WHEN LOUISE HAY DIED IN 2017, her net worth was estimated at around $50 million. Her most popular book, *You Can Heal Your Life*, had sold over 35 million copies. In it, she explained that everything that happens to people—including diseases, accidents, and other misfortunes—is a result of the thoughts they choose to think. She claimed that she had been diagnosed with incurable cancer but had cured herself by changing the way she thought, and she amassed a fortune by promising people that they could learn to do the same thing if only they attended one of her seminars or bought a few of her videos or books. In a 2010 television interview with one of the authors of this textbook, Hay was asked why she believed her techniques were actually effective.

**Gilbert:** How do you know what you’re saying is right?

**Hay:** Oh, my inner ding.

**Gilbert:** Ding?

**Hay:** My inner ding. It speaks to me. It feels right or it doesn’t feel right. Happiness is choosing thoughts that make you feel good. It’s really very simple.

**Gilbert:** But I hear you saying that even if there were no proof for what you believed, or even if there were scientific evidence against it, it wouldn’t change?

**Hay:** Well, I don’t believe in scientific evidence, I really don’t. Science is fairly new. It hasn’t been around that long. We think it’s such a big deal, but it’s, you know, it’s just a way of looking at life.

Louise Hay said that she didn’t “believe in scientific evidence”—but what could that possibly mean? After all, if her techniques really did cure cancer, then cancer victims who practice her techniques should on average live longer than those who don’t. That isn’t “a way of looking at life.” It’s just plain old common sense—exactly the kind of common sense that lies at the heart of science. Science tells us that there is one and only one way to know for sure whether claims like Louise Hay’s are true, and that’s to

Louise Hay claimed that people can cure cancer with their minds. How can we tell whether her claim is right or wrong?
gather evidence. Sorry, but inner dings don’t count. If we really want to know what’s true about the world, then we actually have to go there and take a look around.

But what exactly should we be looking for? Should we show up at a “Hay House” seminar and ask people in the audience whether they think they’ve been healed by her techniques? Should we examine the medical records of people who have and haven’t bought one of her books? Should we invite people to sign up for a class that teaches her techniques and then wait to see how many of them get cancer over the next few years? All of these may strike you as fairly reasonable ways to test Louise Hay’s claim, but in fact, every one of them is utterly useless. It turns out that there are a few very good ways to test claims about the world and a whole lot of bad ways, and the main point of this chapter is to teach you the difference between them. Scientists have developed powerful tools for determining when a claim is right and when it is wrong, and these tools are what makes science so different from all other ways of knowing. As the philosopher Bertrand Russell (1945, p. 527) wrote, “It is not what the man of science believes that distinguishes him, but how and why he believes it.” That goes for women of science too.

WE’LL START BY EXAMINING THE GENERAL PRINCIPLES THAT GUIDE scientific research and distinguish it from other human enterprises. Next, we’ll see how the methods of psychology allow us to answer two basic questions: What do people do, and why? Psychologists answer the first question by measuring stuff, and they answer the second question by looking for relationships between the stuff they measure. We’ll see that scientific evidence allows us to draw certain kinds of conclusions but not others, and that thinking critically about scientific evidence doesn’t come naturally to most people. Finally, we’ll consider some of the unique ethical questions that confront scientists who study human beings and other animals.

Empiricism: How to Know Stuff

Learning Outcomes

- Explain why direct observation is essential to an accurate understanding of nature.
- Outline the process of the scientific method.
- Identify the challenges to studying human behavior.

**Empiricism** The belief that accurate knowledge can be acquired through observation.

When ancient Greeks sprained their ankles, caught the flu, or accidentally set their beards on fire, they had to choose between two kinds of doctors. The dogmatists (from *dogmatikos*, meaning “belief”) thought the best way to understand illness was to develop theories of the body’s functions, and the empiricists (from *empeirikos*, meaning “experience”) thought the best way was to watch sick people and see what happened. The rivalry between these two schools of medicine didn’t last long because the people who went to see dogmatists tended to die a lot, which was bad for business. That’s why today we use the word *dogmatism* to describe people’s tendency to cling to their beliefs and assumptions and we use the word *empiricism* to describe the belief that accurate knowledge can be acquired through observation. The fact that we can answer questions about the natural world by observing probably seems obvious to you, but this obvious fact has only recently gained wide acceptance. For most of human history, people trusted authority to provide answers to life’s important questions, and it is only in the last millennium (and especially in the past three centuries) that people have begun to trust their eyes and ears more than their elders.
Empiricism: How to Know Stuff

35

The Scientific Method

Empiricism is the backbone of the scientific method, which is a procedure for using empirical evidence to establish facts. Essentially, the scientific method suggests that when we have an idea about how something in the world works—about how bats navigate, or where the moon came from, or why people can’t forget traumatic events—we must go out into the world, make observations, and then use those observations to determine whether our idea is true. Scientists generally refer to these “ideas about how something works” as theories, which are hypothetical explanations of natural phenomena. So, for instance, we might theorize that bats navigate by making sounds and then listening for the echo, or that the moon was formed when a small planet collided with the earth, or that the brain responds to traumatic events by producing chemicals that facilitate memory. Each of these theories is an explanation of how something in the natural world works and why it works that way.

How do we decide if a theory is right? A good theory makes specific predictions about what we should observe in the world if the theory is true. For example, if bats really do navigate by making sounds and then listening for echoes, then deaf bats should not be able to navigate very well. That “should” statement is technically known as a hypothesis, which is a falsifiable prediction made by a theory. The word falsifiable is a critical part of that definition. Some theories, such as “Things happen the way they do because that’s how God wants it,” do not tell us what we should observe if the theory is true; therefore, no amount of observation can ever falsify them. That doesn’t mean the theory is wrong. It just means that we can’t use the scientific method to evaluate its veracity.

So good theories give rise to hypotheses that can be falsified, and when that happens the theory is proved wrong. But how can we prove it right? Alas, although a theory can be proved wrong, it can never be proved right. For example, imagine that you decided to test the theory that bats navigate by using sound. The theory gives rise to a hypothesis: Deaf bats should not be able to navigate. Now, if you observed a deaf bat navigating perfectly well, your observation would be clearly inconsistent with the predictions of your theory, which must therefore be wrong. On the other hand, if you observed deaf bats navigating badly, then your observation would be perfectly consistent with the predictions of your theory—but it would not prove your theory right. Why? Because even if you didn’t see a deaf bat navigating perfectly well today, it is always possible that you will see one tomorrow, or the day after that, or maybe 30 years from now. You did not observe every bat that has ever been and will ever be, so even if today’s observation didn’t prove your theory wrong, there is always a chance that some future observation will. The point here is that observations that are consistent with a theory can increase our confidence that the theory is right, but they can never make us absolutely sure it is right. So the next time you see a headline that says “Scientists prove theory X,” you are hereby authorized to roll your eyes. And make a noise like a bat.

The scientific method tells us that the only way to learn the truth about the world is to develop a theory, derive a falsifiable hypothesis from it, and then test that hypothesis by observing the world—or, in fancier language, gathering empirical evidence. You now know all about theories and hypotheses, so let’s see what this “evidence gathering” stuff actually entails.

The Art of Looking

For centuries, people argued about whether all four of a horse’s feet ever leave the ground at the same time. They would all go out and watch the same horse

The astronomer Galileo Galilei (1564–1642) was excommunicated and sentenced to prison for sticking to his empirical observations of the solar system rather than accepting the teachings of the Church. In 1597 he wrote to his friend and fellow astronomer Johannes Kepler (1571–1630), “What would you say of the learned here, who, replete with the pertinacity of the asp, have steadfastly refused to cast a glance through the telescope? What shall we make of this? Shall we laugh, or shall we cry?” As he later learned, the answer was cry.

run, and then some would swear that all four feet left the ground at the same time, some would swear that they didn’t, and some would just swear. Then, in 1877, a man named Eadweard Muybridge invented a technique for taking photographs in rapid succession. His photos showed that when horses gallop . . . why yes, they really do go airborne! And that was that. Never again did a group of friends get to engage in a flying-horse debate because Muybridge’s pictures had settled the matter forever.

But why did it take so long? Human beings had been watching horses gallop for centuries, so why were some of them sure that horses left the ground completely while others were equally sure they didn’t? Because as wonderful as human eyes may be, there are many things they see inaccurately, and many things they don’t see at all. From where you are sitting right now, the earth looks perfectly flat when it is in fact imperfectly round. From where you are sitting right now, you can’t see a microbe, or ultraviolet light, or the 14 moons of Neptune, but all of them are real and all of them are there. As Muybridge knew, empiricism is the right approach, but to do it properly requires more than a set of eyes. It requires an empirical method, which is a set of rules and techniques for observation.

In many sciences, the word method refers to technologies that enhance the powers of the senses. Biologists use microscopes and astronomers use telescopes because the things they want to observe are invisible to the naked eye. Human behavior, on the other hand, is easy to see, so you might expect psychology’s methods to be relatively simple. But they aren’t simple, and that’s because human beings have three qualities that make them more difficult to study than either cells or stars. First, people are highly complex. Although scientists can describe the birth of a star or the death of a cell in exquisite detail, they can barely begin to say how the 100 billion interconnected neurons that constitute the human brain give rise to the thoughts, feelings, and actions that are psychology’s core concerns. Second, people are highly variable. One ribosome may be a lot like another, but no two people ever do, say, think, or feel exactly the same thing under exactly the same circumstances. Third, people are highly reactive. A cesium atom oscillates at the same rate regardless of who might be watching, but people tend to think, feel, and act differently when they are or are not being observed.

The fact that human beings are highly complex, variable, and reactive presents a challenge to the scientific study of their behavior. Psychologists have met this challenge by developing two kinds of methods: methods of observation, which allow them to determine what people do, and methods of explanation, which allow them to determine why people do it. We’ll examine each of these methods in the sections that follow.

**empirical method** A set of rules and techniques for observation.

---

**Build to the Outcomes**

1. What is the scientific method?
2. What is the difference between a theory and a hypothesis?
3. Why can theories be proven wrong but not right?
4. What makes human behavior especially difficult to study scientifically?
A World of Difference

Are Heroes and Sheroes Divided by Zeroes?

Galileo, Newton, Mendel, Darwin, Faraday, Einstein, Turing, Tesla, Skinner—what do all these people have in common? First, they were all brilliant scientists. And second, they were all men. The history of science is pretty much the history of smart men having big ideas and making big discoveries. So where are all the women? There are two answers to this question. The first is that, until quite recently, educational and employment opportunities for women were limited. For most of history, women were either subtly discouraged or actively prohibited from studying science—and you really can’t win the Nobel Prize in Physics if they won’t let you take algebra. In 1914, the sociologist Leta Hollingworth cheekily noted that “women bear and rear the children, and this has had as an inevitable sequel the occupation of housekeeping, a field where eminence is not possible.” Women, it seems, have been right where society put them: in the kitchen.

The second answer is that men and women may have different interests and talents, and the interests and talents that men have may be the ones people need to become great scientists. Is there any truth to this suggestion? A few years ago, a group of male and female scientists got together, surveyed all the scientific evidence on this topic, and drew a striking conclusion: Yes, they concluded, the evidence does show that men and women differ with regard to certain interests and talents (Halpern et al., 2007). First, men are often more interested in the topics that scientists study. Second, men and women differ in scientifically relevant abilities: Specifically, men are more variable than women on most measures of quantitative ability (the ability to use numbers) and visuo-spatial ability (the ability to use images), both of which are important to success in many sciences.

Being more variable doesn’t mean being better, however. It means that although men and women have the same average amount of talent on these dimensions, there are more men at both the very lowest and very highest ends of the spectrum. If great scientists tend to come from the highest end of the spectrum, then men will naturally be over-represented among great scientists. In fact, recent data show that while men are overrepresented in scientific fields that are “math intensive”—such as geoscience, engineering, economics, mathematics, computer science, and the physical sciences—they are not overrepresented in scientific fields that place less emphasis on math, such as biology, psychology, and sociology (Ceci et al., 2014). Indeed, while women earn fewer than 20% percent of the degrees in computer science and engineering, they earn the majority of degrees in the life and social sciences. With regard to having different interests, it is worth noting that the largest gender gaps in the “hard sciences” emerge in the nations with the highest gender equality, suggesting that when women have a choice, they tend not to choose those fields (Stoet & Geary, 2018).

So yes, men are more interested in certain topics and are more variable on certain dimensions. Why? Is it because there is some innate difference in the structure or function of the male and female brains, or is it because men are encouraged to develop their interests and hone their quantitative skills by parents who buy them video games instead of dolls, and by teachers who encourage them to join the math team rather than the debate team? Once again, an expert review of the evidence concludes that “sex differences in spatial and mathematical reasoning need not stem from biological bases, that the gap between average female and male math ability is narrowing (suggesting strong environmental influences), and that sex differences in math ability at the right tail [which is the “high” end of the ability spectrum] show variation over time and across nationalities, ethnicities, and other factors, indicating that the ratio of males to females at the right tail can and does change” (Ceci et al., 2014, p. 75). In other words, men may be more interested in certain scientific topics and may be more likely to be unusually talented (or untalented!) when it comes to math, but there is no compelling evidence to suggest that these differences are innate.

We agree with the experts who concluded that “there are no single or simple answers to the complex questions about sex differences in science and mathematics” (Halpern et al., 2007, p. 75). But we feel confident that someday a brilliant young psychological scientist will discover the true answers to these complex questions. We hope to be here when she does.

In 1834, William Whewell coined the word “scientist” to describe a remarkable astronomer, physicist, and chemist named Mary Somerville. Few people remember that the world’s first scientist was a woman.
Methods of Observation: Finding Out What People Do

When you observe an apple, your brain uses the pattern of light that enters your eyes to make inferences about the apple’s color, shape, and size. That kind of observation is good enough for buying fruit, but not for doing science. Why? First, everyday observations are often inconsistent: The same apple can appear red in the daylight and crimson at night, or spherical to one person and elliptical to another. Second, everyday observations are often incomplete, which is to say that they simply can’t provide a lot of the information we might want. No matter how long and hard you stared at an apple, you would never be able to determine its melting point or pectin content. Luckily, scientists have devised techniques that allow them to overcome these limitations. In the sections that follow, we’ll first see how psychologists measure the things that everyday observation misses (Measurement), and then we’ll see what psychologists do with their measurements once they’ve made them (Description).

Measurement

We live in a world of clocks and calendars, scales and yardsticks, odometers and thermometers. Measurement is not just a part of modern science, it is a part of modern life. But regardless of what we want to measure—the intensity of an earthquake, the size of an elephant, or the attitude of a registered voter—we must always do two things. First, we must define the property we want to measure, and second, we must find a way to detect it (Figure 2.1). For example, to measure a person’s level of happiness, we would start by generating an operational definition, which is a description of a property in measurable terms. So, for example, we might operationally define happiness as “a person’s self-assessment” or “the amount of dopamine in a person’s brain” or “the number of times a person smiles in an hour.” Once we had this definition in hand, we would need to find a detector—that is, some sort of instrument or device that can detect the property as we just defined it—such as a rating scale (to detect a person’s self-assessment), a carbon electrode (to detect dopamine in the brain), or an electromyograph (to detect a smile).

What makes a good operational definition? Construct validity is the extent to which the thing being measured adequately characterizes the property. For example, most of us would consider the frequency with which a person smiles to be a reasonable way to operationally define the property called “happiness” because we all know from our own experience that happy people tend to smile more than unhappy people do. Do they also eat more or talk more or spend more money? Well, maybe. But then again, maybe not. And that’s why most psychologists would consider “smiles per hour” to be a reasonable way to operationally define happiness, but they would not feel the same way about “number of chocolates eaten” or “number of words spoken” or “number of dollars spent.” To a large extent, construct validity is in the eye of the beholder, and an operational definition is said to have construct validity when most beholders agree that it adequately characterizes a property.

What makes a good detector? The two key features of a good detector are power, which refers to a detector’s ability to detect the presence of differences or changes in the magnitude of a property, and reliability, which refers to a detector’s ability to detect the absence of differences or changes in the magnitude of a property. If a person smiles a bit more often on Tuesday than on Wednesday, a powerful smile-detector will detect different amounts of smiling on those two days. If a person smiles exactly as much on Wednesday as she did on Tuesday, then a reliable smile-detector will detect identical amounts of smiling on those two days. A good detector detects differences or changes in the magnitude of a property when they do exist (power), but not when they don’t (reliability).
Demand Characteristics: Doing What Is Expected

Once we have an operational definition that has construct validity, and a detector that is both powerful and reliable, are we finally ready to measure some behavior? Yes—as long as we want to measure the behavior of a worm, a thunderstorm, a stock market, or anything else that doesn’t care about being observed. But if we want to measure the behavior of a human being, then we still have some work to do because when human beings know they are being observed, they will often try to behave as they think the observer wants or expects them to. **Demand characteristics** are **those aspects of an observational setting that cause people to behave as they think someone else wants or expects.** If a friend asked, “Do you think I’m smart?” you would probably say yes whether you meant it or not. You know what your friend is hoping to hear and so you dutifully supply it. Similarly, if a researcher asked, “Do you think it is wrong to cheat on exams?” then you would probably say yes, if only because you know that’s the response the researcher expects of you. A study that asked such a question would be said to have demand characteristics because the question “demands” or requires participants to give a response that may or may not reflect his or her true feelings. How can we avoid demand characteristics?

**Naturalistic Observation** One way that psychologists avoid demand characteristics is by observing people without their knowledge. **Naturalistic observation** is **a technique for gathering scientific information by unobtrusively observing people in their natural environments.** Naturalistic observation has shown that the biggest groups leave the smallest tips in restaurants (Freeman et al., 1975), that golfers are most likely to cheat when they play several opponents at once (Erffmeyer, 1984), that men usually don’t approach the most beautiful woman at a club (Glenwick, Jason, & Elman, 1978), and that Olympic athletes smile more when they win the bronze medal than the silver medal (Medvec, Madey, & Gilovich, 1995). Each of these conclusions is the result of measurements made by psychologists who observed people who didn’t know they were being observed by psychologists. It seems unlikely that the psychologists could have made the same

Power and Reliability at the Olympics
Usain Bolt ran the 100 meter race in 9.58 seconds, and Yohan Blake ran it in 9.75 seconds. If judges did not have powerful speed-detectors, they might have mistakenly concluded that the two men were tied. Carmelita Jeeter and Torie Bowie both ran the race in 10.83 seconds. If judges did not have reliable speed-detectors, they might have mistakenly concluded that one of them ran faster than the other.

Countries that have a faster pace of life tend to have higher rates of heart disease. How do researchers measure the “pace of life”? They make naturalistic observations—in this case, by measuring the average walking speed of pedestrians in different cities. By the way, the fastest pedestrians are the Irish (left) and the slowest are the Romanians (right) (Levine & Norenzayan, 1999).
observations if the diners, golfers, clubbers, and athletes had realized that they were being observed.

Unfortunately, naturalistic observation isn’t always practical. First, some events just don’t occur naturally. If we wanted to know whether people who have undergone sensory deprivation perform poorly on fine-motor tasks, we would have to stand on the street corner for a very long time before we just so happened to spot a bunch of blindfolded people with earplugs trying to text with one hand. Second, some events can only be observed through direct interaction, such as by conducting an interview or by hooking someone up to a heart rate monitor. If we wanted to know how often people worried about dying, how accurately they remember their high school graduations, or how much electrical activity their brains produce when they feel jealous, then hiding in the bushes and watching them with binoculars just won’t do. What can we do instead?

Privacy and Control When naturalistic observation isn’t possible, psychologists have a number of other techniques for avoiding demand characteristics. For instance, people are less likely to be influenced by demand characteristics when they can’t be identified as the authors of their actions, and psychologists often take advantage of this fact by allowing people to respond privately (e.g., by having them complete questionnaires when they are alone) and/or anonymously (e.g., by not collecting personal information, such as people’s names or addresses). In addition, psychologists may avoid demand characteristics by measuring behaviors that are not under a person’s voluntary control. If a psychologist asked whether you were interested in stupid celebrity gossip, you might lie and say no. Because you can control what you say, that measure is susceptible to demand. But if the psychologist instead gauged your interest in stupid celebrity gossip by measuring the dilation of your pupils as you paged through the latest issue of *Us Weekly*, there would be no way for you to hide the fact that you are deeply interested in the details of Cardi B and Offset’s secret wedding. C’mon. You know you are.

Unawareness One of the best ways to avoid demand characteristics is to make sure that the people who are being observed are unaware of the true purpose of the observation. People can’t try to behave how they should behave if they don’t know how they should behave. For example, if you didn’t know that a psychologist was studying the effects of classical music on mood, you wouldn’t feel obligated to smile when Bach’s *Concerto #1 in F Major* started playing in the background. That’s why psychologists typically don’t reveal the true purpose of an observation to the people being observed until the study is over. And it’s also why psychologists sometimes mislead people by telling them that they are studying one thing when they are really studying another, or by giving people pointless tasks or asking pointless questions simply so that people can’t easily guess the study’s true purpose. (We will discuss the ethical implications of misleading people later in this chapter.)

Observer Bias: Seeing What Is Expected

More than half a century ago, students in a psychology class were asked to measure the speed with which a rat learned to navigate a maze (Rosenthal & Fode, 1963). Some students were told that their rat had been specially bred to be a slow learner while others were told that their rat had been specially bred to be a fast learner. The truth was that the rats were all exactly the same breed. Nonetheless, students who thought they were measuring the speed of a slow learner reported that their rats took 3.47 minutes to navigate the maze, whereas students who thought they were measuring the speed of a fast learner reported that their rats took 2.35 minutes to navigate the maze. In other words, the measurements revealed precisely what the experimenters had expected them to reveal, even though those expectations had no basis in reality.

There are two reasons why this happened. First, expectations can influence observations. Just think about all the decisions the students had to make when they were measuring the speed of their rat. If a rat puts one leg over the finish line, does that count as finishing, or does it have to have all four legs over the line? If a rat falls asleep,
should the stopwatch keep running or should it be stopped until the rat wakes up? If a rat runs the maze in 18.5 seconds, should that number be rounded up to 19 or down to 18 before it is recorded? How students answered questions like these may have depended on whether they thought their rats were fast or slow learners. “Micky is a fast learner, so let’s just call it 18” or “Topo sure is a slow learner. When will he ever drag his last leg over the finish line?” Second, expectations can influence reality. Students who expected their rats to be fast learners unwittingly did things that might have helped that learning along. For instance, students handled their rats more often and more gently when they thought they were fast learners, and the rats may well have responded to this superior treatment by turning in a superior performance. The students probably tried their best to be fair and objective, but their expectations nonetheless seem to have influenced both their rat’s behavior and their observations of it.

This problem is so significant that psychologists have given it a name: Observer bias is the tendency for observers’ expectations to influence both what they believe they observed and what they actually observed. To avoid observer bias, psychologists use a number of techniques, the most common of which is called the double-blind study, which is a study in which neither the researcher nor the participant knows how the participants are expected to behave. For example, if we wanted to know whether people smile more when listening to classical music than to hip-hop, we might give participants a task to do while one of these two kinds of music played in the background, and have a research assistant watch them and record how often they smiled. We would take steps to ensure that our participants did not know what we were studying so that they would not feel obliged to behave as we expected them to; but we would also take steps to ensure that the research assistants did not know how the participants were expected to behave, perhaps by having the research assistants wear noise-cancelling headphones so that they wouldn’t know which kind of music was playing as they recorded the participants’ rate of smiling. If the research assistants don’t have expectations, then their expectations cannot influence either their observations or their participants’ behavior. That’s why psychologists typically make sure that the people who are making the observations in a study are as “blind” to the hypothesis as are the people who are being observed.

**Description**

You now know how to measure. That is, you know how to generate an operational definition of a property that has construct validity, how to find a reliable and powerful detector of that property, and how to use that detector while avoiding demand characteristics and observer bias. So who are you going to measure? Psychologists rarely measure the properties of an entire population, which is a complete collection of people—such as the population of human beings (about 7 billion), the population of Californians (about 38 million), or the population of people with Down syndrome (about 1 million). Rather, they tend to measure the properties of a sample, which is a partial collection of people drawn from a population. We’ll talk later about how psychologists typically get their samples, but for now, just imagine you have taken a sample and measured its properties.

Congratulations! You are now the proud owner of some data. Your data probably take the form of a big spreadsheet filled with numbers that represent things like the frequency with which 211 people smiled in an hour, or the amount of time it took 37 rats to run a maze. And if you are like most researchers, you will look at your data and feel precisely as ignorant as you were before you started, because for most researchers, a big spreadsheet full of numbers just isn’t very informative. But that’s okay, because psychologists have two techniques for making sense out of big spreadsheets full of numbers: graphic representations and descriptive statistics. Let’s have a look at each of these.

**Graphic Representations: Picturing the Data**

Ever wonder why a map is so much better than a list of step-by-step directions? Vision is our most sophisticated sense, so most people find it easier to understand facts when...
Figure 2.2

**Frequency Distributions** The graph in the top panel uses bars to show the number of male residents (shown in green) and female residents (shown in orange) of Somerville, MA, who rated their happiness at each value on a 10-point scale. The graph in the middle panel shows the same data using smooth lines. The graph in the bottom panel shows how these real data compare to a hypothetical normal distribution (shown in purple).

**frequency distribution** A graphic representation showing the number of times that the measurement of a property takes on each of its possible values.

Those facts are represented visually rather than by numbers or words. That’s why scientists typically make sense of their data by creating pictures or graphic representations. The most common kind of graphic representation is called a frequency distribution, which is a graphic representation showing the number of times in which the measurement of a property takes on each of its possible values.

The frequency distribution in the top panel of Figure 2.2 shows the results of a city-wide census in which residents of Somerville, Massachusetts, were asked to report their current level of happiness by using a rating scale that ranged from 1 (very unhappy) to 10 (very happy). The property being measured was happiness, the operational definition of happiness was a scale rating, and all the possible values of that rating (1 through 10) are shown on the horizontal axis. The vertical axis shows the number of men and women who responded to the census and used each of these values to rate their happiness. (People who identified as something other than male or female are not shown in this graph.) So, for example, the graph shows that 1,677 women and 646 men rated their happiness as 8 on a 10-point scale. The middle panel shows exactly the same data, but it uses smooth lines rather than bars, which is an equally common way to display a frequency distribution.

In a single glance, these graphs reveal things about the sample that a page full of numbers does not. For instance, when you look at the shape of the distributions, you instantly know that the people in this sample tend to be fairly happy. Your eyes also tell you that far fewer men than women responded to the census, but that the general shapes of the two distributions are roughly the same, which suggests that the men in this sample are generally about as happy as the women. You can see other things too—for instance, that 8 is the most popular rating for both men and women, and that both genders are about 3 times as likely to rate their happiness as 10 than to rate it as 1. All of that information from a simple drawing!

The distributions in the middle panel are negatively skewed (which means that they lean to the right) rather than positively skewed (which means that they lean to the left) because very few people use numbers that are below the midpoint of the scale to rate their happiness. Although a frequency distribution can have just about any shape, a special shape is shown in purple in the bottom panel of Figure 2.2. As you can see, this distribution has a peak exactly in the middle and then trails off in the same way at both ends. This distribution is unskewed or symmetrical, which is to say that the left half is a mirror image of the right half. Although frequency distributions showing data from the real world are rarely as symmetrical as the purple one is, they are often fairly close, especially when the amount of data used to construct them is quite large. That’s why the distribution shown in purple is called the normal distribution, which is a mathematically defined distribution in which the frequency of measurements is highest in the middle and decreases symmetrically in both directions. The normal distribution is often called a bell curve, but if you’d like to be single for the rest of your life you can refer to it in public as a Gaussian distribution.

**Descriptive Statistics: Summarizing the Data**

A frequency distribution depicts every measurement in a sample and thus provides a full and complete picture of that sample. But sometimes a full and complete picture is more than we want to know. When we ask a friend how she’s been doing, we don’t really want her to show us a frequency distribution of her happiness ratings on each day of the previous 6 months. Rather, we want her to provide a brief summary statement that captures the essential information from that graph—for example,
“I’ve been doing pretty well,” or maybe “I’ve been having a few ups and downs.” In psychology, brief summary statements that capture the essential information from a frequency distribution are called descriptive statistics.

The two most common kinds of descriptive statistics are those that describe the central tendency of a frequency distribution, and those that describe the variability in a frequency distribution. What do these terms mean? Descriptions of central tendency are statements about the value of the measurements that tend to lie near the center or midpoint of the frequency distribution. When a friend says that she’s been “doing pretty well,” she is describing the central tendency (or approximate location of the midpoint) of the frequency distribution of her happiness ratings over time (see Figure 2.3). Descriptions of variability, on the other hand, are statements about the extent to which the measurements in a frequency distribution differ from each other. When your friend says that she has been “having some ups and downs,” she is describing the variability among her happiness ratings. Let’s dig a little deeper into each of these concepts.

Central Tendency: Where Is The Middle Of The Distribution? The three most common descriptions of central tendency are: the mode (the value of the most frequently observed measurement); the mean (the average value of all the measurements); and the median (the value that is in the middle, i.e., greater than or equal to half the measurements and less than or equal to half the measurements). Figure 2.4 shows how each of these descriptive statistics is calculated.

Why do we need three different measures of central tendency? When a distribution is normal, we don’t, because these three measures all have exactly the same value. But when a distribution is skewed, the mean gets dragged toward the end of the long tail, the median follows it but doesn’t get as far, and the mode stays put at the hump (see Figure 2.5). When these three measures have different values, then calculating just one of them can provide a misleading picture of the data. For instance, if you measured the annual income of the roughly 400,000 households in Seattle, Washington, you’d find that the mean is about $84,000. But that sample includes a few unusual households—for example, the households of Bill Gates (whose net worth is $95 billion) and Jeff Bezos (whose net worth was $150 billion when you started reading this sentence but is already more).

Seattle is home to a few hundred ridiculously rich people, so considering the mean alone might lead you to overestimate the affluence of the typical Seattleite. But if, in addition to the mean, you also calculated the median, you’d find it has the considerably lower value of $62,000. Each measure can potentially be misleading in isolation, but when considered together they paint a much more accurate picture of Seattle as a middle-class city with a few ultra-wealthy residents. You should be very suspicious whenever you hear some new fact about the “average person” but don’t hear anything about the median, the mode, or the shape of the frequency distribution.

Figure 2.4
Calculating Descriptive Statistics This frequency distribution shows the data from 20 individuals who rated their happiness on a 10-point scale. Descriptive statistics include measures of central tendency (such as the mean, median, and mode) and measures of variability (such as the range and the standard deviation).

---

Two Kinds of Descriptive Statistics

The most common descriptive statistics describe a frequency distribution’s central tendency (where do most of the measurements lie?) and variability (how much do the measurements differ from one another?).

- **Mode** The value of the most frequently observed measurement.
- **Mean** The average value of all the measurements.
- **Median** The value that is greater than or equal to half the measurements and less than or equal to half the measurements.

![Normal Distribution](image)

- Mode = 5 because there are four 5’s, and fewer of every other value.
- Mean = 5.95 because (1+2+3+4+4+5+5+5+6+6+7+7+7+8+8+8+9+9+10) / 20 = 5.95
- Median = 6 because 11 of the values are ≥ 6 and 11 of the values are ≤ 6.
Variability: How Wide Is The Distribution? Descriptions of central tendency are statements about where the measurements in a frequency distribution tend to lie relative to the values on the vertical axis. Descriptions of variability, on the other hand, are statements about where the measurements in a frequency distribution tend to lie relative to each other. Descriptions of variability tell us how much the measurements differ from each other, or roughly how wide the distribution is. The simplest measure of variability is the range, which is the value of the largest measurement in a frequency distribution minus the value of the smallest measurement. When the range is small, the distribution has less variability than when the range is large. The range is easy to compute, but like the mean it can be dramatically influenced by a few extreme measurements. If someone told you that the temperature in San Diego, California, ranged from 25°F to 111°F, you might get the mistaken impression that San Diego has a remarkably varied climate when, in fact, it has a remarkably stable climate, rarely getting warmer than 76°F or colder than 50°F. The temperature in San Diego did hit 25°F and 111°F, but just once in the last 100 years.

Other measures of variability aren’t as easily distorted by extreme values, but they are a little trickier to compute. One such measure is the standard deviation, which is a statistic that describes how each of the measurements in a frequency distribution differs from the mean. (Technically, it is the square root of the averaged squared difference between each measurement and the mean, and if you say that three times backwards you will enter a hypnotic trance that requires medical attention.) In other words, the standard deviation is an estimate of how far, on average, the various measurements are from the center of the distribution. It is worth noting that variability and central tendency are independent features of a frequency distribution. As FIGURE 2.6 shows, two frequency distributions can have the same central tendencies but different amounts of variability, just as they can have the same amounts of variability but different central tendencies.

Figure 2.5
Differently Shaped Distributions
When a frequency distribution is normal, the mean, median, and mode all have the same value, but when it is positively or negatively skewed, these three measures of central tendency have quite different values.

Figure 2.6
Distributions Can Differ In Terms Of Variability Or Central Tendency
The panel on the left shows two distributions with the same central tendency but different amounts of variability. The panel on the right shows two distributions with the same amount of variability but different central tendencies.

Build to the Outcomes
1. What are the essential features of an operational definition?
2. What two properties must a good detector have?
3. What techniques do psychologists use to avoid demand characteristics and observer bias?
4. What is the difference between a population and a sample?
5. What is a frequency distribution, and what makes the normal distribution special?
6. What is the difference between measures of central tendency and measures of variability?
7. Why can a single descriptive statistic (mode, mean, or median) sometimes be misleading?
Methods of Explanation: Figuring Out Why People Do What They Do

In 1639, a pastor named John Clarke suggested that “Early to bed and early to rise, makes a man healthy, wealthy, and wise.” People have been repeating this little rhyme ever since, but is there any truth to it? The methods you’ve learned about so far would allow you to measure the health, wealth, and wisdom of a sample of people, as well as draw some pictures and make some summary statements about the measurements you made. That’s nice, but it isn’t what you want to know. You want to know if all that health and happiness and wisdom that some people seem to have was caused by getting in and out of bed early. Is there any way to use your new measurement skills to answer questions like this one?

Indeed there is, and that’s what this part of the chapter is all about. In the first section (Correlation), we’ll examine a technique that can tell us whether two properties—such as wisdom and bedtime—are in fact related. In the second section (Causation), we’ll examine a technique that can tell us whether one of these properties (e.g., bedtime) actually causes the other (e.g., wisdom). And in the third and final section (Drawing Conclusions) we’ll examine the kinds of conclusions these two techniques do and do not allow us to draw.

Correlation

Speaking of early to bed and early to rise, how much sleep did you get last night? Okay, now how many U.S. presidents can you name without asking Alexa, Siri, or Soledad O’Brien? If you were to ask a dozen college students those two questions, you’d probably find that the students who had gotten a solid eight hours could recall more presidents, on average, than students who had pulled an all-nighter. If you kept careful track of their responses (yes, that means writing them down), you would probably find that you had a series of measurements much like the ones shown in Table 2.1. And from those measurements you would probably conclude that sleep deprivation causes poor memory.

But wait a minute. Those measurements simply tell you how much sleeping and remembering the people in your sample did. So what led you to conclude that there was a causal relationship between sleeping and remembering?

Synchronized Patterns of Variation

When you asked a dozen college students how much they slept the night before and then asked them to name as many presidents as possible, you may not have realized that you were doing three important things:

• First, you were measuring a pair of variables, which are properties that can take on different values. When you took your first algebra course you were probably horrified to learn that everything you’d been taught in grade school about the distinction between letters and numbers was a lie, that mathematical equations could contain Xs and Ys as well as 7s and 4s, and that the letters were called variables because they could have different values under different circumstances. Same idea here. When you asked about sleep, you were measuring a variable whose value could vary from 0 to 24, and when you asked about presidents, you were measuring a second variable whose value could vary from 0 to 44.1

1 If you think this number should actually be 45 because current president Donald Trump is the 45th president, then you are forgetting that Grover Cleveland was both the 22nd and the 24th president, which means that someone who recalls the name of every president will recall just 44 names.
• Second, you did this again. And then again. And then again. That is, you made a series of measurements rather than just one.

• Third and last, you looked at the measurements you made and tried to discern a pattern. If you look at the second column of Table 2.1, you will see that as you move your eyes from top to bottom, the numbers vary from small to large. We could say that this column of numbers has a particular pattern of variation. Now, if you compare the third column with the second, you will notice the numbers there have a similar (though not identical) pattern of variation. With just a few exceptions, they also tend to vary from small to large as you move your eyes from top to bottom. The patterns of variation in these two columns are somewhat synchronized, and this synchrony is known as a correlation, which occurs when variations in the value of one variable are synchronized with variations in the value of the other. When patterns of variation are synchronized, two variables are said to be correlated—or “co-related.”

Now here’s why this matters: When two variables are correlated, you can use your knowledge of the value of one variable to predict the value of the other variable without having to measure it! For example, imagine that after you collected your sleep and memory data, you ran into a friend who lamented that she’d only gotten two hours of sleep the previous night. Without even asking her, you already have a fairly good idea of how many presidents she could name—somewhere in the 15 to 25 range, right? And if you ran into another friend, asked him to name all the presidents, and found that he named them all with no problem, you would have a pretty good idea of how many hours he’d slept the night before—somewhere in the 7 to 9 range, right?

Correlations allow us to make educated guesses about measurements without having to do the hard work of measuring. If you know that a man is 6’7” tall, you can guess that he probably weighs more than 200 pounds because height and weight are correlated. If you know that a woman earns $1 million per year, you can guess that she probably went to college because income and education are correlated. As you can see, correlations are knowledge—and knowledge is power! But how much power?

If you predict that a rich woman went to college, that a tall man is heavy, or that a sleep-deprived person won’t be able to name many presidents, you will be right far more often than you’ll be wrong. But you won’t be right every time. Oprah Winfrey didn’t finish college and she is worth $3 billion; NBA player Terrance Ferguson is 6’7” but weighs just 184 pounds; and even though the authors of this textbook got an average of 6.1 hours of sleep last night, we have just confirmed that they can name an average of 37 presidents. The point is that correlations do allow us to make predictions, but those predictions are not always accurate.

**Measuring the Direction and Strength of a Correlation**

Statisticians have developed a way to estimate just how accurate predictions are likely to be by measuring the direction and strength of the correlation on which the predictions are based. Direction is easy to measure because the direction of a correlation is either positive or negative. A positive correlation exists when two variables have a “more-is-more” relationship. So, for example, consider pastor John Clarke’s suggestion that the properties “healthy, wealthy, and wise” tend to be bundled together such that people who have a lot of one also have a lot of the others. When we say that more health is associated with more wealth, we are describing a positive correlation. Conversely, a negative correlation exists when two variables have a “more-is-less” relationship. When we say that more health is associated with less poverty, we are describing a negative correlation. More wealth and less poverty are, of course, two ways of saying the same thing, and any relationship can be described either positively or negatively. Scientists tend to use whichever description makes the most sense given the particular variables they are studying.
The direction of a correlation is easy to measure, but the strength is a little more complicated. The correlation coefficient is a mathematical measure of both the direction and strength of a correlation, and it is symbolized by the letter $r$. The first thing to know about it is that like most measures, $r$ has a limited range. What does that mean? Well, if you measure the number of hours of sunshine you experience tomorrow, the value of your measurement will definitely be somewhere in the range of 0 to 24. It cannot possibly be $-7$ or 329, which is to say that any numbers outside the limited range of 0 to 24 are meaningless. Similarly, the value of $r$ can range from $-1$ to 1, and numbers outside that range are meaningless.

So what do the numbers inside that range mean? Let’s start by saying what the two most extreme numbers (1 and $-1$) and the middle number (0) mean. For example, consider what it would mean if the correlation between education and wealth were 1, or $-1$, or 0.

- If every time the value of a variable increases by a certain amount, the value of a second variable also increases by a certain amount, then the variables have a perfect positive correlation and $r = 1$ (see the top graph in **FIGURE 2.7**). For example, if every x-unit increase in education were associated with a y-unit increase in wealth, then the correlation between these two variables would be perfectly positive. In this case, if you know exactly how healthy a person is, you can predict down to the penny how wealthy he or she is.

- If every time the value of a variable increases by a certain amount the value of a second variable decreases by a certain amount, then the variables have a perfect negative correlation and $r = -1$ (see the middle graph in Figure 2.7). For example, if every x-unit increase in education were associated with a y-unit decrease in wealth, then the correlation between these two variables would be perfectly negative. In this case, too, if you know exactly how healthy a person is, you can predict down to the penny how wealthy he or she is.

- If every time the value of a variable increases by a certain amount the value of a second variable neither increases nor decreases systematically, then the variables have no correlation and $r = 0$ (see the bottom graph in Figure 2.7). For example, if an x-unit increase in health were sometimes associated with an increase in wealth, sometimes associated with a decrease in wealth, and sometimes associated with no change in wealth at all, then there would be no correlation between the two variables. In this case, knowing how healthy a person is would not allow you to predict anything about his or her wealth.

Correlations of 1 and $-1$ are extremely common—in textbooks. They are almost unheard of in the real world. Yes, in the real world, health and wealth do have a positive correlation—that is, people who have more of one tend to have more of the other—but they have an imperfect positive correlation. A person with an extra unit of health probably also has some extra wealth, but that extra unit of health doesn’t tell us exactly how much extra wealth that person has. Statements like “increases in health are associated with increases in wealth” do not mean that when the health of a group of people increases by 1 unit, every single individual in that group also experiences a 1-unit increase in wealth. Rather, we mean that for most of the people in the group, wealth increases a bit; for many it increases a lot, for a few it doesn’t increase at all, and for some it may actually decrease.

There are always exceptions to the rule, which means that the value of $r$ will lie somewhere between 0 and 1—and the number and size of the exceptions indicate where in that range it will lie. If there are just a few exceptions to the rule, then $r$ will be closer to 1 than to 0, and predictions about someone’s wealth based solely on knowledge of his or her health will be fairly accurate. But as the number of exceptions increases, then the value of $r$ will begin to move toward 0 and such predictions will become less and less accurate, until finally, when $r$ hits 0, they will become completely useless and you might as well just guess.

**FIGURE 2.7**

**Graphing Correlations** This is what the three different kinds of correlations look like when graphed.

---

**correlation coefficient ($r$)** A mathematical measure of both the direction and strength of a correlation, which is symbolized by the letter $r$. 

---

FIGURE 2.8 illustrates three cases in which two variables are positively correlated but have different numbers of exceptions. In each graph, the diagonal line shows the data you would expect to see if the two variables were perfectly positively correlated. In a sense, the diagonal line represents “the rule” that “when one variable increases by 1 unit, then the other variable increases by 1 unit,” therefore every dot that is not on the diagonal line is “an exception to the rule.” The farther a dot is from the line, the greater an exception it is. As you can see, the size and number of the many “exceptions to the rule” change the value of $r$ quite dramatically. The two variables shown in the graph at the top of Figure 2.8 have a strong correlation of $r = .5$, the two variables shown in the graph in the middle have a moderate correlation of $r = .3$, and the two variables shown in the graph at the bottom have a weak correlation of $r = .1$. The only difference between the graphs of the strong and weak correlations is that the latter has more data points that are farther from the line.

A few pages ago, you thought $r$ was just another Sesame Street letter of the day. Now you know that $r$ is a measure of both the direction and strength of the relationship between two variables. The sign of $r$ (plus or minus) tells you the direction of the relationship (positive or negative), and the absolute value of $r$ (between 0 and 1) tells you about the number and size of the exceptions to the rule—hence, about how accurate you are likely to be when using knowledge of one variable to make predictions about the other. Big Bird would be so proud.

Causation

The mathematics of correlation may have been new to you, but the concept surely was not. You’ve been noticing correlations all your life—between height and weight, between sleep and memory, between age and musical preference, and so on. Natural correlations are the correlations we observe in the world around us; and although the mere observation of a synchronized pattern of variation tells us that two variables have a relationship, it cannot tell us what kind of relationship those variables have. For example, many studies (Anderson & Bushman, 2002; C. A. Anderson et al., 2003, 2017; Huesmann et al., 2003) have observed a positive correlation between the aggressiveness of a child’s behavior and the amount of violence to which that child is exposed through media such as television, movies, and video games. The more violence children see, the more aggressively they tend to behave. These two variables clearly have a relationship, but what kind of relationship do they have? In other words, why are they correlated?

The Third-Variable Problem: Correlation Is Not Causation

One possibility is that exposure to media violence (let’s call it variable $X$) causes aggressiveness (let’s call it variable $Y$). For instance, media violence may teach children that aggression is an acceptable way to vent anger or solve problems. We’ll denote this possible relationship as $X \rightarrow Y$, which simply means $X$ causes $Y$. A second possibility is that $Y \rightarrow X$, which is to say that aggressiveness causes exposure to media violence. For example, children who are naturally aggressive may be especially inclined to play violent video games and watch violent movies. When you first heard that exposure to media violence and aggressiveness were positively correlated, you probably thought of these two possibilities yourself.

But there is an additional possibility you might not have considered. It is possible that another variable—a heretofore unnamed “third variable” that we will call variable $Z$—causes children to behave aggressively and also causes children to be exposed to media violence (see FIGURE 2.9). For instance, it might be that a lack of adult supervision ($Z$) allows children to get away with playing violent video games that adults would normally prohibit ($Z \rightarrow X$), and it may also be that a lack of adult supervision allows children to get away with bullying others simply because no adult is around to which that child is exposed through media such as television, movies, and video games. The more violence children see, the more aggressively they tend to behave. These two variables clearly have a relationship, but what kind of relationship do they have? In other words, why are they correlated?
Methods of Explanation: Figuring Out Why People Do What They Do

Experimentation: Establishing Causation

Experimentation is a technique for establishing the causal relationship between variables. The logic underlying experimentation is simple: There are three possible causes of any correlation, so if we can eliminate two of them, then the one that remains must be the real one. If we know that X did not cause Y, and if we know that Z did not cause X and Y, then we know that Y must have caused X. The logic is simple, but exactly how do experiments eliminate two of the three possible causes? They do this by using a pair of techniques called manipulation and random assignment. Let’s explore each of them in turn.

Manipulation: Making Different Conditions

Imagine that one evening you are working on your laptop, when you notice that your internet connection has suddenly slowed to a crawl. Your roommate is upstairs playing with his new Xbox, which you suspect could be sucking up bandwidth, which is making your laptop run slow. How would you test your suspicion? You now know that simply observing a natural correlation won’t be of much help because even if your laptop ran slow every time your roommate used his Xbox and ran fast every time he didn’t, you still couldn’t conclude that the Xbox was causing the slowdown because of the third-variable problem. For instance, it is possible that your roommate only uses his Xbox in the evenings, which happens to be the time when a whole lot of other people in your neighborhood are home watching Netflix, playing video games, downloading music, and otherwise sucking up bandwidth. “Evening” may be a third variable that causes your roommate to fire up the Xbox and that also causes your laptop to slow down.

So how can you tell whether your slowdown is actually being caused by the Xbox? Rather than observing the natural correlation between your roommate’s Xbox usage and your laptop’s speed, you could go upstairs when your roommate isn’t home and turn his Xbox on and off while observing the speed of your laptop. If your laptop slowed down every time you turned the Xbox on and then sped up again every time you turned the Xbox off, you would know that the Xbox was in fact the cause of the slowdown. What you just did is called manipulation, which is a technique for determining the causal power of a variable by actively changing its value. Rather than measuring two variables, as we do when we observe correlations, experiments require that we manipulate one variable—that is, actively change its value—and then measure the other. Changing the value of the Xbox from on to off is an example of manipulation.

The same technique could be used to determine whether exposure to media violence causes aggressiveness. For instance, we could invite some children to our laboratory and give them one of two experiences: Half could be given a violent video game to play, and the other half could be given a nonviolent video game to play. These two experiences are called the conditions of our study, and we could refer to them as “the violent exposure condition” and “the nonviolent exposure condition.” At conclusion of the experiment, we could then observe the aggressiveness among the children who had been given violent and nonviolent video games. If the group that had been given violent video games was more aggressive than the group that had been given nonviolent video games, we could conclude that exposure to media violence causes aggressiveness.

third-variable problem The fact that the natural correlation between two variables cannot be taken as evidence of a causal relationship between them because a third variable might be causing them both.

experimentation A technique for establishing the causal relationship between variables.

manipulation A technique for determining the causal power of a variable by actively changing its value.
Can you see the difference between these two ads? Don’t worry. It’s subtle. The one and only difference is that the ad on the left has a “Learn More” button and the ad on the right has a “Sign Up” button. By manipulating the label on the button, the advertiser was able to determine whether a label can cause people to click. And it can! The “Learn More” label led 15% more Facebook users to click (Karlson, 2016).

At the end of an hour of game playing, we could measure the children’s aggressiveness, perhaps by observing whether they push to get to the front of a line or by asking whether they think it’s okay for someone to kick a growling dog. Then we could compare the measurements of aggressiveness in one condition with the measurements of aggressiveness in the other condition.

When we compare these measurements across conditions, we are essentially asking whether the value of aggressiveness went from low to high when exposure to media violence went from low to high. If you think about it, we would essentially be computing the correlation between the variable we manipulated (exposure) and the variable we measured (aggressiveness). Now here is the cool part: Because we manipulated exposure rather than just measuring it, we could instantly eliminate two of the possible causes of the correlation we had just observed. Specifically, we’d know for sure that aggressiveness did not cause exposure to media violence, and we’d know that lack of adult supervision did not cause exposure to media violence—and we’d know these two things because we know what did cause exposure to media violence: We did! And that would leave just one possibility: Exposure to media violence must have caused aggressiveness, which is what we wanted to know in the first place.

Experimentation is a technique that allows us to establish the causal relationship between variables by taking three simple steps:

1. **Manipulate**: The first step in an experiment is to manipulate a variable. The variable that is manipulated in an experiment is called the independent variable because its value is determined entirely by the experimenter and therefore does not depend on—or is “independent of”—the participants. A manipulation creates at least two conditions: in our example, a violent exposure condition and a nonviolent exposure condition.

2. **Measure**: The second step in an experiment is to measure a variable. The variable that is measured in an experiment is called the dependent variable because its value does “depend on” the participants.

**independent variable** The variable that is manipulated in an experiment.

**dependent variable** The variable that is measured in an experiment.
3. **Compare**: The third step in an experiment is to compare the value of the variable in one condition with the value of the variable in the other. If the values differ on average, then we know that changes to the value of the independent variable caused changes to the value of the dependent variable. **Figure 2.10** shows exactly how manipulation works (and the HOT SCIENCE box on p. 2-22 shows how nature occasionally does experiments by creating manipulations of her own).

**Random Assignment: Making Sure Conditions Differ in Just One Way**

Manipulation is an essential ingredient in experimentation. But what kind of recipe has just one ingredient? In fact, experimentation has a second and equally important ingredient, and to understand what it is and what it does, let’s return to the study we just designed in which children were given violent or nonviolent video games to play and then their aggressiveness was measured an hour later. When the children show up at our laboratory to participate in our study, how should we decide which child will play which game?

One possibility is that we could be polite for a change and ask each child which kind of game he or she would prefer to play. So imagine we did that and that half the children choose to play a violent game and the other half choose to play a nonviolent game. We let them play their preferred game for an hour, then measure their aggressiveness and discover that the children who played the violent game were, in fact, more aggressive than those who played the nonviolent game. Could we now conclude that playing video games causes aggressiveness? No! But why not? After all, we manipulated exposure to media violence, switching it on and off as though it were an Xbox, and then we measured aggressiveness and found that it went on and off too. We manipulated, we measured, and we compared—we did all three of the three steps. So where did we go wrong?

We went wrong by letting the children select which video game to play—because children who select violent video games probably differ in many ways from children who don’t. They may be older or meaner. They may be younger or sweeter. They may have different brain chemistry, different home environments, different genes, different talents, different numbers of siblings, or different levels of adult supervision. The list of possible differences is endless. The whole point of doing the experiment was to create two conditions whose participants differed from each other in one and only one way, namely, in terms of how much media violence they were exposed to in our laboratory. By letting the children decide for themselves which condition of the experiment they would experience, we ended up with two conditions whose participants may have differed in many ways—and every one of those differences is a third variable that could potentially have caused the differences in aggressiveness that we observed when comparing across conditions. Apparently, it just doesn’t pay to be polite. **Self-selection** is a problem that occurs when anything about a participant determines the participant’s condition; and as you can see, it is a problem that leaves us with the same conundrum that led us to do an experiment in the first place.

So how should we determine which video game each child will play? We should flip a coin. **Random assignment** is a procedure that assigns participants to a condition by chance. Just think about what would happen if we used a coin flip to randomly assign each child to play either a violent or nonviolent video game. As **Figure 2.11** shows, the first thing that would happen is that roughly half the children would be assigned to play violent video games and roughly half would be assigned to play nonviolent video games. Coins land heads up about as often as they land tails up, so if we use a coin to randomly assign children to one of two conditions, then the two conditions should have roughly equal numbers of children.

**Figure 2.10**

The Three Steps of Experimentation

Manipulate the independent variable, measure the dependent variable, then compare the values across conditions. Notice that “comparing the values” just means “looking for a synchronized pattern of variation between two variables,” which is why the table to the far right in this figure looks so much like Table 2.1.

There is no evidence that Louise Hay’s techniques can cure cancer. But even if cancer victims who bought her books did show a higher rate of remission than those who didn’t, there would still be no evidence because buyers are self-selected and thus may differ from nonbuyers in countless ways.

---

**self-selection** A problem that occurs when anything about a participant determines the participant’s condition.

**random assignment** A procedure that assigns participants to a condition by chance.
Hate Posts and Hate Crimes: Not Just a Correlation

In 2004, a college student named Mark Zuckerberg built a website called “The Facebook” so that he and his fellow students could learn about each other online. He had no idea that in just 10 years, his company would be one of the most powerful and profitable in human history, and that he would be one of the richest people on earth. Facebook has been very good for its inventor, but has it been very good for the rest of us? It lets us post videos of our pets, share stories of our vacations, and find out what people we barely know had for lunch—and who could possibly argue with the goodness of that? But has it also become a platform in which people’s worst impulses are brought together, amplified, and then unleashed upon the world?

Do Facebook posts lead people to do terrible things in the real world, such as commit arson, or assault, or even murder?

Karsten Müller and Carlo Schwarz (2018) worried that the answer might be yes. So they conducted a study in Germany, where a right-wing anti-immigrant party (the Alternative für Deutschland, or AfD) had recently developed a significant presence on Facebook. The study involved nothing more than measuring the correlation between two variables: the number of anti-refugee Facebook posts that appeared on the AfD’s website each week and the number of violent crimes against refugees that occurred each week. The accompanying figure shows what they found. As you can see, the two variables were indeed positively correlated: The more “hate posts” that appeared on AfD’s Facebook page at any moment in time, the more “hate crimes” happened on German streets.

But wait a minute. What about the third-variable problem? Hate posts and hate crimes may well be correlated, but that doesn’t mean the former caused the latter. Maybe the hate crimes caused the hate posts, or maybe some third variable—such as a change in the weather or the parliament or the stock market—caused them both. Drawing conclusions about the causal relationship between two variables requires that one of those variables be manipulated, right?

Yes, it does, but luckily for the researchers, nature did just that. Internet service in Germany is not as reliable as it is in the United States, and German Facebook users experience frequent interruptions that can knock everyone offline for hours or even days at a time. So what happens when Internet access is randomly turned on and off and then on again? When the researchers analyzed the historical data on service interruptions, they discovered that when Facebook went down, hate crimes went down too, and when Facebook came back up again, so did hate crimes. Because service interruptions are basically random events, they serve as a manipulation of one variable (the number of hate posts) whose impact on the other variable (the number of hate crimes) can then be measured. When the researchers did that, the impact was clear. As the researchers themselves concluded, “Social media has not only become a fertile soil for the spread of hateful ideas but also motivates real-life action” (Müller & Schwarz, 2018, p. 40).

Psychologists are often faced with a tough choice: They can either (a) observe a correlation in the real world but be unsure about causation, or (b) firmly establish causation in the laboratory but be unsure about whether it generalizes to the real world. But every once in a while, nature resolves this dilemma by randomly manipulating a variable in the real world. When that happens, clever researchers may take advantage of the situation and use it to answer pressing social questions—and to produce some very hot science.

Second—and much more important—we could expect the two conditions to have roughly equal numbers of mean children and sweet children, of older children and younger children, of children who have a lot of adult supervision at home and children who have little, and so on. In other words, we could expect the two conditions to have roughly equal numbers of children who are anything-you-can-ever-name-and-everything-you-can’t! Because the children in the two conditions will be the same on average!
in terms of meanness, age, adult supervision, and every other variable in the known universe except the variable we manipulated, we could be sure that the variable we manipulated was the cause of the differences in the variable we measured. Because exposure to media violence would be the only difference between the two conditions when we started the experiment, it must be the cause of any differences in aggressiveness we observed at the end of the experiment.

**Statistical Testing: Making Sure Conditions Don’t Differ by Chance**

Random assignment is a powerful tool. Unfortunately, like a lot of tools, it doesn’t work every time you use it. If we randomly assigned children to two conditions, we could expect the two conditions to have roughly equal numbers of mean children, young children, supervised children, and so on. But the key word in that sentence is *roughly*. When you flip a coin 100 times, you expect it to come up heads roughly 50 of those times. But every once in a very, very long while, it will come up heads 60 times, or 70 times, or even 100 times, by sheer chance alone (see *The Real World: The Surprisingly High Likelihood of Unlikely Coincidences*). This will not happen often, of course, but if you execute 100 coin flips enough times, it will eventually happen.²

Because random assignment is achieved by using a randomizing device such as a coin, every once in a long while the coin will assign more unsupervised kids to play violent video games and more supervised kids to play nonviolent video games by sheer chance alone. When this happens, scientists say that “random assignment has failed” (though they should actually say that the coin has failed to produce random assignment). When random assignment fails, we are right back where we started when we politely allowed the children to self-select their conditions, and as you will recall, that is not a very good place to be right back to.

How can we know whether random assignment has failed? Unfortunately, we can never know for sure. But what we can do is calculate the odds that random assignment has failed each time we conduct an experiment. It’s not important for you to know how to do this calculation, but it is important for you to understand how psychologists interpret its results. Psychologists perform this calculation every time they do an experiment, and they generally do not accept the results of their experiments unless the calculation suggests that there is less than a 5% chance that those results would have occurred if random assignment had failed. Such results are said to be *statistically significant*, and psychologists typically indicate this by writing “\( p < .05 \),” which simply means that the probability \( p \) that the result would have been observed if random assignment had failed is less than 5% (\(< .05\)). Psychologists have several other ways to calculate the odds that random assignment has failed, but this calculation is the most common.

**Drawing Conclusions**

If we applied all the techniques discussed so far, we could design an experiment that had a very good chance of establishing the causal relationship between two variables. That experiment would be said to have **internal validity**, which is an

---

2 How long is “eventually”? Odds are that you would have to execute the 100 coin flips 12,680,000,000,000,000,000,000,000,000,000,000 times in order for the coin to come up heads every time by chance alone. Assuming you could flip one coin per second, this would take you about 4³⁵ years, which is considerably longer than the universe has existed. In other words, don’t try this at home.

---

<table>
<thead>
<tr>
<th>Violent exposure condition</th>
<th>Exposed to media violence?</th>
<th>Adult supervision?</th>
<th>Aggression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Nonviolent exposure condition</td>
<td>Exposed to media violence?</td>
<td>Adult supervision?</td>
<td>Aggression</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>Low</td>
<td></td>
</tr>
</tbody>
</table>

---

**Internal Validity** An attribute of an experiment that allows it to establish causal relationships.
The Real World • The Surprisingly High Likelihood of Unlikely Coincidences

A recent survey found that more than half of college graduates believe in extrasensory perception, or ESP, and it is easy to understand why. Just consider the case of the Truly Amazing Coincidence. One night you dream that a panda is piloting an airplane over the Indian Ocean, and the next day you tell a friend, who says, “Wow, I had exactly the same dream!” One morning you wake up humming an old Katy Perry tune (probably “Part of Me”), and an hour later you hear it playing in the mall. You and your roommate are sitting around watching television when suddenly you turn to each other and say in perfect unison, “Wanna pizza?” Coincidences like these might make anyone believe in strange and spooky supernatural weirdness.

Well, not anyone. While the Nobel laureate Luis Alvarez was reading the newspaper one day, a particular story got him thinking about an old college friend whom he hadn’t seen in years. A few minutes later, he turned the page and was shocked to see the very same friend’s obituary. But before concluding that he had suddenly developed an acute case of ESP, Alvarez decided to use probability theory to determine just how amazing this coincidence really was. He estimated the number of friends an average person has, estimated how often an average person thinks about each of those friends, did a few simple calculations, and determined the probability that someone would think about a friend 5 minutes before learning about that friend’s death. The odds were astonishing. In a country the size of the United States, Alvarez calculated that this coincidence should happen to 10 people every day (Alvarez, 1965). A fellow Nobel laureate put the number closer to 80 people a day (Charpak & Broch, 2004!)

“In ten years there are five million minutes,” says the statistics professor Irving Jack. “That means each person has plenty of opportunity to have some remarkable coincidences in his life” (Neimark, 2004). If 250 million Americans dream for about 2 hours every night, that’s a half billion hours of dreaming, so it isn’t really surprising that two people might have the same dream, or that someone would dream about something that actually happened the next day. As the mathematics professor John Paulos noted, “In reality, the most astonishingly incredible coincidence imaginable would be the complete absence of all coincidence” (Neimark, 2004).

If all of this seems surprising to you, then you are not alone. Research shows that people routinely underestimate the likelihood of coincidences happening by chance (Diaconis & Mosteller, 1989; Falk & McGregor, 1983; Hintzman, Asher, & Stern, 1978). If you want to profit from this fact, just assemble a group of 24 or more people and offer to bet anyone $1 that at least two of the people in the group share a birthday. The odds are in your favor, and the bigger the group, the better those odds are. In fact, in a group of 35 people, the odds are a remarkable 85%. When everyone starts handing you their dollars and asking how in the world you knew that two of the people in your group shared a birthday, you can honestly tell them it was ESP—also known as Especially Sneaky Psychology.
variable, and in the people we studied, and it is likely that. Each phrase describes an important restriction on the conclusions we can draw from this or any other experimental result. Let’s consider them in turn, as well as the kinds of restrictions they place on our conclusions.

**The Representativeness Restriction: “As We Defined That Variable . . .”**

The results of an experiment naturally depend on how the independent and dependent variables are operationally defined. For instance, we are more likely to find that exposure to media violence causes aggressiveness if we operationally define exposure as “watching 2 hours of gory axe murders” rather than “watching 10 minutes of football,” or if we define aggressiveness as “interrupting another person who is talking” rather than “beating someone with a tire iron.” The way we operationally define variables can have a profound influence on what we find, so which of these ways of operationally defining them is the right way?

One common answer is that we should operationally define variables in an experiment as they are defined in the real world. **External validity** is an attribute of an experiment in which variables have been operationally defined in a normal, typical, or realistic way. It seems pretty clear that the kind of aggressive behavior that concerns teachers and parents lies somewhere between an interruption and an assault, and that the kind of media violence to which children are typically exposed lies somewhere between sports and torture. If the goal of an experiment is to determine whether the kinds of media violence to which children are typically exposed causes the kinds of aggressiveness with which societies are typically concerned, then external validity is important. When variables are defined in an experiment as they typically are in the real world, we say that the variables are representative of the real world.

External validity sounds like such a good idea that you may be surprised to learn that most psychology experiments are externally invalid—and that most psychologists don’t consider this a problem (Mook, 1983). To understand why, consider a simple example from physics. Physicists have a theory: Heat is the result of the rapid motion of molecules. This theory gives rise to a hypothesis: When the molecules that constitute an object are slowed, the object should become cooler. Imagine that a physicist tested this hypothesis by performing a laboratory experiment in which she used a laser to slow the motion of the molecules in a steel ball, measured the temperature of the ball before and after, and found that the temperature of the ball decreased by a large amount when its molecules were slowed. Are you worried about external invalidity? Are you thinking, “How can this experiment teach us anything about the real world when in the real world no one uses lasers to slow the movement of the molecules in steel balls?”

Probably not, and that’s because you intuitively understand that the physicist’s theory led to a clear hypothesis about what would happen in the laboratory, and that the events the physicist manipulated and measured in the laboratory therefore tested the hypothesis quite nicely. Similarly, a well-thought-out theory about the causal relationship between exposure to media violence and aggressiveness should lead to hypotheses about how children in a laboratory will behave after playing a violent video game, and thus their behavior tests that hypothesis quite nicely. If children who play Left 4 Dead in a laboratory are more likely to shove each other than are children who play Minecraft, then any theory claiming that media violence cannot influence aggressiveness has just been proved wrong.

In short, theories allow us to generate hypotheses about what can, must, or will happen under particular circumstances; and experiments are usually meant to create some of those circumstances, test the hypotheses, and thereby provide evidence for or against the theories that generated them (see also Hot Science: Hate Posts and Hate Crimes: Not Just a Correlation). Experiments are not usually meant to be miniature versions of everyday life, and as such, external invalidity is not usually a problem.

---

**external validity** An attribute of an experiment in which variables have been defined in a normal, typical, or realistic way.
The Generalizability Restriction: “In the People We Studied . . .”

You already know that psychologists rarely measure an entire population, but instead measure a sample from that population. How big should that sample be? The size of a population is signified by the uppercase letter \( N \), the size of a sample is signified by the lowercase letter \( n \), and therefore \( 0 < n < N \). However, in most studies, \( n \) is a whole lot closer to 0 than it is to \( N \), and indeed, in some cases \( n = 1 \).

For example, some individuals are so remarkable that they deserve close study, and the psychologists who study them are using the case method, which is a procedure for gathering scientific information by studying a single individual. We can learn a lot about, say, memory by studying someone such as Rajveer Meena, who memorized the first 70,000 digits of \( \pi \); about consciousness by studying someone such as Henry Molaison, whose ability to look backward and forward in time was impaired by brain surgery; about intelligence and creativity by studying someone such as Tanishq Abraham, who graduated from college at the age of 11 and began consulting in the aerospace industry. Such people are not only interesting in their own right, but their remarkable abilities can also provide important insights into how the rest of us work.

With that said, the vast majority of studies you will read about in this textbook use samples of 10, 100, 1,000, or even many thousands of people. Psychologists typically use the largest samples they can get because the larger the sample, the more confidence they can have in their findings. Why? Well, if a coin came up heads 7 out of 10 times, you might suspect it was an unfair coin—that is, that the coin was slightly heavier on the tail than the head. But you would not be highly confident in that conclusion because it was based on a small sample of coin flips. If the same coin came up heads 700 out of 1,000 times, however, you would be extremely confident that it was an unfair coin. Because large samples allow us to be more confident in our results, in an ideal world, \( n \) would be very close to \( N \) itself. Alas, psychologists in the real world have limited time, limited research funds, and limited access to participants, so they have to settle for the largest \( n \) that circumstances allow.

So how do psychologists determine which people they will include in their sample and which people they won’t? One method is by random sampling, which is a technique for selecting participants that ensures that every member of a population has an equal chance of being included in the sample. (Note to self: Do not confuse random sampling with random assignment. The only thing they have in common is the word random.) When we randomly sample participants from a population, the sample is said to be representative of the population. This allows us to generalize from the sample to the population—that is, to conclude that what we observed in our sample would also have been observed if we had measured the entire population.

You probably already have solid intuitions about the value of random sampling. For example, if you stopped at a roadside farm stand to buy a bag of cherries and the farmer offered to let you taste a few that she had carefully selected from the bag, you’d be reluctant to generalize from that sample to the population of cherries in the bag because you’d know that the cherries you were offered had been . . . well, cherry-picked. But on the other hand, if the farmer invited you to close your eyes and pull a few cherries from the bag at random, you’d probably consider those cherries to be a representative sample and you would naturally generalize from that sample (“The cherries I tasted were quite sweet”) to the population (“I bet most of the cherries in the bag are sweet too”).

Like external validity, random sampling sounds like an obviously good idea, so you may be surprised once again to find that most psychological studies involve nonrandom samples. Indeed, virtually every participant in every psychology experiment you will ever read about was a volunteer, and a large share were college students who were significantly younger, smarter, healthier, wealthier, and Whiter than the average Earthling. About 96% of the people whom psychologists study come from countries that have just 12% of the world’s population, and 70% come from the United States.
alone (Henrich, Heine, & Norenzayan, 2010). Why do psychologists sample people nonrandomly?

Because they have to. Even if there were a list of all the world’s human inhabitants from which we could randomly sample our research participants (there isn’t), how would we locate the 5-year-old Bedouin girl whose family roams the desert so that we could measure the electrical activity in her brain while she watched cartoons? How would we convince an illiterate farmer in South Sudan to complete a lengthy questionnaire about his political beliefs? How could we sample the Ayoreo-Totobiegosode tribe of Paraguay, which has never had any contact with the outside world? Most psychology experiments are conducted by professors and graduate students at colleges and universities in the Western Hemisphere, and as much as they might like to randomly sample the population of our planet, the fact is that they are pretty much stuck studying the local folks who volunteer for their studies.

Is this a fatal flaw in psychological science? No, and there are two reasons why. First, *sometimes the representativeness of a sample doesn’t matter.* If one pig flew over the Statue of Liberty just one time, it would instantly disprove the Standard Theory of Porcine Locomotion, and it really wouldn’t matter if other pigs could do the same trick. Similarly, if playing *Left 4 Dead* for an hour caused a group of 5th graders from a public school in Ann Arbor, Michigan, to start shoving other kids in the laboratory, then even if the game did not have a similar effect on 9th graders from a private school in Austin, Texas, we would still know that media violence *can* influence aggressiveness—which means that any theory that says it can’t is just plain wrong. Sometimes psychologists aren’t concerned with whether everyone does something; they just want to know if anyone does it.

The second reason why nonrandom sampling is not a fatal flaw is that *sometimes the representativeness of the sample is a reasonable starting assumption.* Instead of asking, “Do I have a compelling reason to believe that my sample is representative of the population?” we could just as well ask, “Do I have a compelling reason to believe that my sample is not representative of the population?” For example, few of us would be willing to take a new medicine if a nonrandom sample of participants took it and died. Indeed, we would probably refuse to take the medicine even if those participants were mice! Although these nonrandomly sampled participants would be different from us in many ways (for example, tails and whiskers), most of us would assume that anything that kills them has some reasonable chance of killing us. Similarly, if a psychology experiment demonstrated that a sample of American children behaved aggressively after playing violent video games, we might ask whether there

It can be a mistake to generalize from a nonrandom sample. In 1948, pollsters mistakenly predicted that Thomas Dewey would beat Harry Truman. Why? Because polling was done by telephone, and Dewey Republicans were more likely to have telephones than were Truman Democrats. In 2004, pollsters mistakenly predicted that John Kerry would beat George Bush. Why? Because pollsters solicited voters as they left the polls, and Kerry supporters were more optimistic and therefore more willing to stop and talk to pollsters. In 2016, pollsters mistakenly predicted that Hillary Clinton would beat Donald Trump. Why? Because pollsters solicited too many of the highly educated voters who were likely to support Clinton and not enough of the less educated voters who were likely to support Trump.
is a compelling reason to suspect that Ecuadorian college students or middle-aged Australians would respond any differently. If the answer is yes, then we can conduct experiments to test that possibility.

The bottom line is that learning about some people does not necessarily tell us about all people, but it can still tell us a lot—and it certainly tells us more than learning about no people at all, which is often the only other choice.

The Reliability Restriction: “It Is Likely That . . .”

A replication is an experiment that uses the same procedures as a previous experiment but with a new sample from the same population. In the last few years, many major media outlets have reported that when psychologists replicate the experiments of other psychologists, they usually fail to replicate their results, which suggests that the initial result was some sort of fluke. Could that be right? Are the headlines true?

Estimating Psychology’s Replication RateIf a researcher finds that children who were randomly assigned to play violent video games are more aggressive than those who were randomly assigned to play nonviolent video games, we would expect that other researchers who use the same video games and the same measures of aggressiveness under the same conditions should get the same result. We might not expect this to happen every single time, but we would expect it to happen most of the time—and if it happened a whole lot less than most of the time, we would have good reason to suspect that the original finding was a fluke. You already know that psychologists publish their results only when the odds that random assignment failed are less than 5%, but that means that 5% of the time, the original result was indeed a fluke and therefore we would not expect it to be replicated. We can live with that. But what if these flukes happen way more than 5% of the time? What if they happen the majority of the time, just as the headlines suggest?

In the last few years, teams of psychologists have tried to estimate the proportion of flukes in the scientific literature by selecting a sample of published studies and then attempting to replicate them. Some teams have found the replication rate in their samples to be frighteningly low (Open Science Collaboration et al., 2015) while others have found it to be reasonably high (Klein et al., 2014). But do any of these findings tell us the actual replication rate of studies in psychology? Probably not (Gilbert et al., 2016). First, the teams who did these replications did not randomly sample the population of published studies in psychology. Rather, they chose studies of particular kinds (e.g., those that are easy for anyone to do rather than those that require a lot of time, money, or expertise) from particular areas of psychology (e.g., almost always from social, and almost never from neuroscience, developmental, or clinical). Because the studies they chose are not representative of psychology as a whole, their results may tell us something about the replicability of the specific studies in the samples, but they are unlikely to tell us much about the replication rate of the entire discipline.

Second, the teams did not always use the same methods that were used in the original studies—sometimes because details of the original methods were not known, and sometimes because the teams made mistakes. This means that some of their studies were not really replications at all; and yet, the results of these studies are typically included when journalists report that there is a “replication crisis” in psychology. Indeed, when the National Academies of Sciences, Engineering, and Medicine recently considered all the evidence, they concluded that contrary to what you may read in the news, “it is not helpful, or justified, to refer to psychology as in a state of ‘crisis’” (National Academies of Sciences, Engineering, and Medicine, 2019, p. 124). The bottom line is that, no one knows the actual replication rate of experiments in psychology.

Type I And Type II Errors If we can’t answer the question about the real replication rate, then maybe we can answer an easier one: What is the ideal replication rate? What percentage of the results published in scientific journals do we wish would replicate
perfectly? You are probably tempted to say 100%, but that answer is wrong. To understand why that answer is wrong requires understanding the two errors researchers can make when they draw conclusions from evidence.

A **Type I error** occurs when researchers conclude that there is a causal relationship between two variables when in fact there is not. If a psychologist concluded that playing violent video games had increased aggressiveness in a sample of children when in fact it hadn't, that conclusion would be a Type I error, also known as a false positive. A **Type II error** occurs when researchers conclude that there is not a causal relationship between two variables when in fact there is. If a psychologist concluded that playing violent video games had not increased aggressiveness in a sample of children when in fact it had, that conclusion would be a Type II error, also known as a false negative. You can think of these mistaken conclusions as flukes and flunks: A fluke happens when we detect something that isn't really there, and a flunk happens when we fail to detect something that is.

Now, the annoying thing about flukes and flunks is that anything you do to reduce the likelihood of one will increase the likelihood of the other. Why? Well, just consider a home security system that sounds an alarm whenever the motion detector is triggered. If you set the sensitivity of the motion detector to high, the alarm will always ring when a burglar opens a window, but sometimes it will ring when a houseplant drops a leaf. You will always detect a burglar when he is really there, but sometimes you’ll also detect him when he’s not. On the other hand, if you set the detector’s sensitivity to low, the alarm will never mistakenly ring for a houseplant, but sometimes it will fail to ring when a burglar opens a window. You will never detect a burglar when he isn’t really there, but sometimes you’ll fail to detect him when he is. See the problem? If you want to catch every single burglar, then you’ll just have to put up with a few false alarms, and if you want to eliminate every single false alarm then you’ll just have to put up with a few uncaught burglars. When an alarm rings even though there is no burglar, the alarm is making a Type I error; and when an alarm fails to ring when there is a burglar, the alarm is making a Type II error. So when setting the sensitivity of your motion detector, you just have to decide which of these errors worries you the most.

Psychologists worry too—not about burglars and houseplants, but about flukes and flunks. They don’t want to draw false conclusions from their results, but they also don’t want to miss the chance to draw true ones. If psychologists designed their experiments to avoid all Type II errors, they would rarely miss discovering a real and important fact about human behavior, but a sizeable chunk of those “facts” would turn out to be flukes and the replication rate would be very low. On the other hand, if psychologists designed their experiments to avoid all Type I errors, they would never be fooled by flukes and the replication rate would be very high, but a lot of their studies would be flunks that failed to discover real and important facts about human behavior.

Neither of these are happy outcomes, so psychologists do exactly what you would do if you were setting up a home security system: They try to strike a balance between the two mistakes they are in danger of making by minimizing the risk of whichever mistake seems worse in a particular situation, and accepting an elevated risk of its opposite (Finkel, Eastwick, & Reis, 2017). The National Academies of Sciences, Engineering, and Medicine noted that “even if there were a definitive estimate of replicability in psychology, no one knows the expected level of non-replicability in a healthy science” (2019, p. 124). In other words, we do not know what the replication rate should be, but we do know that if it were 100%, psychologists would be failing to make the most important discoveries.

Replication serves an important function in psychology, as it does in every science. And it reminds us that even the best experimental evidence does not allow us to conclude that two variables are causally related; rather, it allows us to conclude that two variables are likely to be causally related. The more easily and more often that evidence is reproduced, the more confident we can be in the causal relationship between the variables. We can never be completely confident in any result, of course, but replications can move us ever closer to certainty.

---

**Type I error** An error that occurs when researchers conclude that there is a causal relationship between two variables when in fact there is not.

**Type II error** An error that occurs when researchers conclude that there is not a causal relationship between two variables when in fact there is.
Build to the Outcomes

1. What is a natural correlation and how can it be measured?
2. What does it mean for a correlation to be positive or negative, strong or weak?
3. What is the correlation coefficient and what do its values mean?
4. What is the third-variable problem?
5. How do manipulation and random assignment solve the third-variable problem?
6. Why is self-selection a problem?
7. What is the difference between a dependent variable and an independent variable?
8. What is the difference between internal validity and external validity?
9. What can be learned from nonrandom samples?
10. What are Type I and Type II errors?

Thinking Critically About Evidence

In 1620, Sir Francis Bacon published a book called *Novum Organum*, in which he described a new method for discovering facts about the natural world. What was once called the “Baconian method” is now widely known as the scientific method, and that method has allowed human beings to learn more about themselves and the world in the last four centuries than in all the previous centuries combined.

As you’ve seen in this chapter, the scientific method produces empirical evidence. But empirical evidence is useful only if we know how to think about it—and the fact is that most of us don’t. Using evidence requires critical thinking, which involves asking ourselves tough questions about whether we have interpreted the evidence in an unbiased way, and about whether the evidence tells not just the truth, but the whole truth. Research suggests that most of us have trouble doing these two things and that educational programs designed to teach or improve our critical thinking skills are not particularly effective (Willingham, 2007). Why do we have so much trouble thinking critically?

Consider the armadillo. Some animals freeze when threatened; others duck, run, or growl. Armadillos jump. This natural tendency served armadillos quite well for millennia because for millennia the most common threat to an armadillo’s well-being was a rattlesnake. Alas, this natural tendency serves armadillos rather poorly today, because when they wander onto a Texas highway and are threatened by a speeding car, they jump up and hit the bumper. No armadillo makes this mistake twice. Human beings also have natural tendencies that once served them well but no longer do. Our natural and intuitive ways of thinking about evidence, for example, worked just fine when we were hunter-gatherers living in small groups on the African savannah. But today, most of us live in large, complex societies, and our natural ways of thinking about the world don’t serve us so well. Sir Francis Bacon understood this. In the very same book in which he developed the scientific method, he argued that two natural human tendencies—the tendency to see what we expect or want to see, and the tendency to ignore what we can’t see—are the enemies of critical thinking. Let’s examine each of these tendencies and see how they thwart critical thinking.

We See What We Expect and Want to See

When two people are presented with the same evidence, they often draw different conclusions, and Sir Francis Bacon knew why: “The human understanding, once it has adopted opinions . . . draws everything else to support and agree with them . . . [and therefore our] first conclusion colors and brings into conformity with itself all
Thinking Critically About Evidence

that come after.” In other words, our preexisting beliefs color our view of new evidence, causing us to see what we expect to see. As such, evidence often seems to confirm what we believed all along.

This tendency has been widely documented in psychological science. For instance, participants in one study (Darley & Gross, 1983) learned about a little girl named Hannah. One group of participants was told that Hannah came from an affluent family, while the other group was told that Hannah came from a poor family. All participants were then shown some evidence about Hannah’s academic abilities (specifically, they watched a video of Hannah taking a reading test) and were then asked to rate Hannah. Although the video was exactly the same for all participants, those who believed that Hannah’s family was affluent rated her performance more positively than did those who believed that her family was poor. What’s more, both groups of participants defended their conclusions by citing evidence from the video! Experiments like this one suggest that when we consider evidence, what we see depends on what we expected to see.

Our beliefs aren’t the only things that color our views of evidence. They are also colored by our preferences and prejudices, our ambitions and aversions, our hopes and needs and wants and dreams. As Bacon noted, “The human understanding is not a dry light, but is infused by desire and emotion which give rise to wishful science. For man prefers to believe what he wants to be true.” Research suggests that Bacon was right about this, too. For example, participants in one study (Lord, Ross, & Lepper, 1979) were shown some scientific evidence about the effectiveness of the death penalty. Some of the evidence suggested that the death penalty deterred crime, whereas some suggested it did not. What did participants make of this mixed bag of evidence? Participants who originally supported the death penalty became even more supportive, and participants who originally opposed the death penalty became even more opposed. In other words, when presented with exactly the same evidence, participants saw what they wanted to see and ended up feeling even more sure about their initial views. Subsequent research has shown that the same pattern emerges when professional scientists are asked to rate the quality of scientific studies that either confirm or disconfirm what they want to believe (Koehler, 1993).

Exactly how do beliefs and desires shape our view of the evidence? One way is that we tend to hold different kinds of evidence to different standards (McPhetres & Zuckerman, 2017). When evidence confirms what we believe or want to believe, we ask ourselves, “Can I believe this evidence?” But when evidence disconfirms what we believe or want to believe, we ask ourselves, “Must I believe this evidence?” (Gilovich, 1991). The problem is that can is a low standard and must is a high one: You can believe almost anything, but you must believe almost nothing. For instance, can you believe that people with college degrees are happier than people without them? Of course. Plenty of surveys show that just such a relationship exists, and a reasonable person who studied the evidence could easily defend this conclusion. Now, must you believe it? Well, no. After all, those surveys didn’t measure every single person on earth, did they? And if the survey questions had been asked differently, they might well have produced different answers, right? A reasonable person who studied the evidence could easily conclude that the relationship between education and happiness is not yet clear enough to warrant such a strong conclusion.

A second way in which our beliefs and desires shape our view of the evidence is by influencing which evidence we consider in the first place. Most people surround themselves with others who believe the same things they believe and want the same things they want, which means that our friends and families are much more likely to validate our beliefs and desires than to challenge them. Studies also show that when given the opportunity to search for evidence, people preferentially search for evidence that confirms their beliefs and fulfills their desires (Hart et al., 2009). What’s more, when people find evidence that confirms their beliefs and fulfills their desires, they tend to stop looking; yet when they find evidence that does the opposite, they keep searching for more evidence (Kunda, 1990).
What all of these studies have in common is that they suggest that most evidence leaves room for interpretation, and that's the room in which our beliefs and desires love to hang out. Because it is so easy to see what we expect to see, or to see what we want to see, the first rule of critical thinking is this: **Doubt your own conclusions.** One of the best ways to find the truth about the world is to seek out people who don’t see the world your way and then listen carefully to what they have to say. Most of us find this painful—those of us who watch MSNBC can’t bear to turn on FOX, and vice versa—but the outcome is often worth the pain. Scientists go out of their way to expose themselves to criticism by sending their papers to the colleagues who are most likely to disagree with them or by presenting their findings to audiences full of critics, and they do this largely so that they can achieve a more balanced view of their own conclusions (as well as their self-worth). If you want to be happy, take your friend to lunch; but if you want to be right, take your enemy.

**We Don’t Consider What We Don’t See**

In another part of his remarkable book, Francis Bacon recounted an old story about a man who visited a Roman temple. The priest showed the man a portrait of several sailors who had taken religious vows and then miraculously survived a shipwreck, and suggested that this was clear evidence of the power of the gods. The visitor paused a moment and then asked precisely the right question: “But where are the pictures of those who perished after taking their vows?” According to Bacon, most of us never think to ask this kind of question. We consider the evidence we can see and forget about the evidence we can’t. Bacon claimed that “little or no attention is paid to things invisible” and argued that this natural tendency was “the greatest impediment and aberration of the human understanding.”

Bacon was right when he claimed that people rarely pay attention to what is missing (Kardes & Sanbonmatsu, 2003). For example, participants in one study (Newman, Wolff, & Hearst, 1980) played a game in which they were shown a set of trigrams, which are three-letter combinations such as SXY, GTR, BCG, and EVX. On each trial, the experimenter pointed to one of the trigrams in the set and told the participants that this trigram was the special one. The participants’ job was to figure out what made the special trigram so special. How many trials did it take before participants figured it out? It depended on the trigram’s special feature. For half the participants, the special trigram was always the one that contained the letter T, and participants in this condition needed to see about 34 sets of trigrams before they figured out that the presence of T was what made the trigram special. But for the other half of the participants, the special trigram was always the one that lacked the letter T. How many trials did it take before participants figured it out? They never figured it out—never. It is much easier to think about what is there than what isn’t.

The tendency to ignore missing evidence can lead to erroneous conclusions (Wainer & Zwerling, 2006). For instance, consider the red map in **FIGURE 2.12**, which shows the U.S. counties with the lowest rates of kidney cancer. As you can see, they are predominantly rural counties in the South, West, and Midwest. It isn’t hard to imagine why the places with the lowest populations might also have the lowest rates of kidney cancer: People who live in these counties probably eat more farm-grown foods, breathe less polluted air, drink less polluted water, engage in more outdoor activities, and so on. Given the obvious health benefits of “country living,” it is no wonder that the most rural counties in America have the lowest kidney cancer rates.

That’s a reasonable hypothesis based on the evidence you saw. But it is utterly wrong, and you would have known it was utterly wrong if only you had stopped to think about the evidence you were not shown. That evidence is shown in the green map in Figure 2.12, which shows the U.S. counties with the highest rates of kidney cancer. As you can see, they too are predominantly rural and predominantly in the South, West, and Midwest. Indeed, except for their colors, the two maps in Figures 2.12 look pretty much the same.
Why? Because as it turns out, rural counties tend to have extreme rates of kidney cancer—that is, they have some of the lowest rates, but they also have some of the highest rates—and that’s because rural counties have fewer people in them. Imagine flipping a coin either 3 times or 3,000 times. Which of those series of flips is most likely to produce the extreme outcomes “all heads” or “all tails”? The series with fewer flips, of course. For precisely the same reason, counties with few people are more likely to produce the extreme outcomes “high cancer rate” and “low cancer rate” than are counties with many people.

Someone who is shown the evidence in the red map and who forgets to ask about the missing evidence in the green map would draw the wrong conclusion about the relationship between kidney cancer and population density. And yet, forgetting to ask for missing evidence is something that happens all the time. When the Gates Foundation decided to spend $1.7 billion to improve schools, they began by identifying the characteristics of America’s best-performing schools. They discovered that the best schools had small class sizes. So they spent a lot of money to create schools with small class sizes and, much to their chagrin, discovered that the small classrooms weren’t any better on average than were their larger counterparts. What went wrong? The Gates Foundation began by considering the characteristics of the best-performing schools, but they forgot to also consider the characteristics of the worst-performing schools. If they had looked at this missing evidence, they would have seen that schools with small classrooms are indeed over-represented among the best schools, but they are also overrepresented among the worst. Just as small counties produce extreme outcomes, so do small classrooms (Wainer & Zwerling, 2006). If the first rule of critical thinking is to doubt what you do see, then the second rule is to consider what you don’t see.

The Skeptical Stance

Science is a human enterprise and humans make mistakes: They see what they expect to see, they see what they want to see, and they often fail to consider what they can’t see at all. What makes science different from most other human enterprises is that scientists actively seek to discover and remedy their mistakes. Scientists are constantly striving to make their observations more accurate and their reasoning more rigorous, and they invite anyone and everyone to examine their evidence and challenge their conclusions. As a result, science is the ultimate democracy in which the lowliest nobody can triumph over the most celebrated somebody. When an unknown Swiss patent clerk with a vivid imagination challenged the greatest physicists of his day, he didn’t have a famous father, a fancy degree, powerful friends, or a fat wallet. Albert Einstein won the scientific debate for one reason and one reason alone: He was right.

So think of the remaining chapters in this textbook as a report from the field—a description of the work that psychological scientists have done so far as they stumble toward knowledge. These chapters tell the story of the men and women who have put their faith in Francis Bacon’s method and used it to pry loose small pieces
of the truth about who we are, how we think, feel, and behave, and what we are all doing here together on the third rock from the sun. Some of their reports will turn out to be flukes, others will turn out to be flunks, but every one of them is somebody’s best guess about the way people work. Read these reports with interest but think critically about their claims—and for that matter, about everything else.

**Build to the Outcomes**

1. According to Francis Bacon, what are the two enemies of critical thinking?
2. What does it mean to say, “If you want to be happy, take your friend to lunch; if you want to be right, take your enemy?”
3. What makes science different from most other human enterprises?

---

**The Ethics of Science: Doing What’s Right**

Somewhere along the way, someone probably told you that it isn’t nice to treat people like objects. And yet, psychologists may appear to be doing just that when they create experimental situations that cause people to feel fearful or sad, to do things that are embarrassing or immoral, and to learn things about themselves and others that they might not really want to know. Don’t be fooled by appearances. The fact is that psychologists go to great lengths to protect the well-being of their research participants, and they are bound by a code of ethics that is as detailed and demanding as the professional codes that bind physicians, lawyers, and accountants. That code requires psychologists to show respect for people, for animals, and for the truth. Let’s examine each of these obligations in turn.

### Respecting People

During World War II, Nazi doctors performed barbaric experiments on human subjects, such as removing organs without anesthesia or submerging people in ice water just to see how long it would take them to die. After the war ended, the international community developed the Nuremberg Code of 1947 and then the Declaration of Helsinki in 1964, which spelled out rules for the ethical treatment of the people who participate in experiments. Unfortunately, not everyone obeyed the rules. For example, from 1932 until 1972, the U.S. Public Health Service conducted the infamous Tuskegee experiment, in which 399 African American men with syphilis were denied treatment so that researchers could observe the progression of the disease. As one journalist noted, the government “used human beings as laboratory animals in a long and inefficient study of how long it takes syphilis to kill someone” (Jones, 1993, p. 10).

So in 1974, the U.S. Congress created the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. In 1979, the U.S. Department of Health, Education and Welfare released what came to be known as the Belmont Report, which described three basic principles that all research involving human participants must follow. First, research should show respect for persons and their right to make decisions for and about themselves without undue influence or coercion. Second, research should be beneficent, which means that it should attempt to maximize benefits and reduce risks to the participant. Third, research...
should be *just*, which means that it should distribute benefits and risks equally to participants without prejudice toward particular individuals or groups.

The specific ethical code that psychologists follow incorporates these basic principles and expands them. (You can find the American Psychological Association’s *Ethical Principles of Psychologists and Code of Conduct* [2017] at http://www.apa.org/ethics/code.) Here are a few of the most important rules that govern the conduct of psychological research:

- **Informed consent**: Participants may not take part in a psychological study unless they have given *informed consent*, which is a *verbal agreement to participate in a study made by an adult who has been informed of all the risks that participation may entail*. This doesn’t mean that the person must know everything about the study (e.g., the hypothesis), but it does mean that the person must know about anything that might potentially be harmful or painful. If people cannot give informed consent (e.g., because they are minors or are mentally incapable), then informed consent must be obtained from their legal guardians. And even after people give informed consent, they always have the right to withdraw from the study at any time without penalty.

- **Freedom from coercion**: Psychologists may not coerce participation. Coercion not only means physical and psychological coercion but monetary coercion as well. It is unethical to offer people large amounts of money to do something that they might otherwise decline to do. College students may be invited to participate in studies as part of their training in psychology, but they are ordinarily offered the option of learning the same things by other means.

- **Protection from harm**: Psychologists must take every possible precaution to protect their research participants from physical or psychological harm. If there are two equally effective ways to study something, the psychologist must use the safer method. If no safe method is available, the psychologist may not perform the study.

- **Risk–benefit analysis**: Although participants may be asked to accept small risks, such as a minor shock or a small embarrassment, they may not even be asked to accept large risks, such as severe pain, psychological trauma, or any risk that is greater than the risks they would ordinarily take in their everyday lives. Furthermore, even when participants are asked to take small risks, the psychologist must first demonstrate that these risks are outweighed by the social benefits of the new knowledge that might be gained from the study.

- **Deception**: Psychologists may use deception only when it is justified by the study’s scientific, educational, or applied value and when alternative procedures are not feasible. They may never deceive participants about any aspect of a study that could cause them physical or psychological harm or pain.

- **Debriefing**: If a participant is deceived in any way before or during a study, the psychologist must provide a *debriefing*, which is a *verbal description of the true nature and purpose of a study*. If the participant was changed in any way (e.g., made to feel sad), the psychologist must attempt to undo that change (e.g., ask the person to do a task that will make him or her happy) and restore the participant to the state he or she was in before the study.

- **Confidentiality**: Psychologists are obligated to keep private and personal information obtained during a study confidential.

These are just some of the rules that psychologists must follow. But how are those rules enforced? Almost all psychology studies are performed by psychologists who work at colleges and universities. These institutions have institutional review boards (IRBs) that are composed of instructors and researchers, university staff, and laypeople from the community (e.g., business leaders or members of the clergy). If the research is federally funded, the law requires that the IRB include at least one nonscientist and one person who is not affiliated with the institution. (See Other Voices: Can We Afford...
Science? for more about the federal funding of psychological science.) A psychologist may conduct a study only after the IRB has reviewed and approved it.

As you can imagine, the code of ethics and the procedure for approval are so strict that many studies simply cannot be performed anywhere, by anyone, at any time. For example, psychologists would love to know how growing up without exposure to language affects a person’s subsequent ability to speak and think, but they cannot ethically manipulate that variable in an experiment. They can only study the natural correlations between language exposure and speaking ability, and so may never be able to firmly establish the causal relationships between these variables. Indeed, there are many questions that psychologists will never be able to answer definitively because doing so would require unethical experiments that violate basic human rights.

Respecting Animals

Not all research participants have human rights because not all research participants are human. Some are chimpanzees, rats, pigeons, or other nonhuman animals. The American Psychological Association’s code specifically describes the special rights of these nonhuman participants, and some of the more important ones are as follows:

- All procedures involving animals must be supervised by psychologists who are trained in research methods and experienced in the care of laboratory animals and who are responsible for ensuring appropriate consideration of the animals’ comfort, health, and humane treatment.
- Psychologists must make reasonable efforts to minimize the discomfort, infection, illness, and pain of animals.
- Psychologists may use a procedure that subjects an animal to pain, stress, or privation only when an alternative procedure is unavailable and when the procedure is justified by the scientific, educational, or applied value of the study.
- Psychologists must perform all surgical procedures under appropriate anesthesia and must minimize an animal’s pain during and after surgery.

That’s good—but is it good enough? Some people don’t think so. For example, philosopher Peter Singer (1975) has argued that all creatures capable of feeling pain have the same fundamental rights, and that treating nonhumans differently from humans is a form of speciesism that is every bit as abhorrent as racism or sexism. Singer’s philosophy has inspired groups such as People for the Ethical Treatment of Animals to call for an end to all research involving nonhuman animals. Unfortunately, it has also inspired some groups to attack psychologists who legally conduct such research. As two researchers (Ringach & Jentsch, 2009, p. 11417) reported:

We have seen our cars and homes firebombed or flooded, and we have received letters packed with poisoned razors and death threats via e-mail and voicemail. Our families and neighbors have been terrorized by angry mobs of masked protesters who throw rocks, break windows, and chant that “you should stop or be stopped” and that they “know where you sleep at night.” Some of the attacks have been cataloged as attempted murder. Adding insult to injury, misguided animal-rights militants openly incite others to violence on the Internet, brag about the resulting crimes, and go as far as to call plots for our assassination “morally justifiable.”

Where do most people stand on this issue? The majority of Americans consider it morally acceptable to use nonhuman animals in research (Gallup, 2018). Indeed, most Americans eat meat, wear leather, and support the rights of hunters, which is to say that most Americans see a sharp distinction between animal and human rights. Science is not in the business of resolving moral controversies, and every individual must draw his or her own conclusions about this issue. But whatever position you take,
Who pays for all the research described in textbooks like this one? The answer is you. By and large, scientific research is funded by government agencies, such as the National Science Foundation, which give scientists grants (also known as money) to do particular research projects that the scientists have proposed. Of course, this money could be spent on other things, such as feeding the poor, housing the homeless, caring for the ill and elderly, and so on. Does it make sense to spend taxpayer dollars on psychological science when some of our fellow citizens are cold and hungry?

The legal scholar Cass Sunstein argues that research in the behavioral sciences is not an expenditure—it is an investment. Behavior research raises two legitimate concerns. The first is complexity and poor design. We can solve those problems—sometimes without spending a penny.

Behavioral research shows that efforts at simplification, or slight variations in wording, can make all the difference. Since 2010, Britain has had its own Behavioral Insights Team, which experimented with a brief addition to a letter to late-paying taxpayers: “The great majority of people in your local area pay their tax on time.” The change, which is being introduced nationally, produced a 15 percent increase in on-time payments and is projected to bring in millions of dollars worth of revenue. When government programs aren’t working, those on the left tend to support more funding, while those on the right want to scrap them altogether. It is better to ask whether the problem is complexity and poor design. We can solve those problems—sometimes without spending a penny.

What do you think? Is Sunstein right? Is psychological science a wise use of public funds? Or is it a luxury that we simply can’t afford?

it is important to note that only a small percentage of the people who champion animal rights engage in abusive or illegal behavior. It is also important to note that only a small percentage of psychological studies involve animals, and that only a small percentage of those studies cause animals pain or harm. Psychologists mainly study people, and when they do study animals, they mainly study their behavior.

Respecting Truth

Institutional review boards ensure that data are collected ethically. But once the data are collected, who ensures that they are ethically analyzed and reported? No one. Psychology, like all sciences, works on the honor system. No authority is charged with monitoring what psychologists do with the data they’ve collected, and no authority is charged with checking to see if the claims they make are true. You may find that a bit odd. After all, we don’t use the honor system in stores (“Take the microwave home and pay us next time you’re in the neighborhood”), banks (“I don’t need to look up your account, just tell me how much money you want to withdraw”), or courtrooms (“If you say you’re innocent, well then, that’s good enough for me”), so why would we expect it to work in science? Are scientists more honest than everyone else?

Definitely! Okay, we just lied. But the honor system doesn’t depend on scientists being especially honest; it depends on the fact that science is a community enterprise. When scientists claim to have discovered something important, other scientists don’t just applaud: They start studying it too. When the physicist Jan Hendrik Schön announced in 2001 that he had produced a molecular-scale transistor, other physicists were deeply impressed—that is, until they tried to replicate his work and discovered that Schön had fabricated his data (Agin, 2007). Schön lost his job and his PhD, but the important point is that such frauds can’t last long because one scientist’s conclusion is the next scientist’s research question.

This doesn’t mean that all frauds are uncovered swiftly, however. The psychologist Diederik Stapel lied, cheated, and made up his data for decades before people became suspicious enough to investigate (Leveit Committee, Noort Committee, Drenth Committee, 2012), and that’s mainly because the discoveries he claimed to have made were not particularly important to begin with. Not all frauds are uncovered, but all of the important ones are. The psychologist who fraudulently claims to have shown that chimps are smarter than goldfish may never get caught because no one is likely to follow up on such an obvious finding, but the psychologist who fraudulently claims to have shown the opposite will soon have a lot of explaining to do.

What exactly are psychologists on their honor to do? At least three things. First, when writing reports of their studies and publishing them in scientific journals, psychologists are obligated to report truthfully on what they did and what they found. They can’t fabricate results (e.g., by claiming to have performed studies that they never really performed) or fudge results (e.g., by changing records of data that were actually collected), and they can’t mislead by omission (e.g., by reporting only the results that confirm their hypothesis and saying nothing about the results that don’t). Second, psychologists are obligated to share credit fairly by including as co-authors of their reports the other people who contributed to the work, as well as by mentioning in their reports the other scientists who have done related work. And third, psychologists are obligated to share their data. The American Psychological Association’s code of conduct states that ethical psychologists “do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis.” Most scientific frauds have been uncovered by fellow scientists who became suspicious when they looked closely at the fraudster’s data. The fact that anyone can check up on anyone else is part of why the honor system works as well as it does.
Empiricism: How to Know Stuff

- Empiricism is the belief that the best way to understand the world is to observe it firsthand. It is only in the past few centuries that people have begun to systematically collect and evaluate evidence to test the accuracy of their beliefs about the world.
- Observation doesn’t just mean “looking.” It requires a method. The scientific method involves (a) developing a theory that gives rise to a falsifiable hypothesis; and then (b) making observations that serve to test that hypothesis. Although these tests may prove that a theory is false, they can never prove that it is true.
- The methods of psychology are special because human beings are more complex, variable, and reactive than almost anything else that scientists study.

Methods of Observation: Finding Out What People Do

- Measurement involves (a) defining a property in measurable terms and then (b) using a device that can detect that property. A good definition has construct validity (the concrete condition being measured adequately characterizes the property), and a good detector has both power (it can tell when properties are different) and reliability (it can tell when properties are the same).
- When people know they are being observed, they may behave as they think they should. Demand characteristics are aspects of an observational setting that cause people to behave as they think someone else wants or expects them to. Psychologists try to reduce or eliminate demand characteristics by observing participants in their natural habitats or by hiding their expectations from the participant. Observer bias is the tendency for observers’ expectations to influence both what they believe they observed and what actually happened. Psychologists try to avoid observer bias by conducting double-blind studies.
- Psychologists usually measure samples rather than entire populations. They often describe their measurements with a graphic representation called a frequency distribution, which often has a special shape known as the normal distribution. They also describe their measurements with descriptive statistics; the most common are descriptions of central tendency (such as the mean, median, and mode) and descriptions of variability (such as the range and the standard deviation).

Methods of Explanation: Figuring Out Why People Do What They Do

- To determine whether two variables are causally related, we must first determine whether they are related at all. This can be done by measuring each variable many times and then comparing the patterns of variation within each series of measurements. If the patterns are synchronized, then the variables are correlated. Correlations allow us to predict the value of one variable from knowledge of the value of the other. The direction and strength of a correlation are measured by the correlation coefficient (r).
- Even when we observe a correlation between two variables, we can’t conclude that they are causally related because a “third variable” could be causing them both. Experiments solve this third-variable problem by manipulating an independent variable, randomly assigning participants to the conditions that this manipulation creates, and then measuring a dependent variable. These measurements are then compared across conditions. If calculations show that the results would only happen 5% of the time if random assignment had failed, then the differences in the measurements across conditions are assumed to have been caused by the manipulation.
- An internally valid experiment establishes the likelihood of a causal relationship between variables as they were defined and among the participants who were studied. When an experiment mimics the real world, it is externally valid. But most psychology experiments are not attempts to mimic the real world; rather, they test hypotheses derived from theories.
- Random sampling allows researchers to generalize from their samples to the populations from which the samples were drawn, but most psychology studies cannot use random sampling and therefore there are restrictions on the conclusions that can be drawn from them. One restriction is that we can never be absolutely sure that a result is not a fluke, yet replications do help to increase our confidence.
- Replication is an attempt to reproduce a result by using the same procedures and sampling from the same population as...
the original study. Although no one knows the real replication rate in psychological science, the fact that researchers must balance the risk of Type I and Type II errors means that we would not expect the rate to be—or want it to be—100%.

Thinking Critically About Evidence

- Thinking critically about evidence is difficult because people have a natural tendency to see what they expect to see, to see what they want to see, and to consider what they see but not what they don’t see.
- Critical thinkers consider evidence that disconfirms their own opinions. They also consider the evidence that is absent, not just the evidence that is present.
- What makes science different from most other human enterprises is that science actively seeks to discover and remedy its own biases and errors.

The Ethics of Science: Doing What’s Right

- Institutional review boards ensure that the rights of human beings who participate in scientific research are based on the principles of respect for persons, beneficence, and justice.
- Psychologists are obligated to uphold these principles by getting informed consent from participants, not coercing their participation, protecting participants from harm, weighing benefits against risks, avoiding deception, and keeping information confidential.
- Psychologists are obligated to respect the rights of animals and to treat them humanely. Most people are in favor of using animals in scientific research.
- Psychologists are obligated to tell the truth about their studies, to share credit appropriately, and to grant others access to their data.

Key Concept Quiz

1. The belief that accurate knowledge can be acquired through observation is the definition of
   a. critical thinking.
   b. dogmatism.
   c. empiricism.
   d. correlation.

2. Which of the following is the best definition of a hypothesis?
   a. empirical evidence
   b. a scientific investigation
   c. a falsifiable prediction
   d. a theoretical idea

3. If a detector is used to measure the same property twice but produces different measurements, then it lacks
   a. validity.
   b. reliability.
   c. power.
   d. concreteness.

4. Aspects of an observational setting that cause people to behave as they think someone wants or expects them to be called
   a. observer biases.
   b. Type I errors.
   c. Type II errors.
   d. demand characteristics.

5. Which of the following describes the average value of all the measurements in a particular distribution?
   a. mean
   b. median
   c. mode
   d. range

6. What does the sign of \( r \) (the correlation coefficient) show?
   a. the strength of a correlation
   b. the direction of a correlation
   c. the likelihood that random assignment failed
   d. the degree of replicability

7. When two variables are correlated, what keeps us from concluding that one is the cause and the other is the effect?
   a. the third-variable problem
   b. observer bias
   c. the strength of the manipulation
   d. the failure of random assignment

8. A researcher administers a questionnaire concerning attitudes toward tax increases to people of all genders and ages who live all across the United States. The dependent variable in the study is the ____________ of the participants.
   a. age
   b. gender
   c. attitude
   d. geographic location

9. An experiment that defines variables as they are defined in the real world is
   a. externally valid.
   b. internally valid.
   c. operationally defined.
   d. statistically significant.

10. When people find evidence that confirms their beliefs, they often
    a. stop looking.
    b. seek more evidence.
    c. refuse to believe it.
    d. take their enemies to lunch.

Don’t stop now! Quizzing yourself is a powerful study tool. Go to LaunchPad to access the LearningCurve adaptive quizzing system and your own personalized learning plan. Visit launchpadworks.com.
Key Terms

- empiricism (p. 34)
- scientific method (p. 35)
- theories (p. 35)
- hypothesis (p. 35)
- empirical method (p. 36)
- operational definition (p. 38)
- construct validity (p. 38)
- power (p. 38)
- reliability (p. 38)
- demand characteristics (p. 39)
- naturalistic observation (p. 39)
- observer bias (p. 41)
- double-blind study (p. 41)
- population (p. 41)
- sample (p. 41)
- frequency distribution (p. 42)
- normal distribution (p. 43)
- mode (p. 43)
- mean (p. 43)
- median (p. 43)
- range (p. 45)
- standard deviation (p. 45)
- variable (p. 45)
- correlation (p. 46)
- correlation coefficient ($r$) (p. 47)
- natural correlation (p. 48)
- third-variable problem (p. 49)
- experimentation (p. 49)
- manipulation (p. 49)
- independent variable (p. 50)
- dependent variable (p. 50)
- self-selection (p. 51)
- random assignment (p. 51)
- internal validity (p. 53)
- external validity (p. 55)
- case method (p. 56)
- random sampling (p. 56)
- replication (p. 58)
- Type I error (p. 59)
- Type II error (p. 59)
- informed consent (p. 65)
- debriefing (p. 65)

Changing Minds

1. A research study shows that getting a good night’s sleep increases people’s performance on almost any kind of task. You tell a classmate about this study and she shrugs. “Who didn’t already know that? If you ask me, psychology is just common sense. Why conduct experiments to show what everyone already knows?” How would you explain the value of studying something that seems like “common sense”?

2. Your friend texts you a link to a study showing that Europeans who work longer hours are less happy than those who work shorter hours, but that in the United States it’s the other way around. The text reads “Cool experiment!” so you reply “Study—not experiment,” either because you are very wise or because you don’t know much about friendship. Why aren’t all research studies experiments? What can’t you learn from this study that you could learn from an experiment?

3. After the first exam, your professor says she’s noticed a positive correlation between the location of students’ seats and their exam scores: “The closer students sit to the front of the room, the higher their scores on the exam,” she announces. After class, your friend suggests that the two of you should sit up front for the rest of the semester to improve your grades. Having read about correlation and causation, should you be skeptical? What are some possible reasons for the correlation between seating position and good grades? Could you design an experiment to test whether sitting up front actually causes good grades?

4. A classmate in your criminal justice course suggests that mental illness is a major cause of violent crimes in the United States. As evidence, he mentions a highly publicized murder trial in which the convicted suspect was diagnosed with schizophrenia. What scientific evidence would he need to support this claim?

5. You ask your friend if she wants to go to the gym with you. “No,” she says, “I never exercise.” You tell her that regular exercise has all kinds of health benefits, including greatly reducing the risk of heart disease. “I don’t believe that,” she replies. “I had an uncle who got up at 6 a.m. every day of his life to go jogging, and he still died of a heart attack at age 53.” What would you tell your friend? Does her uncle’s case prove that exercise really doesn’t protect against heart disease after all?

Answers To Key Concept Quiz

1. c; 2. c; 3. b; 4. d; 5. a; 6. b; 7. a; 8. c; 9. a; 10. a

LaunchPad features the full e-Book of Psychology, the LearningCurve adaptive quizzing system, videos, and a variety of activities to boost your learning. Visit LaunchPad at launchpadworks.com.