteem respond to manipulations.

When researchers use preexisting differences in an experimental context, they are referred to as **quasi-independent variables**—“quasi,” or “nearly,” because they are being measured, not manipulated, by the experimenter, and thus do not meet the criteria for a regular independent variable. In fact, variables used in this way are things that *cannot* be manipulated by an experimenter—either for practical or ethical reasons—including gender, race, age, eye color, religion, and so forth. Instead, these are treated as independent variables in that participants are divided into groups along these variables (e.g., male versus female; Catholic versus Protestant versus Muslim).

Because these variables are not manipulated, an experimenter cannot make causal statements about them. For a study to count as an experiment, these quasi-independent variables would have to be combined with a true independent variable. This could be as simple as comparing how men and women respond to a new antidepressant drug—gender would be quasi-independent while drug type would be a true independent variable.

Sometimes the line between true and quasi-experiments can be subtle. Imagine we want to study the effects on people’s persistence at a second task based on winning versus losing a contest. In a quasi-experimental approach, we could have two participants play a game, resulting in a natural winner and loser, and then compare how long each one stuck with the next game. The approach’s limitation is that some preexisting condition might have affected winning and losing the first game. Perhaps the winners had more self-confidence and patience at the start. However, we could improve the design to be a true experiment by having participants play a rigged game against a confederate, thereby causing participants either to win or lose. In this case, we would be manipulating winning and losing, and preexisting differences would be averaged out across the groups (more on this later in the chapter).

Controlling the Environment

The second important element of experimental designs is the researcher’s high degree of control over the environment. In addition to manipulating variables, an experimenter has to ensure that the other aspects of the environment are the same for all participants. For instance, if we were interested in the effects of temperature on people’s mood, we could manipulate temperature levels in the laboratory so that some people experienced warmer temperatures and other people cooler temperatures. However, it is equally important to make sure that *other* potential influences on mood are the same for both groups. That is, we would want to make sure that the “warm” and “cool” groups were tested in the same room, around the same time of day, and by similar experimenters.

The overall goal, then, is to control **extraneous variables**, or variables that add noise to the hypothesis test. In essence, the more researchers can control extraneous variables, the more confidence they can have in the results of the hypothesis test. As the section “Validity and Control” will discuss, these extraneous variables can have different degrees of impact on a study. Imagine we conduct the study on temperature and mood, and all of our participants are in a windowless room with a flickering fluorescent light. This environment would likely influence people’s mood—making everyone a little bit grumpy—but it causes fewer problems for our hypothesis test because it affects everyone equally. Table 5.1 shows hypothetical data from two variations of this study, using a 10-point scale to measure mood ratings. In the top row, participants were in a well-lit room; notice that participants in the cooler room reported being in a better mood (i.e., an 8 versus a 5). In the bottom row, all participants were in the windowless room with flickering lights. These numbers suggest that people were still in a better mood in the cooler room (5) than a warm room (2), but the flickering fluorescent light had a constant dampening effect on everyone’s mood.

Table 5.1: Influence of an extraneous variable

	Cool Room	Warm Room
Variation 1: Well-Lit	8	5
Variation 2: Flickering Fluorescent	5	2

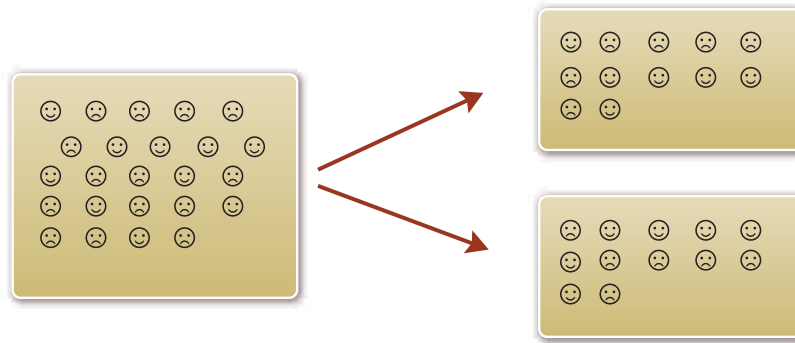
Assigning People to Conditions

The third key feature of experimental designs is that the researcher can assign people to receive different conditions, or versions, of the independent variable. This is an important piece of the experimental process: Experimenters not only control the options—warm versus cool room, \$100 versus no money, etc.—but they also control which participants get each option. Whereas a correlational design might assess the relationship between current mood and choosing the warm room, an experimental design will assign some participants to the warm room and then measure the effects on their mood. In other words, experimenters are able to make causal statements because they cause things to happen to a particular group of people.

The most common, and most preferable, way to assign people to conditions is through a process called **random assignment**. An experimenter who uses random assignment makes a separate decision for each participant as to which group he or she will be assigned to before the participant arrives. As the term implies, this decision is made randomly—by flipping a coin, using a random number table (for an example, see <http://stattrek.com/tables/random.aspx>), drawing numbers out of an envelope, or even simply alternating back and forth between experimental conditions. The overall goal is to try to balance preexisting differences among people, as Figure 5.2 illustrates. So, for example, some people might generally be more comfortable in warm rooms, while others might be more comfortable in cold rooms. If each person who shows up for the study has an equal chance of being in either group, then the groups in the sample should reflect the same distribution of differences as the population.

Figure 5.2: Random assignment

The 24 participants in our sample consist of a mix of happy and sad people. The goal of random assignment is to have these differences distributed equally across the experimental conditions. Thus, the two groups on the right each consist of six happy and six sad people, and our random assignment was successful.



Forming groups through random assignment also has the significant advantage of helping to avoid bias in the selection and assignment of subjects. For example, it would be a bad idea to assign people to groups based on a first impression of them because participants might be placed in the cold room if they arrived at the laboratory dressed in warm clothing. Experimenters who make decisions about condition assignments ahead of time can be more confident that the independent variable is responsible for changes in the dependent variable.

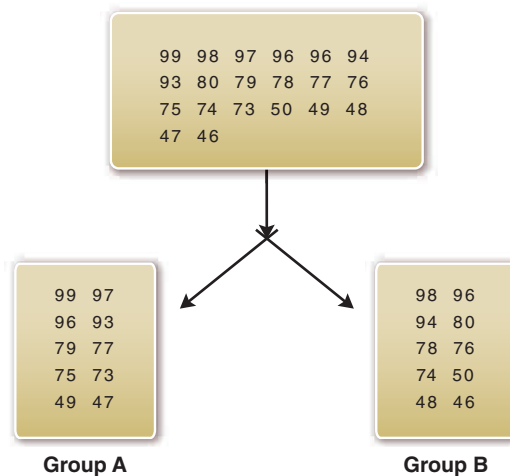
Worth highlighting here is the difference here between random selection and random assignment (discussed in Chapter 4). *Random selection* means that the sample of participants is chosen at random from the population, as with the probability sampling methods discussed in Chapter 4. However, most psychology experiments use a convenience sample of individuals who volunteer to complete the study. This means that the sample is often far from fully random. However, a researcher can still make sure that the study involves *random assignment* to groups, so that each condition contains an equal representation of the sample.

In some cases—most notably, when samples are small—random assignment may not be sufficient to balance an important characteristic that might affect the results of a particular study. Imagine conducting a study that compared two strategies for teaching students complex math skills. In this example, it would be especially important to make sure that both groups contained a mix of individuals with, say, average and above-average intelligence. For this reason, the experimenter would necessarily take extra steps to ensure that intelligence was equally distributed between the groups, which can be accomplished with a variation on random assignment called **matched random assignment**. This kind of assignment requires the experimenter to obtain scores on an important matching variable—in this case, intelligence—rank participants based on the matching variable, and then randomly assign people

to conditions. Figure 5.3 shows how this process would unfold in our math-skills study. First, the researcher gives participants an IQ test to measure preexisting differences in intelligence. Second, the experimenter ranks participants based on these scores, from highest to lowest. Third, the experimenter moves down this list in order and randomly assigns each participant to one of the conditions. This process still contains an element of random assignment, but adding the extra step of rank ordering ensures a more balanced distribution of intelligence test scores across the conditions.

Figure 5.3 Matched random assignment

The 20 participants in our sample represent a mix of very high, average, and very low intelligence test scores (measured 1–100). The goal of matched random assignment is to ensure that this variation is distributed equally across the two conditions. The experimenter would first rank participants by intelligence test scores (top box), and then distribute these participants alternately between the conditions. The end result is that both groups (lower boxes) contain a good mix of high, average, and low scores.



5.3 Experimental Validity

Chapter 2 discussed the concept of *validity*, or the degree to which the measures used in a study capture the constructs that they were designed to capture. That is, a measure of happiness needs to capture differences in people's levels of happiness. This section returns to the subject of validity in an experimental context, assessing whether experimental results demonstrate the causal relationships that researchers think they are demonstrating. We will

discuss two types of validity that are relevant to experimental designs. The first is **internal validity**, which assesses the degree to which results can actually be attributed to the independent variables. The second is **external validity**, which assesses how well the results generalize to situations beyond the specific conditions laid out in the experiment. Taken together, internal and external validity provide a way to assess the merits of an experiment. However, each kind has its own threats and remedies, as the following sections explain.

Internal Validity

To have a high degree of internal validity, experimenters strive for maximum control over extraneous variables. That is, they try to design experiments so that the independent variable is the only cause of differences between groups. But, of course, no study is ever perfect, and some degree of error is always in place. In many cases, errors are the result of unavoidable random causes, such as the health or mood of the participants on the day of the experiment. In other cases, errors are due to factors that are, in fact, within the experimenter's control. This section focuses on several of these more manageable threats to internal validity and discusses strategies for reducing their influence.

Experimental Confounds

To avoid threats to the internal validity of an experiment, it is important to control and minimize the influence of extraneous variables that might add noise to a hypothesis test. In many cases, extraneous variables can be considered relatively minor nuisances, as when the mood experiment was inadvertently run in a depressing room. Now, though, suppose we conduct our study on temperature and mood, and due to a lack of careful planning, accidentally place all of the “warm room” participants in a sunny room, and the “cool room” participants in a windowless room. We might very well find that the warm-room participants are in a much better mood. Still, is this the result of warm temperatures or the result of exposure to sunshine? Unfortunately, we would be unable to tell the difference because of a **confounding variable** (or **confound**)—a variable that changes systematically with the independent variable. In this example, room lighting is confounded with room temperature because all of the warm-room participants are also exposed to sunshine, and all of the cool-room participants are not. This confounding combination of variables leaves us unable to determine which variable actually has the effect on mood. In other words, because our groups differ in more than one way, we cannot clearly say that the independent variable of interest (the room) caused the dependent variable (mood) to change.

This observation may seem oversimplified, but the way to avoid confounds is to be very careful in designing experiments. By ensuring that groups are alike in every way but the experimental condition, an experimenter can generally prevent confounds. Nevertheless, avoiding confounds is somewhat easier said than done because they can come from unexpected places. For example, most studies involve the use of multiple research assistants who manage data collection and interact with participants. Some of these assistants might be more or less friendly than others, so it is important to make sure each of them interacts with participants in all conditions. The friendliest assistant's always running participants in the warm-room

group, for example, would result in a confounding variable (friendly versus unfriendly assistants) between room and research assistant. Consequently, the experimenter would be unable to separate the influence of the independent variable (the room) from that of the confound (the research assistant).

Selection Bias

Internal validity can also be threatened when groups differ before the manipulation, a condition known as **selection bias**. Selection bias causes problems because these preexisting differences might be the driving factor behind the results. Imagine someone is investigating a new program that will help people stop smoking. The experimenter might decide to ask for volunteers who are ready to quit smoking and put them through a six-week program. But by asking for volunteers—a remarkably common error—the researcher gathers a group of people who are already somewhat motivated to stop smoking. Thus, it is difficult to separate the effects of the new program from the effects of this preexisting motivation.

One easy way to avoid this problem is through either random or matched random assignment. In the stop-smoking example, a researcher could still ask for volunteers, but then randomly assign these volunteers either to the new program or to a control group. Both groups consisting of people motivated to quit smoking would help to cancel out the effects of motivation. Another way to minimize selection bias is to use the same people in both conditions so that they serve as their own control. In the stop-smoking example, the experimenter could assign volunteers first to one program and then to the other. However, this approach might present a problem: Participants who successfully quit smoking in the first program would not benefit from the second program. This technique is known as a **within-subject design**, and we will discuss its advantages and disadvantages in the section “Within-Subject Designs.”

Differential Attrition

Despite researchers’ best efforts at random assignment, they could still have a biased sample at the end of a study as a result of **differential attrition**. The problem of differential attrition occurs when subjects drop out of experimental groups for different reasons. Say we are conducting a study of the effects of exercise on depression levels. We manage to randomly assign people either to one week of regular exercise or to one week of regular therapy. At first glance, it appears that the exercise group shows a dramatic drop in depression symptoms. But then we notice that about one-third of the people in this group dropped out before completing the study. Chances are we are left with the participants who are most motivated to exercise, to overcome their depression, or both. Thus, it is difficult to isolate the effects of



Digital Vision/Photodisc/Thinkstock

Friendliness of the research assistant is a variable that can affect the outcome of an experiment.

the independent variable on depression symptoms. Although we cannot prevent people from dropping out of our study, we can look carefully at those who do. In many cases, researchers can spot a pattern and use it to guide future research. For example, it may be possible to create a profile of people who dropped out of the exercise study and use this knowledge to increase retention for the next attempt.

Outside Events

As much as experimenters strive to control the laboratory environment, participants are often influenced by events in the outside world. These events—sometimes called history effects—are often large scale and include political upheavals and natural disasters. History effects threaten research because they make it difficult to tell whether participants' responses are due to the independent variable or to the historical event(s). A paper published by social psychologist Ryan Brown, now a professor at the University of Oklahoma, offers a remarkable example. Brown et al.'s paper discussed the effects of receiving different types of affirmative action as people were selected for a leadership position. The goal was to determine the best way to frame affirmative action to avoid undermining the recipient's confidence (Brown, Charnsangavej, Keough, Newman, & Rentfrow, 2000). For about a week during the data-collection process, students at the University of Texas where the study was being conducted protested on the school's main lawn about a controversial lawsuit regarding affirmative-action policies. One side effect of these protests was that participants arriving for Brown's study had to pass through a swarm of people holding signs that either denounced or supported affirmative action. These types of outside events are difficult, if not impossible, to control. But, because these researchers were aware of the protests, they made a decision to exclude data gathered from participants during the week of the protests from the study, thus minimizing the effects of outside events.

Expectancy Effects

One final set of threats to internal validity results from the influence of expectancies on people's behavior. This influence can cause trouble for experimental designs in three related ways. First, **experimenter expectancies** can cause researchers to see what they expect to see, leading to subtle bias in favor of their hypotheses. In a clever demonstration of this phenomenon, the psychologist Robert Rosenthal asked his graduate students at Harvard University to train groups of rats to run a maze (Rosenthal & Fode, 1963). He also told them that based on a pretest, the rats had been classified as either bright or dull. As might be surmised, these labels were pure fiction, but they still influenced the way that the students treated the rats. Those labeled bright were given more encouragement and learned the maze much more quickly than rats labeled dull. Rosenthal later extended this line of work to teachers' expectations of their students (Rosenthal & Jacobson, 1992) and found support for the same conclusion: People often bring about the results they expect by behaving in a particular way.

One common way to avoid experimenter expectancies is to have participants interact with a researcher who is "blind" to (i.e., unaware of) the condition in which each participant is. Blind researchers may be fully aware of the general research hypothesis, but their behavior is less likely to affect the results if they are unaware of the specific conditions. In the Rosenthal and

Fode (1963) study, the graduate students' behavior only influenced the rats' learning speed because they were aware of the labels bright and dull. If these had not been assigned, the rats would have been treated fairly equally across the conditions.

Second, participants in a research study often behave differently based on their own expectancies about the goals of the study. These expectancies often develop in response to **demand characteristics**, or cues in the study that lead participants to guess the hypothesis. In a well-known study conducted at the University of Wisconsin, psychologists Leonard Berkowitz and Anthony LePage (1967) found that participants would behave more aggressively—by delivering electric shocks to another participant—if a gun was in the room than if there were no gun present. This finding has some clear implications for gun-control policies, suggesting that the mere presence of guns increases the likelihood of gun violence. However, a common critique of this study contends that participants may have quickly clued in to its purpose and figured out how they were “supposed” to behave. That is, the gun served as a demand characteristic, possibly making participants act more aggressively because they thought the researchers expected them to do so.

To minimize demand characteristics, researchers use a variety of techniques, all of which attempt to hide the true purpose of the study from participants. One common strategy is to use a **cover story**, or a misleading statement about what is being studied. Chapter 1 (1.3) discussed Milgram's famous obedience studies, which discovered that people were willing to obey orders to deliver dangerous levels of electric shocks to other people. To disguise the purpose of the study, Milgram described it to participants as a study of punishment and learning. To give another example, Ryan Brown and colleagues (2000) presented their affirmative-action study as a study of leadership styles. These cover stories aimed to give participants a compelling explanation for what they experienced during the study and to direct their attention away from the research hypothesis.

Another strategy for avoiding demand characteristics is to use the **unrelated-experiments technique**, which leads participants to believe that they are completing two different experiments during one laboratory session. The experimenter can use this bit of deception to present the independent variable during the first experiment and then measure the dependent variable during the second experiment. For example, a study by Harvard psychologist Margaret Shih and colleagues (Shih, Pittinsky, & Ambady, 1999) recruited Asian-American females and asked them to complete two supposedly unrelated studies. In the first, they were asked to read and form impressions of one of two magazine articles; these articles were designed to make them focus on either their Asian-American identity or their female identity. In the second experiment, they were asked to complete a math test as quickly as possible. The goal of this study was to examine the effects of priming different aspects of identity on math performance. Based on previous research, these authors predicted that priming an Asian-American identity would remind participants of positive stereotypes regarding Asians and math performance, whereas priming a female identity would remind participants of negative stereotypes regarding women and math performance. As researchers expected, priming an Asian-American identity led this group of participants to do better on a math test than did priming a female identity. The unrelated-experiments technique was especially useful for this study because it kept participants from connecting the independent variable (magazine article prime) with the dependent variable (math test).

A final way in which expectancies shape behavior is the **placebo effect**, meaning that change can result from the mere expectation that change will occur. Imagine we want to test the hypothesis that alcohol causes people to become aggressive. One relatively easy way to do this would be to give alcohol to a group of volunteers (aged 21 and older) and then measure how aggressively they behave in response to being provoked. The problem with this approach is that people also expect alcohol to change their behavior, and so we might see changes in aggression simply because of these expectations. Fortunately, the problem has an easy solution: add a **placebo control** group to the study that mimics the experimental condition in every way but one. In this case, we might tell all participants that they will be drinking a mix of vodka and orange juice but only add vodka to half of the participants' drinks. The orange-juice-only group serves as our placebo control. Any differences between this group and the alcohol group can be attributed to the alcohol itself.



Martin Poole/Digital Vision/Thinkstock

The placebo effect can test whether alcohol affects behavior, or whether people just expect it to and exhibit changed behavior based on their expectations.

External Validity

To have a high degree of external validity in experiments, researchers strive for maximum realism in the laboratory environment. External validity means that the results extend beyond the particular set of circumstances created in a single study. Recall that science is a cumulative discipline and that knowledge grows one study at a time. Thus, each study is more meaningful: 1) to the extent that it sheds light on a *real phenomenon*; and 2) to the extent that the results *generalize to other studies*. This section examines each of these criteria separately.

Mundane Realism

The first component of external validity is the extent to which an experiment captures the real-world phenomenon under study. Inspired by a string of school shootings in the 1990s, one popular question in the area of aggression research asks whether rejection by a peer group leads to aggression. That is, when people are rejected from a group, do they lash out and behave aggressively toward the members of that group? Researchers must find realistic ways to manipulate rejection and measure aggression without infringing on participants' welfare. Given the need to strike this balance, how real can conditions be in the laboratory? How do we study real-world phenomena without sacrificing internal validity?

The answer is to strive for **mundane realism**, meaning that the research replicates the psychological conditions of the real-world phenomenon (sometimes referred to as *ecological*

validity). In other words, we need not recreate the phenomenon down to the last detail; instead, we aim to make the laboratory setting feel like the real-world phenomenon. Researchers studying aggressive behavior and rejection have developed some rather clever ways of doing this, including allowing participants to administer loud noise blasts or serve large quantities of hot sauce to those who reject them. Psychologically, these acts feel like aggressive revenge because participants are able to lash out against those who rejected them—with the intent of causing harm—even though the behaviors themselves may differ from the ways people exact revenge in the real world.

In a 1996 study, Tara MacDonald and her colleagues at Queen's University in Ontario, Canada, examined the relationship between alcohol and condom use (MacDonald, Zanna, & Fong, 1996). The authors were intrigued by a puzzling set of real-world data: Most people self-reported that they would use condoms when engaging in casual sex, but actual rates of unprotected sex (i.e., having sexual intercourse without a condom) were also remarkably high. In this study, the authors found that alcohol was a key factor in causing “common sense to go out the window” (p. 763), resulting in a decreased likelihood of condom use. But how on earth might they study this phenomenon in the laboratory? In the authors' words, “even the most ambitious of scientists would have to conclude that it is impossible to observe the effects of intoxication on actual condom use in a controlled laboratory setting” (p. 765).

To solve this dilemma, MacDonald and colleagues developed a clever technique for studying people's intentions to use condoms. Participants were randomly assigned to either an alcohol or placebo condition, and then they viewed a video depicting a young couple faced with the dilemma of whether to have unprotected sex. At the key decision point in the video, the tape was stopped and participants were asked what they would do in the situation. As predicted, participants who were randomly assigned to consume alcohol said they would be more willing to proceed with unprotected sex. While this laboratory study does not capture the full experience of making decisions about casual sex, it does a pretty nice job of capturing the psychological conditions involved.

Generalizing Results

The second component of external validity, **generalizability**, refers to the extent to which the results extend to other studies by using a wide variety of populations and a wide variety of operational definitions (sometimes referred to as *population validity*). If we conclude that rejection causes people to become more aggressive, for example, this conclusion should ideally carry over to other studies of the same phenomenon, studies that use different ways of manipulating rejection and different ways of measuring aggression. If we want to conclude that alcohol reduces the intention to use condoms, we would need to test this relationship in a variety of settings—from laboratories to nightclubs—using different measures of intentions.

Thus, each single study researchers conduct is limited in its conclusions. For a particular idea to take hold in the scientific literature, it must be replicated, or repeated in different contexts. **Replication** can take one of four forms. First, **exact replication** involves trying to recreate the original experiment as closely as possible to verify the findings. This type of replication is often the first step following a surprising result, and it helps researchers to gain more confidence in the patterns.

The second and much more common method, **conceptual replication**, involves testing the relationship between conceptual variables using new operational definitions. Conceptual replications would include testing aggression hypotheses using new measures or examining the link between alcohol and condom use in different settings. For example, rejection might be operationalized in one study by having participants be chosen last for a group project. A conceptual replication might take a different approach, operationalizing rejection by having participants be ignored during a group conversation or voted out of the group. Likewise, a conceptual replication might change the operationalization of aggression, with one study measuring the delivery of loud blasts of noise and another measuring the amount of hot sauce that people give to their rejecters. Each variation studies the same concept (aggression or rejection) but uses slightly different operationalizations. If all of these variations yield similar results, this further supports the underlying ideas—in this case, that rejection causes people to be more aggressive.

The third method, **participant replication**, involves repeating the study with a new population of participants. These types of replication are usually driven by a compelling theory as to why the two populations differ. For example, we might reasonably hypothesize that the decision to use condoms is guided by a different set of considerations among college students than among older, single adults. Or, we might hypothesize that different cultures around the world might have different responses to being rejected from a group.

Finally, **constructive replication** re-creates the original experiment but adds elements to the design. These additions are typically designed to either rule out alternative explanations or extend knowledge about the variables under study. Our rejection and aggression example might compare the impact of being rejected by a group versus by an individual.

Internal Versus External Validity

This chapter has focused on two ways to assess validity in the context of experimental designs. Internal validity assesses the degree to which results can be attributed to independent variables; external validity assesses how well results generalize beyond the specific conditions of the experiment. In an ideal world, studies would have a high degree of both of these. That is, we would feel completely confident that our independent variable was the only cause of differences in our dependent variable, and our experimental paradigm would perfectly capture the real-world phenomenon under study.

Reality, though, often demands a trade-off between internal and external validity. In MacDonald et al.'s (1996) study on condom use, the researchers sacrificed some realism in order to conduct a tightly controlled study of participants' intentions. In Berkowitz and LePage's (1967) study on the effect of weapons, the researchers risked the presence of a demand characteristic in order to study reactions to actual weapons. These types of trade-offs are always made based on the goals of the experiment.

Research: Applying Concepts

Balancing Internal Versus External Validity

To give you a better sense of how researchers make the compromises involving internal and external validity, consider the following fictional scenarios.

Scenario 1—Time Pressure and Stereotyping

Dr. Bob is interested in whether people are more likely to rely on stereotypes when they are in a hurry. In a well-controlled laboratory experiment, he asks participants to categorize ambiguous shapes as either squares or circles, and half of these participants are given a short time limit to accomplish the task. The independent variable is the presence or absence of time pressure, and the dependent variable is the extent to which people use stereotypes in their classification of ambiguous shapes. Dr. Bob hypothesizes that people will be more likely to use stereotypes when they are in a hurry because they will have fewer cognitive resources to consider carefully all aspects of the situation. Dr. Bob takes great care to have all participants meet in the same room. He uses the same research assistant every time, and the study is always conducted in the morning. Consistent with his hypothesis, Dr. Bob finds that people seem to use shape stereotypes more under time pressure.

The internal validity of this study appears high—Dr. Bob has controlled for other influences on participants' attention span by collecting all of his data in the morning. He has also minimized error variance by using the same room and the same research assistant. In addition, Dr. Bob has created a tightly controlled study of stereotyping through the use of circles and squares. Had he used photographs of people (rather than shapes), the attractiveness of these people might have influenced participants' judgments. The study, however, has a trade-off: By studying the social phenomenon of stereotyping using geometric shapes, Dr. Bob has removed the social element of the study, thereby posing a threat to mundane realism. The psychological meaning of stereotyping shapes is rather different from the meaning of stereotyping people, which makes this study relatively low in external validity.

Scenario 2—Hunger and Mood

Dr. Jen is interested in the effects of hunger on mood; not surprisingly, she predicts that people will be happier when they are well fed. She tests this hypothesis with a lengthy laboratory experiment, requiring participants to be confined to a laboratory room for 12 hours with very few distractions. Participants have access to a small pile of magazines to help pass the time. Half of the participants are allowed to eat during this time, and the other half is deprived of food for the full 12 hours. Dr. Jen—a naturally friendly person—collects data from the food-deprivation groups on a Saturday afternoon, while her grumpy research assistant, Mike, collects data from the well-fed group on a Monday morning. Her independent variable is food deprivation, with participants either not deprived of food or deprived for 12 hours. Her dependent variable consists of participants' self-reported mood ratings. When Dr. Jen analyzes the data, she is shocked to discover that participants in the food-deprivation group are much happier than those in the well-fed group.

(continued on next page)

Compared to our first scenario, this study seems high on external validity. To test her predictions about food deprivation, Dr. Jen actually deprives her participants of food. One possible problem with external validity is that participants are confined to a laboratory setting during the deprivation period with only a small pile of magazines to read. That is, participants may be more affected by hunger when they do not have other things to distract them. In the real world, people are often hungry but distracted by paying attention to work, family, or leisure activities. Dr. Jen, though, has sacrificed some external validity for the sake of controlling how participants spend their time during the deprivation period. The larger problem with her study has to do with internal validity. Dr. Jen has accidentally confounded two additional variables with her independent variable: Participants in the deprivation group have a different experimenter and data are collected at a different time of day. Thus, Dr. Jen's surprising results most likely reflect that everyone is in a better mood on Saturday than on Monday and that Dr. Jen is more pleasant to spend 12 hours with than Mike is.

Scenario 3—Math Tutoring and Graduation Rates

Dr. Liz is interested in whether specialized math tutoring can help increase graduation rates among female math majors. To test her hypothesis, she solicits female volunteers for a math-skills workshop by placing fliers around campus, as well as by sending email announcements to all math majors. The independent variable is whether participants are in the math skills workshop, and the dependent variable is whether participants graduate with a math degree. Those who volunteer for the workshop are given weekly skills tutoring, along with informal discussion groups designed to provide encouragement and increase motivation. At the end of the study, Liz is pleased to see that participants in the workshops are twice as likely as nonparticipants to stick with the major and graduate.

The obvious strength of this study is its external validity. Dr. Liz has provided math tutoring to math majors, and she has observed a difference in graduation rates. Thus, this study is very much embedded in the real world. However, this external validity comes at a cost to internal validity. The study's biggest flaw is that Dr. Liz has recruited *volunteers* for her workshops, resulting in selection bias for her sample. People who volunteer for extra math tutoring are likely to be more invested in completing their degree and might also have more time available to dedicate to their education. Dr. Liz would also need to be mindful of how many people drop out of her study. If significant numbers of participants withdraw, she could have a problem with differential attrition, so that the most motivated people stayed with the workshops. Dr. Liz can fix this study with relative ease by asking for volunteers more generally and then randomly assigning these volunteers to take part in either the math tutoring workshops or a different type of workshop. While the sample might still be less than random, Dr. Liz would at least have the power to assign participants to different groups.

5.4 Experimental Design

The process of designing experiments boils down to deciding *what* to manipulate and *how* to do it. This section covers two broad issues related to experimental design: deciding how to structure the levels, or different versions of an independent variable, and deciding on the *number* of independent variables necessary to test the hypotheses. While these decisions may seem tedious, they are at the crux of designing successful experiments, and are, therefore, the key to performing successful tests of hypotheses.

Levels of the Independent Variable

The primary goal in designing experiments is to ensure that the levels of independent variables are equivalent in every way but one. This is what allows researchers to make causal statements about the effects of that single change. These levels can be formed in one of two broad ways: representing two distinct groups of people or representing the same group of people over time.

Between-Subject Designs

In most of the examples discussed so far, the levels of independent variables have represented two distinct groups—participants are in either the control group or the experimental group. This type of design is referred to as a **between-subject design** because the levels differ between one subject and the next. Each participant who enrolls in the experiment is exposed to only one level of the independent variable—for example, *either* the experimental *or* the control group. Most of the examples so far have been illustrations of between-subject designs: participants receive either alcohol or a placebo; students read an article designed to prime either their Asian or their female identity; and graduate students train rats that are falsely labeled either bright or dull. The “either-or” between-subject approach is common and has the advantage of using distinct groups to represent each level of the independent variable. In other words, participants who are asked to consume alcohol are completely distinct from those asked to consume the placebo drink. However, the between-subject approach is only one option for structuring the levels of the independent variable. This section examines two additional ways to structure these levels.

Within-Subject Designs

In some cases, the levels of the independent variable can represent the same participants at different time periods. This type of design is referred to as a **within-subject design** because the levels differ within individual participants. Each participant who enrolls in the experiment would be exposed to all levels of the independent variable. That is, every participant would be in *both* the experimental *and* the control group. Within-subject designs are often used to compare changes over time in response to various stimuli. For example, a researcher might measure anxiety symptoms before and after people are locked in a room with a spider, or measure depression symptoms before and after people undergo drug treatment.

Within-subject designs have two main advantages over between-subject designs. First, because the same people constitute both levels of the IV, these designs require fewer participants. Suppose we decide to collect data from 20 participants at each level of an IV. In a between-subject design with three levels, we would need 60 people. However, if we run the same experiment as a within-subject design—exposing the same group of people to three different sets of circumstances—we would need only 20 people. Thus, within-subject designs are often a good way to conserve resources.

Second, participants also serve as their own control group, allowing the researcher to minimize a major source of error variance. Remember that one key feature of experimental design is the researcher’s power to assign people to groups to distribute subject differences randomly across the levels of the IV. Using a within-subject design solves the problem of subject

differences in another way, by examining changes within people. For instance, in the study of spiders and anxiety, some participants are likely to have higher baseline anxiety than others. By measuring changes in anxiety in the same group of people before and after spider exposure, we are able to minimize the effects of individual differences.

Disadvantages of Within-Subject Designs

Within-subject designs also have two clear disadvantages compared to between-subject designs. First, they pose the risk of **carryover effects**, in which the effects of one level are still present when another level is introduced. Because the same people are exposed to all levels of the IV, it can be difficult to separate the effects of one level from the effects of the others. One common paradigm in emotion research is to show participants several film clips that elicit different types of emotion. People might view one clip showing a puppy playing with a blanket, another showing a child crying, and another showing a surgical amputation. Even without seeing these clips in full color, we can imagine that it would be hard to shake off the disgust triggered by the amputation to experience the joy triggered by the puppy.

When researchers use a within-subject design, they take steps to minimize carryover effects. In studies of emotion, for example, researchers typically show a brief neutral clip—like waves rolling onto a beach—after each emotional clip, so that participants experience each emotion after viewing a benign image. Another simple technique is to collect data from the baseline control condition first whenever possible. In the study of spiders and anxiety, it would be important to measure baseline anxiety at the start of the experiment before exposing people to spiders. Once people have been surprised by a spider, it will be hard to get them to relax enough to collect control ratings of anxiety.



Wavebreakmedia/iStock/Thinkstock

Carryover effects can be understood through the example of monitoring people's reactions to different film clips. How they feel about one image may influence how they react to the next image.

Second, within-subject designs risk **order effects**, meaning that the order in which levels are presented can moderate their effects. Order effects fall into two categories. The **practice effect** happens when participants' performance improves over time simply due to repeated attempts. This is a particular problem in studies that examine learning. Say we use a within-subject design to compare two techniques for teaching people to solve logic problems. Participants would learn technique A, then take a logic test, then learn technique B, and then take a second logic test. The possible problem is that participants will have had more opportunities to practice logic problems by the time they take the second test. This makes it difficult to separate the effects of practicing the logic problems from the effects of using different teaching techniques.

The flipside of practice effects is the phenomenon of the **fatigue effect**, which happens when participants' performance decreases over time due to repeated testing. Imagine running a variation of the above experiment, teaching people different ways to improve their reaction time. Participants might learn each technique and have their reaction time tested several times after each one. The problem is that people gradually start to tire, and their reaction times slow down due to fatigue. Thus, it would be difficult to separate the effects of fatigue from the effects of the different teaching techniques.

The result of both types of order effects is in confounding the order of presentation with the level of the independent variable. Fortunately, researchers have a relatively easy way to avoid both carryover and fatigue effects: a process called **counterbalancing**. Counterbalancing involves varying the order of presentation to groups of participants. The simplest approach is to divide participants into as many groups as combinations of levels in the experiment. That is, we create a group for each possible order, allowing us to identify the effects of encountering the conditions in different orders. In the examples above, the learning experiments involved two techniques, A and B. To counterbalance these techniques across the study, we divide the participants into two groups. We expose one group to A and then B; we expose the other group to B and then A. When it is time to analyze the data, we will be able to examine the effects of both presentation order and teaching technique. If the order of presentation made a difference, then the A/B group would differ from the B/A group in some way.

Mixed Designs

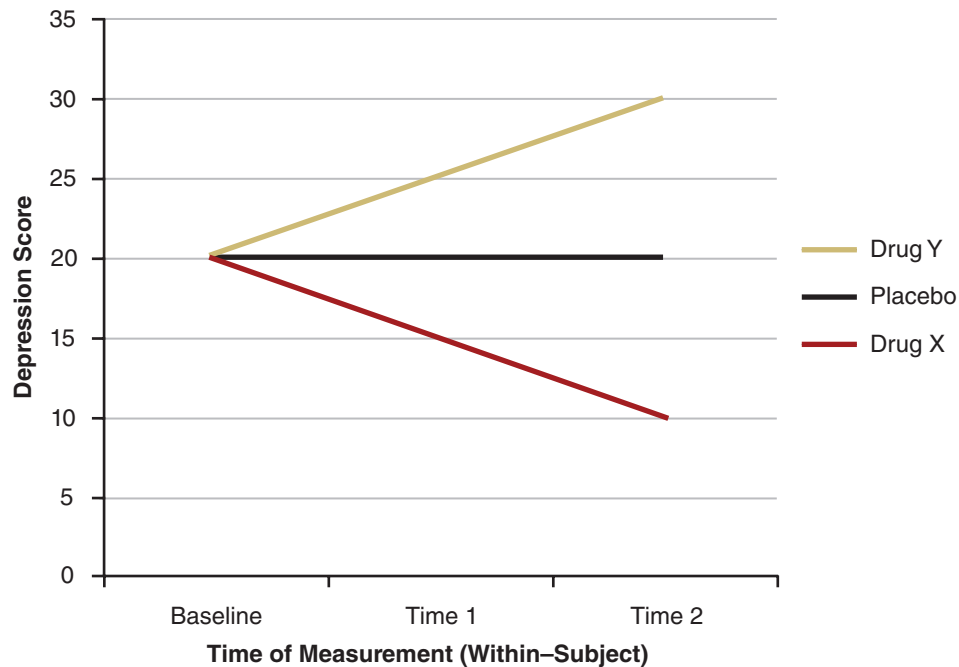
The third common way to structure the levels of an IV is using a **mixed design**, which contains at least one between-subject variable and at least one within-subject variable. So, in the previous example, participants would be exposed to *both* teaching techniques (A and B) but in only *one* order of presentation. In this case, teaching technique is a within-subject variable because participants experience both levels, and presentation order is a between-subject variable because participants experience only one level. Because we have one of each in the overall experiment, it is a mixed design.

Studies that compare the effects of different drugs commonly use mixed designs. Imagine we want to compare three new drugs—Drug X, Drug Y, and a placebo control—to determine which has the strongest effects on reducing depression symptoms. To perform this study, we would want to measure depression symptoms on at least three occasions: before starting drug treatment, after a few months of taking the drug, and then again after a few months of stopping the drug (to assess relapse rates). So, our participants would be given one of three possible drugs and then measured at each of three time periods. In this mixed design, measurement time is a *within*-subject variable because participants are measured at *all* possible times, while the drug is a between-subject variable because participants experience only *one* of three possible drugs.

Figure 5.4 shows the hypothetical results of this study. Observe that the placebo pill has no effect on depression symptoms; depression scores in this group are the same at all three measurements. Drug X appears to cause significant improvement in depression symptoms; depression scores drop steadily across measurements in this group. Strangely, Drug Y seems

to make depression worse; depression scores increase steadily across measurements in this group. The mixed design allows us both to track people over time and to compare different drugs in one study.

Figure 5.4: Example of a mixed-subjects design



Research: Thinking Critically

Outwalking Depression

Follow the link below to an article from *Psychology Today*, describing a 2011 research study from the *Journal of Psychiatric Research*. The study provides new evidence of the benefits of exercise for people with depression. As you read the article, consider what you have learned so far about the research process, and then respond to the questions below.

<https://www.psychologytoday.com/blog/exercise-and-mood/201107/outwalking-depression>

Think About It

1. Identify the following essential aspects of this experimental design:
 - a) What are the IV and DV in this study?
 - b) How many levels does the IV have?
 - c) Is this a between-subjects, within-subjects, or mixed design?
 - d) Draw a simple table labeling each condition.

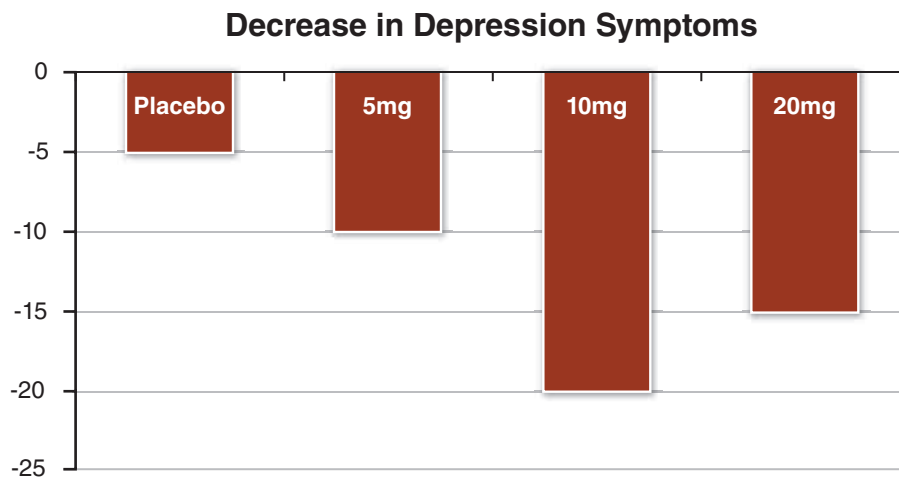
2.
 - a) What preexisting differences between groups should the researchers be sure to take into account? Name as many as you can.
 - b) How should the researchers assign participants to the conditions in order to ensure that preexisting differences cannot account for the results?
3. How might expectancy effects influence the results of this study? Can you think of any ways to control for this?
4. Briefly state how you would replicate this study in each of the following ways:
 - a) exact replication
 - b) conceptual replication
 - c) participant replication
 - d) constructive replication

One-Way Versus Factorial Designs

The second big issue in creating experimental designs is to decide how many independent variables to manipulate. In some cases, we can test our hypotheses by manipulating a single IV and measuring the outcome—such as giving people either alcohol or a placebo drink and measuring the intention to use condoms. In other cases, hypotheses involve more complex combinations of variables. Earlier, the chapter discussed research findings that people tend to act more aggressively after a peer group has rejected them—a single independent variable. Researchers could, however, extend this study and ask what happens when people are rejected by members of the same sex versus members of the opposite sex. We could go one step further and test whether the attractiveness of the rejecters matters, for a total of three independent variables. These examples illustrate two broad categories of experimental design, known as one-way and factorial designs.

One-Way Designs

If a study involves assigning people to either an experimental or control group and measuring outcomes, it has a **one-way design**, or a design that has only one independent variable with two or more levels to the variable. These tend to be the simplest experiments and have the advantage of testing manipulations in isolation. The majority of drug studies use one-way designs. These types of study compare the effects on medical outcomes for people randomly assigned, for instance, to take the antidepressant drug Prozac or a placebo. Note that a one-way design can still have multiple levels—in many cases it is preferable to test several different doses of a drug. So, for example, we might test the effects of Prozac by assigning people to take doses of 5 mg, 10 mg, 20 mg, or a placebo control. The independent variable would be the drug dose, and the dependent variable would be a change in depression symptoms. This one-way design would allow us to compare all three of the drug doses to a placebo control, as well as to test the effects of varying doses of the drug. Figure 5.5 shows hypothetical results from this study. We can see that even those receiving the placebo showed a drop in depression symptoms, with the 10-mg dose of Prozac producing the maximum benefit.

Figure 5.5: Comparing drug doses in a one-way design

Factorial Designs

Despite the appealing simplicity of one-way designs, experiments conducted in the field of psychology with only one IV are relatively rare. The real world is much more complicated, so studies that focus on people's thoughts, feelings, and behaviors must somehow capture this complexity. Thus, the rejection-and-aggression example above is not that farfetched. If a researcher wanted to manipulate the occurrence of rejection, the gender of the rejecters, and the attractiveness of the rejecters in a single study, the experiment would have a **factorial design**. Factorial designs are those that have two or more independent variables, each of which has two or more levels. When experimenters use a factorial design, their purpose is to observe both the effects of individual variables and the combined effects of multiple variables.

Factorial designs have their own terminology to reflect the fact that they include both individual variables and combinations of variables. The beginning of this chapter explained that the versions of an independent variable are referred to as both levels and conditions, with a subtle difference between the two. This difference becomes relevant to the discussion of factorial designs. Specifically, *levels* refer to the versions of each IV, while *conditions* refer to the groups formed by combinations of IVs. Consider one variation of the rejection-and-aggression example from this perspective: The first IV has two levels because participants are either rejected or not rejected. The second IV also has two levels because members of the same sex or the opposite sex do the rejecting. To determine the number of conditions in this study, we calculate the number of different experiences that participants can have in the study. This is a simple matter of multiplying the levels of separate variables, so two multiplied by two, for a total of four conditions.

Researchers also have a way to quickly describe the number of variables in their design: A two-way design has two independent variables; a three-way design has three independent variables; an eight-way design has eight independent variables, and so on. Even more useful, the system of **factorial notation** offers a simple way to describe both the number of variables

and the number of levels in experimental designs. For instance, we might describe our design as a 2×2 (pronounced “two by two”), which instantly communicates two things: (1) the study uses two independent variables, indicated by the presence of two separate numbers and (2) each IV has two levels, indicated by the number 2 listed for each one.

The 2×2 Design

One of the most common factorial designs also happens to be the simplest one—the 2×2 design. As noted above, these designs have two independent variables, with two levels each, for a total of four experimental conditions. The simplicity of these designs makes them a useful way to become more comfortable with some of the basic concepts of experiments. This section will explore an example of a 2×2 and analyze it in detail.

Beginning in the late 1960s, social psychologists developed a keen interest in understanding the predictors of helping behavior. This interest was inspired, in large part, by the tragedy of Kitty Genovese, who was killed outside her apartment building while none of her neighbors called the police (Gansberg, 1964). As Chapter 2 (2.1) discussed, in one representative study, Princeton psychologists John Darley and Bibb Latané examined people’s likelihood of responding to a staged emergency. Participants were led to believe that they were taking part in a group discussion over an intercom system, but in reality, all of the other participants were prerecorded. The key independent variable was the number of other people supposedly present, ranging from two to six. A few minutes into the conversation, one participant appeared to have a seizure. The recording went like this (actual transcript; Darley & Latané, 1968):

I could really-er-use some help so if someone would-er-give me a little h-hel-puh-er-er-er c-could somebody er-er-hel-er-uh-uh-uh [*choking sounds*] . . . I’m gonna die-er-er-I’m . . . gonna die-er-hel-er-er-seizure-er [*chokes, then quiet*].

What do people do in this situation? Do they help? How long does it take? Darley and Latané discovered that two things happen as the group became larger: People were less likely to help at all, and those who did help took considerably longer to do so. Researchers concluded from this and other studies that people are less likely to help when other people are present because the responsibility for helping is “diffused” among the members of the crowd (Darley & Latané, 1968).

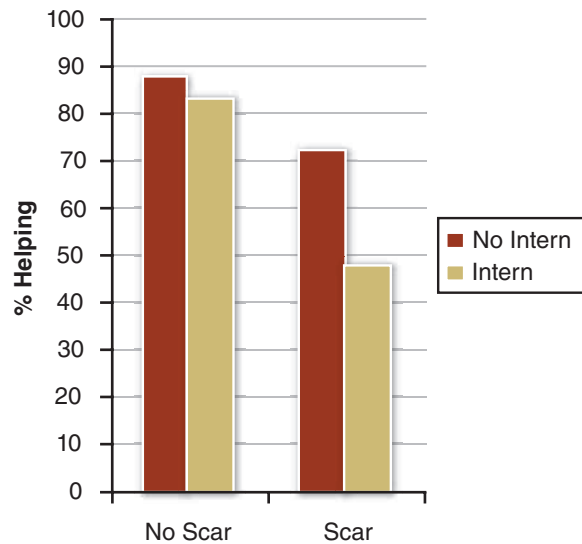
Building on this earlier conclusion, the sociologist Jane Piliavin and her colleagues (Piliavin, Piliavin, & Rodin, 1975) explored the influence of two additional variables on helping behavior. The experimenters staged an emergency on a New York City subway train in which a person who was in on the study appeared to collapse in pain. Piliavin and her team manipulated two variables in their staged emergency. The first independent variable was the presence or absence of a nearby medical intern, who could be easily identified in blue scrubs. The second independent variable was the presence or absence of a large disfiguring scar on the victim’s face. The combination of these variables resulted in four conditions, as Table 5.2 shows. The dependent variable in this study was the percentage of people taking action to help the confederate.

Table 5.2: 2×2 Design of the Piliavin et al. study

	No intern	Intern
No scar	1	2
Scar	3	4

The authors predicted that bystanders would be less likely to help if a perceived medical professional was nearby since he or she was considered more qualified to help the victim. They also predicted that people would be less likely to help when the confederate had a large scar because previous research had demonstrated convincingly that people avoid contact with those who are disfigured or have other stigmatizing conditions (e.g., Goffman, 1963). As Figure 5.6 reveals, the results supported these hypotheses. Both the presence of a scar and the presence of a perceived medical professional reduced the percentage of people who came to help. Nevertheless, something else is apparent in these results: When the confederate was not scarred, having an intern nearby led to a small decrease in helping (from 88% to 84%). However, when the confederate had a large facial scar, having an intern nearby decreased helping from 72% to 48%. In other words, it seems these variables are having a *combined* effect on helping behavior. The next section examines these combined effects more closely.

**Figure 5.6: Sample 2×2 design:
Results from Piliavan et al. (1975)**



Piliavan et al. (1975)

Main Effects and Interactions

When experiments involve only one independent variable, the analyses can be as simple as comparing two group means—as did the example in Chapter 1, which compared the happiness levels of couples with and without children. But what about cases where the design has more than one independent variable?

A factorial design has two types of effects: A **main effect** refers to the effect of each independent variable on the dependent variable, averaging values across the levels of other variables. A 2×2 design has two main effects; a $2 \times 2 \times 2$ design has three main effects because there are three IVs. An **interaction** occurs when the variables have a combined effect; that is, the effects of one IV are different depending on the levels of the other IV. So, applying this new terminology to the Piliavin et al. (1975) “subway emergency” study, produces three possible results (“possible,” because we would need to use statistical analyses to verify them):

1. *The main effect of scar:* Does the presence of a scar affect helping behavior?
Yes. More people help in absence of a facial scar. Figure 5.6 indicates that the bars on the left (no scar) are, on average, higher than those on the right (scar).

2. *The main effect of intern:* Does the presence of an intern affect helping behavior? Yes. More people help when no medical intern is on hand. Note that in Figure 5.6, the red bars (no intern) are, on average, higher than the tan bars (intern).
3. *The interaction between scar and intern:* Does the effect of one variable depend on the effect of another variable? Yes. Refer to Figure 5.6 and observe that the presence of a medical intern matters more when the victim has a facial scar. In visual terms, the gap between red and tan bars is much larger in the bars on the right. This indicates an interaction between scar and intern.

Consider a fictional example. Imagine we are interested in people's perceptions of actors in different types of movies. We might predict that some actors are better suited to comedy and others are better suited to action movies. A simple experiment to test this hypothesis would show four movies in a 2×2 design, using the same two actors in two movies (for a total of four conditions). The first IV would be the movie type, with two levels: action and comedy. The second IV would be the actor, with two levels: Will Smith and Arnold Schwarzenegger. The dependent variable would be the ratings of each movie on a 10-point scale. This design produces three possible results:

1. *The main effect of actor:* Do people generally prefer Will Smith or Arnold Schwarzenegger, regardless of the movie?
2. *The main effect of movie type:* Do people generally prefer action or comedy movies, regardless of the actor?
3. *The interaction between actor and movie type:* Do people prefer each actor in a different kind of movie? (i.e., are ratings affected by the combination of actor and movie type?)

After collecting data from a sample of participants, we end up with the following average ratings for each movie, which Table 5.3 shows.

Table 5.3: Main effects and marginal means: the actor study

	Arnold	Will	Marginal Mean
Action	6	1.5	3.75
Comedy	5	8	6.5
Marginal Mean	5.5	4.75	

These two means let us compare movie ratings, ignoring the actor. People have a slight preference for Comedy (6.5) over Action (3.75).

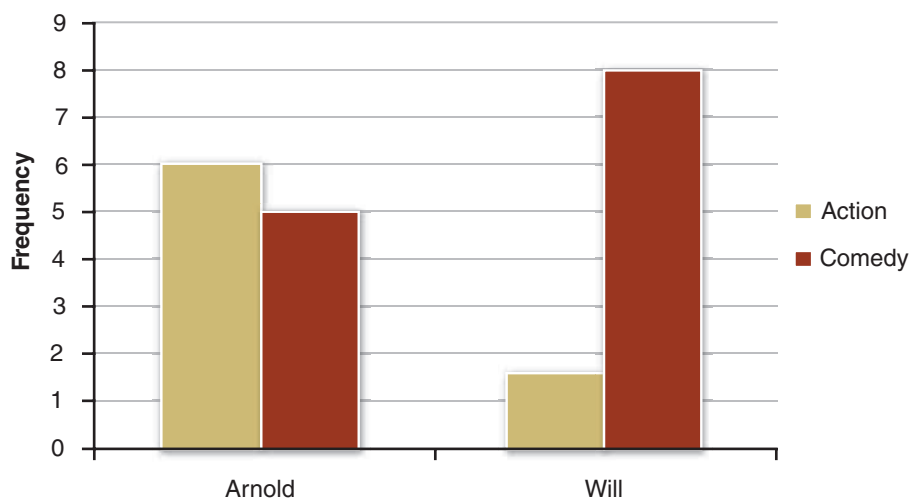
These two means let us compare actor ratings, ignoring the movie type. People have a slight preference for Arnold (5.5) over Will (4.75).

Remember that main effects represent the effects of one IV, averaging across the levels of the other IV. To average across levels, we calculate the **marginal means**, or the combined mean across levels of another factor. In other words, the marginal mean for action movies is calculated by averaging together the ratings of both Arnold Schwarzenegger and Will Smith

in action movies. The marginal mean for Arnold Schwarzenegger is calculated by averaging together ratings of Arnold Schwarzenegger in both action and comedy movies. Performing these calculations for our 2×2 design results in four marginal means, which are presented alongside the participant ratings in Table 5.3. To verify these patterns would require statistical analyses, but it appears that people have a slight preference for comedy over action movies, as well as a slight preference for Arnold Schwarzenegger's acting over Will Smith's acting.

What about the interaction? The main hypothesis here posits that some actors perform best in some genres of movies (e.g., action or comedy) than they do in other genres, which suggests that the actor and the movie type have a *combined effect* on people's ratings of the movies. Examining the means in Table 5.3 conveys a sense of this finding, but it is much easier to appreciate in a graph. Figure 5.7 shows the mean of participants' ratings across the four conditions. If we focus first on the ratings of Arnold Schwarzenegger, we can see that participants did have a slight preference for him in action (6) versus comedy (5) roles. Then, examining ratings of Will Smith, we can see that participants had a strong preference for him in comedy (8) versus action (1.5) roles. Together, this set of means indicates an interaction between actor and movie type because the effects of one variable depend on another. In plain English: People's perceptions of an actor depend on the type of movie in which he or she performs. This pattern of results nicely fits for the hypothesis that certain actors are better suited to certain types of movie: Arnold should probably stick to action movies, and Will should definitely stick to comedies.

Figure 5.7: Interaction in the actor study



Before moving on to the logic of analyzing experiments, consider one more example from a published experiment. A large body of research in social psychology suggests that stereotypes can negatively affect performance on cognitive tasks (e.g., tests of math and verbal skills). According to Stanford social psychologist Claude Steele and his colleagues, individuals' fear of confirming negative stereotypes about their group acts as a distraction. This distraction—which the researchers term *stereotype threat*—makes it hard to concentrate and perform well, and thus leads to lower scores on a cognitive test (Steele, 1997). One of the primary

implications of this research is that ethnic differences in standardized-test scores can be viewed as a situational phenomenon—change the situation, and the differences go away. In the first published study of stereotype threat, Claude Steele and Josh Aronson (1995) found that when African-American students at Stanford were asked to indicate their race before taking a standardized test, this was enough to remind them of negative stereotypes, and they performed poorly. When the testing situation was changed, however, and participants were no longer asked their race, the students performed at the same level as Caucasian students. Worth emphasizing is that these were Stanford students and had therefore met admissions standards for one of the best universities in the nation. Even this group of elite students was susceptible to situational pressure but performed at their best when the pressure was eliminated.

In a great application of stereotype threat, social psychologist Jeff Stone at the University of Arizona asked both African-American and Caucasian college students to try their hands at putting on a golf course (Stone, Lynch, Sjomeling, & Darley, 1999). Putting was described as a test of natural athletic ability to half of the participants and as a test of sports intelligence to the other half. Thus, the experiment had two independent variables: the race of the participants (African-American or Caucasian) and the description of the task (“athletic ability” or “sports intelligence”). Note that “race” in this study is technically a quasi-independent variable because it is not manipulated. This design resulted in a total of four conditions, and the dependent variable was the number of putts that participants managed to make. Stone and colleagues hypothesized that describing the task as a test of athletic ability would lead Caucasian participants to worry about the stereotypes regarding their poor athletic ability. In contrast, describing the task as a test of intelligence would lead African-American participants to worry about the stereotypes regarding their lower intelligence.

Consistent with their hypotheses, Stone and colleagues found an interaction between race and task description but no main effects. That is, neither race was better at the putting task overall, and neither task description had an overall effect on putting performance. The combination of these variables, though, proved fascinating. When researchers described the task as measuring sports intelligence, the African-American participants did poorly due to fear of confirming negative stereotypes about their overall intelligence. Conversely, when researchers described the task as measuring natural athletic ability, the Caucasian participants did poorly due to fear of confirming negative stereotypes about their athleticism. This study beautifully illustrates an interaction; the effects of one variable (*task description*) depend on the effects of another (*race of participants*). The results further confirm the power of the situation: Neither group did better or worse overall, but both were responsive to a situationally induced fear of confirming negative stereotypes.



Comstock Images/Stockbyte/Thinkstock

Skill on the golf course was used to study stereotypes in an experiment conducted by Jeff Stone at the University of Arizona.

5.5 Analyzing Data From Experiments

So far, we have been drawing conclusions about experimental findings using conceptual terms. But naturally, before we actually make a decision about the status of our hypotheses, we have to conduct statistical analyses. This section provides a conceptual overview of the most common statistical techniques for analyzing experimental data.

Dealing With Multiple Groups

Why do researchers need a special technique for experimental designs? After all, we learned in Chapter 2 (2.4) that we can compare two pairs of means using a t test; why not use several t tests to analyze our experimental designs? For the movie ratings study, we could analyze the data using a total of six t tests to capture every possible pair of means:

- Arnold Schwarzenegger in a comedy versus Will Smith in a comedy;
- Arnold Schwarzenegger in an action movie versus Will Smith in an action movie;
- Arnold Schwarzenegger in a comedy versus an action movie;
- Will Smith in a comedy versus an action movie;
- Will Smith in a comedy versus Arnold Schwarzenegger in an action movie;
- and finally Will Smith in an action movie versus Arnold Schwarzenegger in a comedy.

This approach, however, presents a problem. The odds of making a Type I error (getting excited about a false positive) increase with every statistical test. Researchers typically set their alpha level at 0.05 for a t test, meaning that they are comfortable with a 5% chance of a Type I error. Unfortunately, if we conduct six t tests, *each one* has a 5% chance of a Type I error, meaning that we have a greater chance of a false-positive result somewhere in the study. In short, we need a statistical approach that reduces the number of comparisons we perform. Fortunately, a statistical technique called the **analysis of variance (ANOVA)** tests for differences by comparing the amount of variance explained by the independent variables to the variance explained by error.

The Logic of ANOVA

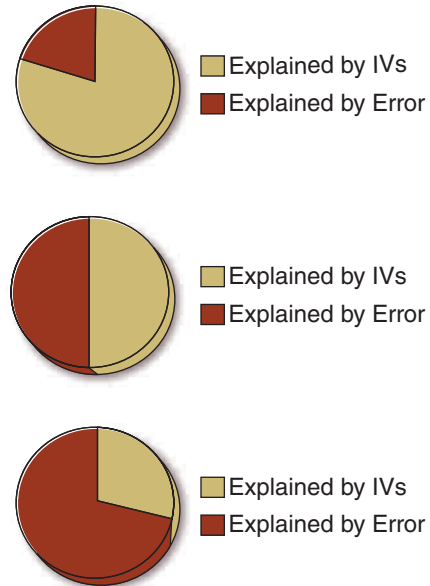
The logic behind an analysis of variance is rather straightforward. As the course has discussed throughout, variability in a dataset can be divided into systematic and error variance. That is, we can attribute some of the variability to the factors being studied, but a degree of random error will always be present. In our movie ratings study, some of the variability in these ratings can be attributed to the independent variables (differences in actors and movie types), while some of the variability is due to other factors—perhaps some people simply like movies more than other people.

The ANOVA works by comparing the influence of these different sources of variance. We always want to explain as much of the variance as possible through the independent variables. If the independent variables have more influence than random error does, this is good news. If, on the other hand, error variance has more influence than the independent variables, this is bad news for the hypotheses. Comparing the three pie charts in Figure 5.8 conveys a sense of this

problem. The proportion of variance explained by our independent variables is shaded in tan, while the proportion explained by error is shaded in red. In the top graph, the independent variables explain approximately 80% of the variance, which we can view as a good result. In the middle graph, however, variance is explained equally by the independent variables and by error, and in the bottom graph, the independent variables explain only 20% of the variance. Thus, in the latter two graphs, the independent variables do no better than random error at explaining the results.

One more analogy may be helpful. In the field of engineering, the term *signal-to-noise ratio* is used to describe the amount of light, sound, energy, etc., that is detectable above and beyond background noise. This ratio is high when the signal comes through clearly and low when it is mixed with static or other interference. Likewise, when someone tries to tune in a favorite radio station, the goal is to find a clear signal that is not covered up by static. Believe it or not, the ANOVA statistic (symbolized F) is doing the same thing. That is, the analysis tells us whether differences in experimental conditions (signal) are detectable above and beyond error variance (noise).

Figure 5.8: Comparing sources of variance



Research: Thinking Critically

Love Ballad Leaves Women More Open to a Date

Follow the link below to a press release describing a 2010 study from the journal *Psychology of Music*. The study suggests that listening to love ballads may make women more likely to give their phone number to someone they have just met. As you read the article, consider what you have learned so far about the research process, and then respond to the questions below.

<http://www.sciencedaily.com/releases/2010/06/100618112139.htm>

Think About It

1. In this experiment, the type of song (love song or neutral song) is confounded with at least one other variable. Try to identify one. Do you think that this confounded variable would make a difference? How would you design a study that overcomes this?
2. Describe how demand characteristics might compromise the internal validity of this study. Can you think of any ways around this?
3. Toward the end of the article, the authors suggest that one explanation for these results is that the romantic music put the women into a more positive mood, and that this in turn made them more receptive to the men. How could you design a study that tests this hypothesis?
4. Given the nature of the DV in this study, would an ANOVA test be appropriate? What would be the more appropriate statistical test, and why?

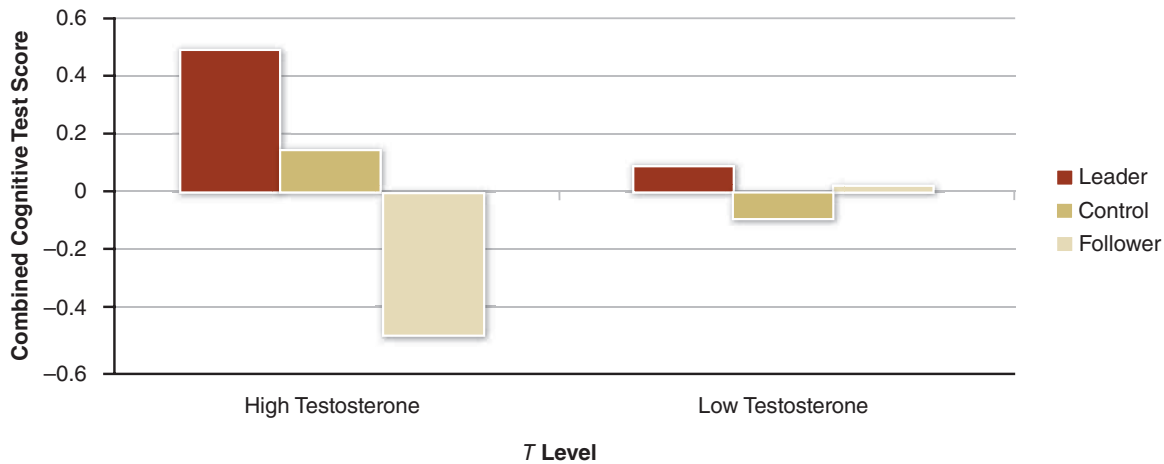
Exploring the Data

Statistics courses cover ANOVA in more detail, but, despite its elegant simplicity, the test has a notable limitation. After conducting an ANOVA, we have a yes-or-no answer to the following question: *Do our experimental groups have a systematic effect on the dependent variable?* The answer lets us decide whether to reject the null hypothesis, but it does not tell us everything we want to know about the data. In essence, a significant F value tells us that the groups have a significant difference, but it does not tell us *what the difference is*. Conducting an ANOVA on our movie-ratings study would reveal a significant interaction between actor and movie, but we would need to take additional steps to determine the meaning of this interaction.

This section will describe the process of exploring and interpreting ANOVA results to make sense of the data. The example is drawn from a published study by Newman, Sellers, and Josephs (2005), which was designed to explore the effects of testosterone on cognitive performance. Previous research had suggested that testosterone was involved in two types of complex human behavior. On one hand, people with higher testosterone tend to perform better on tests of spatial skills, such as having to rotate objects mentally, and perform worse on tests of verbal skills, such as listing all the synonyms for a particular word. These patterns are thought to reflect the influence of testosterone on developing brain structures. On the other hand, people with higher testosterone are also more concerned with gaining and maintaining high status relative to other people. Testosterone correlates with a person's position in the hierarchy and tends to rise and fall when people win and lose competitions, respectively. Sociologist Alan Mazur and his colleagues measured testosterone levels before, during, and after a series of professional chess matches. They found that testosterone rose in both players in anticipation of the competition, then rose even further in the winners, but plummeted in the losers (Mazur, Booth, & Dabbs, 1992).

Newman and colleagues (2005) set out to test the combination of these variables. Based on previous research, they hypothesized that people with higher testosterone would be uncomfortable when they were placed in a low-status position, leading them to perform worse on cognitive tasks. The researchers tested this hypothesis by randomly assigning people to a high status, low status, or control condition, and then administering a spatial and a verbal test. The resulting between-subjects design was a 2 (testosterone: high or low) $\times$ 3 (condition: high status, low status, control), for a total of six groups. Note that "testosterone" in this study is a quasi-independent variable, because it is measured rather than manipulated by the experimenters.

Once the results were in, the ANOVA revealed an interaction between testosterone and status but no main effects. Figure 5.9 shows the results of the study. These bars represent z scores that combine the spatial and verbal tests into one number. So, what do these numbers mean? How do we make sense out of the patterns? Doing so involves a combination of comparing means and calculating effect sizes, as we discuss next.

Figure 5.9: Exploring the data: Results from Newman et al. (2005)

Newman et al. (2005)

Mean Comparisons

The first step in interpreting results is to compare the various pairs of means within the design. This might seem counterintuitive, since the whole point of the ANOVA was to test for effects without comparing individual means. Our goal, therefore, is to somehow explore differences in conditions without inflating Type I error rates. Achieving this balance involves two strategies.

Planned comparisons (also called **a priori comparisons**) involve comparing only the means for which differences were predicted by the hypothesis. In the experiment by Newman et al. (2005), the hypothesis explicitly stated that high-testosterone people should perform better in a high-status position than a low-status position. So, a planned comparison for this prediction would involve comparing two means with a t test: high T , high status (the highest red bar); and high T , low status (the lowest tan bar). Consistent with the researchers' hypothesis, high-testosterone people did perform higher on both tests, $t(27) = 2.35$, $p = 0.01$, but only in a high-status position. Type I errors are of less concern with planned comparisons because only a small number of theoretically driven comparisons are being conducted.

Referring to the graph of these results in Figure 5.9 and comparing high- with low-testosterone people reveals another interesting pattern: In a high-status position, high-testosterone people do better than low-testosterone people, but in a low-status position, this pattern is reversed, and high-testosterone people do worse. However, the researchers did not predict these mean comparisons, so to do planned contrasts would be cheating. Instead, they would use a second strategy called a **post hoc comparison**, which controls the overall alpha by taking into account the fact that multiple comparisons are being performed. In most cases, research only permits post hoc tests if the overall F test is significant.

One popular way to conduct post hoc tests while minimizing the error rate is to use a technique called a **Bonferroni correction**. This technique, named after the Italian mathematician who developed it, involves simply adjusting the alpha level by the number of comparisons that are performed. For example, imagine we want to conduct 10 follow-up post hoc tests to explore the data. The Bonferroni correction would involve dividing the alpha level (0.05) by the number of comparisons (10), for a corrected alpha level of 0.005. Then, rather than using a cutoff of 0.05 for each test, we use this more conservative Bonferroni-corrected value of 0.005. Translation: Rather than accepting a Type I error rate of 5%, we are moving to a more conservative 0.5% cutoff to correct for the number of comparisons that we are performing.

Another popular alternative to the Bonferroni correction is called **Tukey's HSD** (for *Honestly Significant Difference*). This test works by calculating a critical value for mean comparisons (the HSD), and then using this critical value to evaluate whether mean comparisons are significantly different. The test manages to avoid inflating Type I error because the HSD is calculated based on the sample size, the number of experimental conditions, and the *MSWG*, which essentially tests all the comparisons at once. In the study by Newman et al. (2005), both of these post hoc tests were significant: Compared to those low in testosterone, high-testosterone people did better in a high-status position but worse in a low-status position, suggesting that high testosterone magnifies the effect of testing situations on cognitive performance.

Effect Size

Statistical significance is only part of the story; researchers also want to know how big the effects of their independent variables are. Researchers can calculate effect size using several ways, but in general, bigger values mean a stronger effect. One of these statistics, Cohen's d , is calculated as the difference between two means divided by their pooled standard deviation. The resulting values can therefore be expressed in terms of standard deviations; a d of 1 means that the means are one standard deviation apart. How big should we expect our effects to be? Based on Cohen's analyses of typical effect sizes in the social sciences, he suggests the following benchmarks: $d = 0.20$ is a small effect; $d = 0.40$ is a moderate effect; and $d = 0.60$ is a large effect. In addition to these qualitative categories, effect-size values can be interpreted in terms of standard deviation units. So, a d of 1 is equivalent to a standard deviation of 1. In other words, a large effect in social and behavioral sciences accounts for a little more than half of a standard deviation.

In interpreting the results of their testosterone experiment, Newman and colleagues (2005) computed effect-size measurements for two of the key mean comparisons. First, they compared high-testosterone people in the high- and low-status conditions; the size of this effect was a $d = 0.78$. Second, they compared the high- and low-testosterone people in the low-status condition; the size of this effect was a $d = 0.61$. Both of these effects fall in the "large" range based on Cohen's benchmarks. More important, taken together with the mean comparisons, they help us to understand the way testosterone affects behavior. The authors conclude that cognitive performance stems from an interaction between biology (testosterone) and environment (assigned status) such that high-testosterone people are more responsive to their status in a given situation. When they are placed in a high-status position, they relax and perform well. Conversely, when placed in a low-status position, they become distracted and perform poorly. Researchers reach this nuanced conclusion only through an exploration of the data, using mean comparisons and effect-size measures.

5.6 Wrap-Up: Avoiding Error

As this final chapter concludes, it is worth thinking back to one of the key concepts in Chapter 2 (2.4): Type I and Type II errors. Regardless of the research question, the hypothesis, or the particulars of the research design, all studies have the goal of making accurate decisions about the hypotheses. That is, we need to be able to correctly reject the null hypothesis when it is false, and fail to reject the null when it is true. Still, from time to time and despite our best efforts, we make mistakes when we draw conclusions about our hypotheses, as Table 5.4 summarizes. A Type I error, or “false positive,” involves falsely rejecting a null hypothesis and becoming excited about an effect that is due to chance. A Type II error, or “false negative,” involves failing to reject the null hypothesis and missing an effect that is real and interesting. (For a refresher on these terms, refer back to Chapter 2.)

Table 5.4: Review of Type I and Type II errors

	Researcher's Decision	
	Reject Null	Fail to Reject Null
Null is FALSE	Correct Decision	Type II Error
Null is TRUE	Type I Error	Correct Decision

This section takes a problem-solving approach to minimizing both of these errors in an experimental context. It turns out that each error is primarily under the researcher's control at different stages in the research process, which means reducing each error calls for different strategies.

Avoiding Type I Error

Type I errors occur when results are due to chance but are mistakenly interpreted as significant. We can generally reduce the odds of this happening by setting our alpha level at $p < 0.05$, meaning that we will only be excited about results that have less than a 5% chance of Type I error. However, Type I errors can still occur as a result of either extremely large samples or large numbers of statistical comparisons. Large samples can make small effects seem highly significant, so it is important to set a more conservative alpha level in large-scale studies. And, this chapter has discussed, the odds of Type I error are compounded with each statistical test we conduct.

What this means is that Type I error is primarily under researchers' control during statistical analysis—the smarter the statistics, the lower the odds of Type I error. This chapter has discussed several examples of “smart” statistics: Instead of conducting lots of t tests, we use an ANOVA to test for differences across the entire design simultaneously. Instead of conducting t tests to compare means after an ANOVA, we use a mix of planned contrasts (for comparisons that we predicted) and post hoc tests (for other comparisons we want to explore). More advanced statistical techniques take this a step further. For example, the **multivariate analysis of variance (MANOVA)** statistic analyzes sets of dependent variables to reduce further

the number of individual tests. Researchers use this approach when dependent variables represent different measures of a related concept, such as using heart rate, blood pressure, and muscle tension to capture the stress response. The MANOVA works, broadly speaking, by computing a weighted sum of these separate DVs (called a *canonical variable*) and using this new variable as the dependent variable. To learn more about this and other advanced statistical techniques, see the excellent volume by James Stevens (2002), *Applied Multivariate Statistics*, <http://goo.gl/zgGDj>.

Avoiding Type II Error

Type II errors occur when a real underlying relationship exists between the variables, but the statistical tests are nonsignificant. The primary sources of this error are small samples and bad design. Small samples may fail to capture enough variability and may therefore lead to nonsignificant p values in testing an otherwise significant effect. Both large and small mistakes in experimental designs can add noise to the dataset, making it difficult to detect the real effects of independent variables.

This means that Type II error is primarily under the researcher's control during the design process—the smarter the research designs, the lower the odds of Type II error. First, as Chapter 2 discussed, it is relatively simple to estimate the sample size needed for our research using a power calculator. These tools take basic information about the number of conditions in the research design and the estimated size of the effect and then estimate the number of people needed to detect this effect. (See Chapter 2, Figure 2.5, for an annotated example using one of these online calculators.)

Second, as every chapter has discussed, it is the experimenter's responsibility to take steps to minimize extraneous variables that might interfere with the hypothesis test. Whether researchers are conducting an observation, a survey study, or an experiment, the overall goal is to ensure that the variables of interest are the main cause of changes in the dependent variable. This is perhaps easiest in an experimental context because these designs are usually conducted in a controlled setting where the experimenter has control over the independent variables. Nonetheless, as the chapter discussed earlier, many factors can threaten the internal validity of an experiment—from confounds to sample bias to expectancy effects. In essence, the more we can control the influence of these extraneous variables, the more confidence we can have in the results of the hypothesis test.

Table 5.5 presents a summary of the information in this section, listing the primary sources of Type I and Type II errors, as well as the time period when these are under experimenter control.

Table 5.5: Summary—avoiding error

Error	Definition	Main Source	When You Can Control
Type I	False-positive	Lots of tests; lots of people	Conducting stats
Type II	False-negative	Bad measures; not enough people	Designing experiments

Summary and Resources

Chapter Summary

This chapter focused on experimental designs, in which the primary goal is to explain behavior in causal terms. The chapter began with an overview of experimental terminology and the key features of experiments. Three key features distinguish experiments from other research designs. First, researchers manipulate a variable, giving them a fair amount of confidence that the independent variable (IV) causes changes in the dependent variable (DV). Second, researchers control the environment, ensuring that everything about the experimental context is the same for different groups of participants—*except* for the level of the independent variable. Finally, the researchers have the power to assign participants to conditions using random assignment. This process helps to ensure that preexisting differences among participants (e.g., in mood, motivation, intelligence, etc.) are balanced across the experimental conditions.

Next, the chapter explained the concept of experimental validity. When evaluating experiments, researchers must take into account both internal validity—or the extent to which the IV is the cause of changes in the DV—and external validity—or the extent to which the results generalize beyond the specific laboratory setting. Several factors can threaten internal validity, including experimental confounds, selection bias, and expectancy effects. The common thread among these threats is that they add noise to the hypothesis test and cast doubt on the direct connection between IV and DV. External validity involves two components, the realism of the study and the generalizability of the findings. Psychology experiments are designed to study real-world phenomena, but sometimes compromises have to be made to study these phenomena in the laboratory. Research often achieves this balance via mundane realism, or replicating the psychological conditions of the real phenomenon. Last, researchers have more confidence in the findings of a study when they can be replicated, or repeated in different settings with different measures.

In designing the nuts and bolts of experiments, researchers have to make decisions about both the nature and number of independent variables. First, designs can be described as between-subject, within-subject, or mixed. In a between-subject design, participants are in only one experimental condition and receive only one combination of the independent variables. In a within-subject design, participants are in all experimental conditions and receive all combinations of the independent variables. Finally, a mixed design contains a combination of between- and within-subject variables. In addition, research designs can be described as either one-way or factorial. One-way designs consist of only one IV with at least two levels; factorial designs consist of at least two IVs, each having at least two levels. A factorial design produces several results to examine: the main effect of each IV plus the interaction, or combination, of the IVs.

The chapter also discussed the logic of analyzing experimental data, using the analysis of variance (ANOVA) statistic. This test works by simultaneously comparing sources of variance and therefore avoids the risk of inflated Type I error. The ANOVA (or F) is calculated as a ratio of systematic variance to error variance, or, more specifically, of between-groups variance to within-groups variance. The bigger this ratio, the more experimental manipulations contribute to overall variability in scores. However, the F statistic suggests only that differences exist in the design; further analyses are necessary to explore these differences. The chapter

described an example from a published study, discussing the process of comparing means and calculating effect sizes. In comparing means, researchers use a mix of planned contrasts (for comparisons that they predicted) and post hoc tests (for other comparisons they want to explore).

Finally, the chapter concluded by referring to two recurring concepts, Type I error (false positive) and Type II error (false negative). These errors interfere with the broad goal of making correct decisions about the status of a hypothesis. Thus, the purpose of this final section was to review ways to minimize errors. Type I errors are primarily inflated by large samples and lots of statistical analyses. Consequently, this error is under the experimenter's control at the data-analysis stage. Type II errors are primarily inflated by small samples and flaws in the experimental design. Consequently, this error is under the experimenter's control at the design and planning stage.

Key Terms

analysis of variance (ANOVA) A statistical procedure that tests for differences by comparing the variance explained by systematic factors to the variance explained by error.

between-subject design Experimental design in which each group of participants is exposed to only one level of the independent variable.

Bonferroni correction A post hoc test that involves adjusting the alpha level by the number of comparisons to set a more conservative cutoff.

carryover effect Effects of one level are present when another level is introduced, making it difficult to separate the effects of different levels.

conceptual replication Testing the relationship between conceptual variables using new operational definitions.

condition One of the versions of an independent variable, forming different groups in the experiment; in a factorial design, refers to the groups formed by combinations of IVs.

confounding variable (or confound) A variable that changes systematically with the independent variable.

constructive replication Recreation of the original experiment that adds elements to the design; usually designed to rule out alternative explanations or extend knowledge about the variables under study.

control condition Group within the experiment that does not receive the experimental treatment.

counterbalancing Variation of the order of presentation among participants to reduce order effects.

cover story A misleading statement to participants about what is being studied to prevent effects of demand characteristics.

demand characteristic Cue in the study that leads participants to guess the hypothesis.

differential attrition Loss of participants, who drop out of experimental groups for different reasons.

environmental manipulation Changing some aspect of the experimental setting.

exact replication Recreation of the original experiment as closely as possible to verify the findings.

experimental condition Group within the experiment that receives a treatment designed to test a hypothesis.

experimental design Design whose primary goal is to explain causes of behavior.

experimenter expectancy Researchers see what they expect to see, leading to subtle bias in favor of their hypotheses; threat to internal validity.

external validity A metric that assesses generalizability of results beyond the specific conditions of the experiment.

extraneous variable Variable that adds noise to a hypothesis test.

factorial design A design that has two or more independent variables, each with two or more levels.

factorial notation A system for describing the number of variables and the number of levels in experimental designs.

fatigue effect Decline of participants' performance as a result of repeated testing.

generalizability The extent to which results extend to other studies, using a wide variety of populations and of operational definitions.

instructional manipulation Changing the way a task is described to change participants' mind-sets.

interaction The combined effect of variables in a factorial design; the effects of one IV are different depending on the levels of the other IV.

internal validity A metric that assesses the degree to which results can be attributed to independent variables.

invasive manipulation Taking measures to change internal, physiological processes; usually conducted in medical settings.

level Another way to describe the versions of an independent variable; describes the specific circumstances created by manipulating a variable.

main effect The effect of each independent variable on the dependent variable, collapsing across the levels of other variables.

marginal mean The combined mean of one factor across levels of another factor.

matched random assignment A variation on random assignments; ensures that an important variable is equally distributed between or among the groups; the experimenter obtains scores on an important matching variable, ranks participants on this variable, and then randomly assigns participants to conditions.

mixed design Experimental design that contains at least one between-subject variable and at least one within-subject variable.

multivariate analysis of variance (MANOVA) A statistic that analyzes sets of dependent variables to reduce the number of individual tests.

mundane realism Research that replicates the psychological conditions of the real-world phenomenon; criterion for judging external validity.

one-way design A design that has only one independent variable, with two or more levels to the variable.

order effect Moderation of the effects because of the order in which levels occur.

participant replication Repetition of the study with a new population of participants; usually driven by a compelling theory as to why the two populations differ.

placebo control Group added to a study to reduce placebo effects; mimics the experimental condition in every way but one.

placebo effect Change resulting from the mere expectation that change will occur.

planned comparison (or a priori comparison) Comparisons that involve comparing only the means for which differences were predicted by the hypothesis.

post hoc comparison Comparison that controls the overall alpha by taking into account that multiple comparisons are being performed; usually allowed only if the overall *F* test is significant.

practice effect Improvement of participants' performance as a result of repeated testing.

quasi-independent variable Preexisting difference used to divide participants in an experimental context; referred to as "quasi" because variables are being measured, not manipulated, by the experimenter.

random assignment A technique for assigning participants to conditions; before participants arrive, the experimenter makes a random decision for each participant's placement in a group.

replication Repetition of research results in different contexts and/or different laboratories.

selection bias Occurs when groups are different before the manipulation; problematic because preexisting differences might be the driving factor behind the results.

Tukey's HSD (*Honestly Significant Difference*) A post hoc test that calculates a critical value for mean comparisons (the HSD) and then uses this critical value to evaluate whether mean comparisons are significantly different.

unrelated-experiments technique A strategy for preventing the effects of demand characteristics, leading participants to believe that they are completing two experiments during one session; experimenter can use this to present the independent variable during the first experiment and measure the dependent variable during the second experiment.

within-subject design Experimental design in which each group of participants is exposed to all levels of the independent variable.

Apply Your Knowledge

1. List and briefly describe the three distinguishing features of an experiment.
 - a.
 - b.
 - c.
2. List the three types of expectancy effect that can affect experimental results, and name one way to avoid each type.
 - a.
 - b.
 - c.
3. The following designs are described using factorial notation. For each one, state (a) the number of variables in the design, (b) the number of levels each variable has, and (c) the total number of experimental conditions.

$$3 \times 3 \times 3$$

- a.
- b.
- c.

$$2 \times 3 \times 4$$

- a.
- b.
- c.

$$4 \times 4$$

- a.
- b.
- c.

$$2 \times 2 \times 2 \times 2$$

- a.
- b.
- c.

4. Forty students were asked to rate two authors according to their knowledge of certain topic areas. Each student was given two passages to read. In one passage ("Brain"), the author discussed the roles of various brain structures in perceptual-motor coordination. In the second passage ("Motivation"), the author described ways to enhance motivation in preschool children. For half the students, both passages were written by male authors. For the other half of the students, both passages were written by a female author. After reading the passages, students rated the authors' knowledge of their topic areas on a scale ranging from 1 (displays very little knowledge) to 10 (displays a thorough knowledge).

5.

	Male Author	Female Author
Brain	9	4
Motivation	6	7

- (1) Identify the following information about the design:
 - (2) Describe the design using factorial notation (e.g., 4×3).
 - (3) Identify the total number of conditions.
 - (4) Identify the design (circle one): between-subject within-subject mixed
6. For each of the following scenarios, identify what a Type I error and a Type II error would look like. Then, determine which type would be a bigger problem for that scenario.
- a. A large international airport has received a bomb threat. In response, the airport police have tightened security and now check every piece of luggage manually.
 - (1) Type I:
 - (2) Type II:
 - (3) Bigger problem:
 - b. Your friend purchases a pregnancy test.
 - (1) Type I:
 - (2) Type II:
 - (3) Bigger problem:

Critical Thinking Questions

1. Explain the advantages and disadvantages of a within-subject design.
2. Compare and contrast the following terms. Your answers should demonstrate that you understand each term. Be sure to give some kind of context (e.g., "both are types of . . .") or provide an example, and state how they are different.
 - a. internal versus external validity
 - b. between-subjects versus within-subject design
 - c. level versus condition
3. Explain the difference between Type I and Type II errors. How can each type of error be minimized?

Research Scenarios: Try It

Imagine that you are working with a pharmaceutical company to determine whether exercise enhances the effect of a new antidepressant medication. Prior research by this company has suggested that the drug appears to work better among people who exercise regularly than those who are sedentary. To determine this, the researchers recruited depressed individuals who either had a regular exercise routine or were sedentary and randomly assigned them to an active drug condition or a placebo control condition. After six months of treatment, they found that the medication reduced depression far more dramatically among individuals who exercised than among the sedentary individuals.

1. The pharmaceutical researchers concluded on this basis that exercise enhances the effect of their drug, but you spot a flaw in their design that compromises its internal validity. What is it?
 - a. Confounding variables
 - b. Differential attrition
 - c. Expectancy effects
 - d. Placebo effect
2. Suppose you design a constructive replication of the original study which corrects the problem you found by recruiting only individuals without regular exercise programs and randomly assigning them to the exercise and no-exercise groups. Therefore, participants are randomly assigned to one of four conditions: (1) exercise + drug, (2) no exercise + drug, (3) exercise + placebo, (4) no exercise + placebo. You measure depression before and after treatment. What are the independent variable(s) (IV) and dependent variable(s) (DV) of your experiment?
 - a. 1 DV: depression; 3 IV: exercise, drug, time
 - b. 1 DV: depression; 2 IV: exercise, drug
 - c. 2 DV: depression, time; 2 IV: exercise, drug
 - d. 3 DV: exercise, depression, drug; 1 IV: time
3. To ensure that patients in the exercise conditions actually work out regularly, you require them to keep an exercise journal and come into the lab three times a week to use a treadmill for 45 minutes under medical supervision. What should you instruct the no-exercise condition do?
 - a. Not to begin exercise.
 - b. Record all physical activity.
 - c. Come into the lab three times a week for 45 minutes of quiet sitting under medical supervision.
 - d. All of the above.
4. Suppose you complete your study and begin to analyze your data. To simplify the analysis, you analyze the change in depression levels from before to after treatment. You obtain the following results for your change in depression scores (more negative numbers mean greater drops in depression levels):

5.

	Drug	Placebo
Exercise	-20	-10
No exercise	-10	-10

Assuming that your statistical tests are significant, what are the main effects of exercise and drug condition in this study?

- a. Main effect of drug, no main effect of exercise
 - b. Main effect of exercise, no main effect of drug
 - c. No main effect of either drug or exercise
 - d. Main effects of both drug and exercise
6. Looking at the results more closely, you realize that the two main effects are not the whole story. Select the option that best states the nature of the interaction:
- a. The effect of the drug depends on whether the participants exercised or not.
 - b. The effect of exercise depends on whether the participants had the active drug or placebo.
 - c. Both (a) and (b) are true.
 - d. The effect of the drug is independent of the effect of exercise.

See Appendix A for answers to Research Scenarios: Try It questions.