chapter 2

Research Design, Measurement, and Testing Hypotheses

Tony Burns/Getty Images

Chapter Contents

- Overview of Research Designs
- Reliability and Validity
- Scales and Types of Measurement
- Hypothesis Testing

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

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

This story illustrates two key points about the scientific process. First, as we discussed in Chapter 1, it is always good to be attentive to your apparent mistakes because they can lead to valuable insights. Second, it is absolutely vital to measure what you think you are measuring. In this chapter, we get more concrete about what it means to do research, beginning with a broad look at the three types of research design. Our goal at this stage is to get a general sense of what these designs refer to, when they are used, and the main differences among them. (Chapters 3, 4, and 5 are each dedicated to different types of research design and elaborate further on each one.) Following our overview of designs, this chapter covers a set of basic principles that are common to all quantitative research designs. Regardless of the particulars of your design, all quantitative research studies involve making sure our measurements are accurate and consistent and that they are captured using the appropriate type of scale. Finally, we will discuss the general process of hypothesis testing, from laying out predictions to drawing conclusions.

2.1 Overview of Research Designs

As you learned in Chapter 1, scientists can have a wide range of goals going into a research project, from describing a phenomenon to attempting to change people's behavior. It turns out that these goals lend themselves to different approaches to answering a research question. That is, you will approach the problem differently when you want to describe voting patterns than when you want to explain them or predict them. These approaches are called **research designs**, or the specific methods that are used to collect, analyze, and interpret data. The choice of a design is not one to be made lightly; the way you collect data trickles down to the kinds of conclusions that you can draw

about them. This section provides a brief introduction to the four main types of design: qualitative, descriptive, correlational, and experimental.

Qualitative Research

You will recall from Chapter 1 that qualitative research is used to gain a deep and thorough understanding of particular cases and contexts. It is often used when the researcher wants to obtain more detailed and rich data about personal experiences, events, and behaviors in their natural environment. If your research question seeks to obtain insight into and to thoroughly understand people's attitudes, behaviors, value systems, concerns, motivations, aspirations, culture, or lifestyles from *their* perspective, then your research design will fall under the category of *qualitative research*. Qualitative research can be very time-consuming because it delves into great detail about the phenomena of interest, such as people's reactions to a particular situation, how a group interacts over time, or how a person behaves in certain environments and circumstances. The following are examples of qualitative research questions:

- How do women in a psychology doctoral program describe their decision to attend an online program versus a campus-based program?
- What is it like to live with a family member who has Alzheimer's disease?
- What are the familial experiences of teenagers who join gangs?
- How do women who have lost their spouse from a tragic accident experience grief?
- What is the nature of the culture of people living on the island of Niihau?

What these five questions have in common is that they use the words *What* and *How* in an attempt to discover, understand, explore, and describe experiences. They are not trying to explain the causes of a phenomenon or to predict cause and effect.

Unlike the other designs that will be discussed in this chapter, qualitative research produces data in the form of words, transcripts, pictures, and stories and generally cannot (or at least not easily) be converted into numerical data. Thus, qualitative research focuses on building holistic and largely narrative descriptions to provide an understanding of a social or cultural phenomenon.

As we will review in Chapter 3, qualitative research is conducted in a natural setting and involves building a complex and holistic picture of the phenomenon of interest. The researcher immerses him- or herself into the study and interacts with participants to obtain a better understanding of their experiences. The goal of qualitative research is not to test hypotheses but rather to uncover patterns that help explain a phenomenon of interest. Thus, qualitative research begins with research questions and may offer hypotheses *after* the study has been conducted. Because of these traits, qualitative research is often conducted on topics that have not been well researched or on topics that are fairly new.

Descriptive Research

Recall from Chapter 1 that one of the basic goals of research is to describe a phenomenon. If your research question centers around description, then your research design falls under the category of **descriptive research**, in which the primary goal is to describe thoughts, feelings, or behaviors. Descriptive research provides a static picture of what

people are thinking, feeling, and doing at a given moment in time, as seen in the following examples of research questions:

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

What these five questions have in common is that they attempt to describe a phenomenon without trying to delve into its causes.

Jennifer Graylock/Associated Press

Dr. Oliver Sacks studies how people with neurological damage form and retain memories.

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

Descriptive research can be either qualitative or quantitative. Descriptions are quantitative when they include hypotheses and attempt to make comparisons and/or to present a random sampling of people's opinions. The majority of our sample questions above would fall into this group because they quantify opinions from samples of households, or cities, or college students. Good examples of quantitative description appear in the "snapshot" feature on the front page of *USA Today*. The graphics represent poll results from various sources; the snapshot for August 3, 2011, reveals that only 61% of Americans turn off the water while they brush their teeth (i.e., behavior).

Descriptive designs are qualitative when they include research questions and attempt to provide a rich description of a particular set of circumstances. A great example of this approach can be found in the work of neurologist Oliver Sacks. Sacks has written several books exploring the ways that people with neurological damage or deficits are able to navigate the world around them. In one selection from *The Man Who Mistook His Wife*

for a Hat (1998), Sacks relates the story of a man he calls William Thompson. As a result of chronic alcohol abuse, Thompson developed Korsakov's syndrome, a brain disease marked by profound memory loss. The memory loss was so severe that Thompson had effectively "erased" himself and could remember only scattered fragments of his past.

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

> I am a grocer, and you're my customer, right? Well, will that be paper or plastic? No, wait, why are you wearing that white coat? You must be Hymie, the kosher butcher. Yep. That's it. But why are there no bloodstains on your coat? (Sacks, 1998, p. 112)

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

Correlational Research

The second goal of research that we discussed in Chapter 1 was to predict a phenomenon. If your research question centers around prediction, then your research design falls under the category of **correlational research**, in which the primary goal is to understand the relationships among various thoughts, feelings, and behaviors. Examples of correlational research questions include:

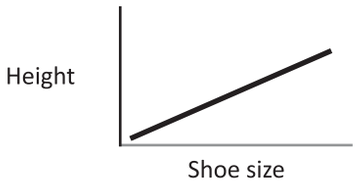
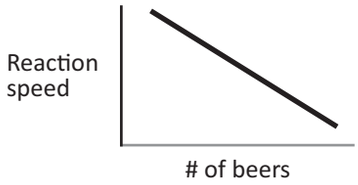
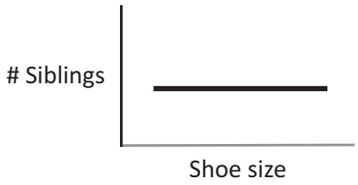
- Are people more aggressive on hot days?
- Are people more likely to smoke when they are drinking?
- Is income level associated with happiness?
- What is the best predictor of success in college?
- Does television viewing relate to hours of exercise?

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

These predictive relationships can turn out in one of three ways (more detail on each one when we get to Chapter 4): A **positive correlation** means that higher values of one variable predict higher values of the other variable. As in, more money predicts higher levels of happiness, and less money predicts lower levels of happiness. The key is that these variables move up and down together, as shown in the first row of Table 2.1. A **negative correlation** means that higher values of one variable predict lower values of the other variable. As in, more television viewing predicts fewer hours of exercise, and fewer hours of television predict more hours of exercise. The key is that one variable increases while the other decreases, as seen in the second row of Table 2.1. Finally, it is worth noting a

third possibility, which is to have no correlation between two variables, meaning that you cannot predict one variable based on another. The key is that changes in one variable are not associated with changes in the other, as seen in the third row of Table 2.1.

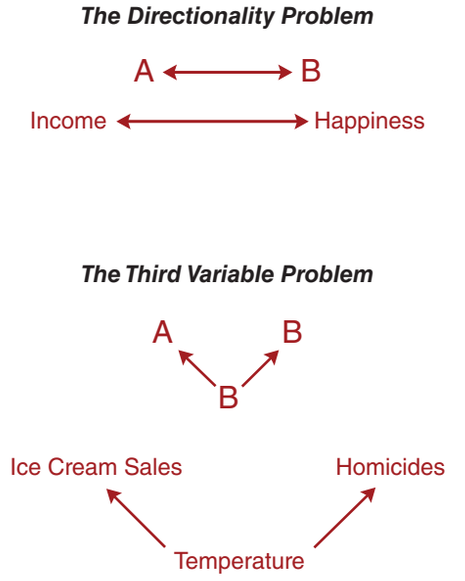
Table 2.1: Three possibilities for correlational research

Outcome	Description	Visual
Positive Correlation	Variables go up and down together For example: Taller people have bigger feet, and shorter people have smaller feet	
Negative Correlation	One variable goes up and the other goes down For example: as the number of beers consumed goes up, speed of reactions go down	
No Correlation	The variables have nothing to do with one another For example: shoe size and number of siblings are completely unrelated	

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

First, we do not know the direction of the relationship; A could cause B or B could cause A. For example, money could cause people to be happier, or happiness could give people the confidence to find higher-paying jobs. Second, there could be a third variable that causes both of our variables to change. For example, increases in temperature could lead to increases in both homicide rates and ice cream sales, making it seem like these variables are related to one another.

Figure 2.1: Correlation is not causation



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

Second, when we measure two variables as they naturally occur, there is always the possibility of a third variable that actually causes both of them. For example, imagine we find a correlation between the number of churches and the number of liquor stores in a city. Do people build more churches to offset the threat of vice encouraged by liquor stores? Or do people build more liquor stores to rebel against the moral code of churches? Most likely, the link involves the third variable of population: The more people there are living in a city, the more churches and liquor stores they can support.

Or, consider this example from analyses of posts on the recommendation website Hunch.com. One of the cofounders of the website conducted extensive analyses of people's activity and brand preferences and found a positive correlation between how much people liked to dance and how likely they were to prefer Apple computers (Fake, 2009). Does this mean that owning a Mac makes you want to dance? Does dancing make you think highly of Macs? Most likely, the link here involves a third variable of personality: People who are more unconventional may be more likely to prefer both Apple computers and dancing.

Experimental Research

Finally, recall that the most powerful goal of research is to attempt to explain and make cause-and-effect statements about a phenomenon. When your research goal involves explanation, then your research design falls under the category of experimental research, in which the primary goal is to explain thoughts, feelings, and behaviors and to make causal statements. Examples of experimental research questions include:

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

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

Experimental designs are able to address the shortcomings of correlational designs because the researcher has more control over the environment. We will cover this in great detail in Chapter 5, but for now, experiments are a relatively simple process: A researcher has to control the environment as much as possible so that all participants in the study have the same experience. He or she will then manipulate, or change, one key variable and then measure outcomes in another key variable. The variable that gets manipulated by the experimenter is called the *independent variable*. The outcome variable that is measured by the experimenter is called the *dependent variable*. The combination of controlling the setting and changing one aspect of this setting at a time allows the researcher to state with some certainty that the changes caused something to happen.

Let's make this a little more concrete. Imagine that you wanted to test the hypothesis that meditation causes improvements in health. In this case, meditation would be the independent variable and health would be the dependent variable. One way to test this hypothesis would be to take a group of people and have half of them meditate 20 minutes per day for several days while the other half did something else for the same amount of time. The group that meditates would be the **experimental group** because it provides the test of our hypothesis. The group that does not meditate would be the **control group** because it provides a basis of comparison for the experimental group. You would want to make sure that these groups spent the 20 minutes in similar conditions so that the only difference would be the presence or absence of meditation. One way to accomplish this would be to have all participants sit quietly for the 20 minutes but give the experimental group specific instructions on how to meditate. Then, to test whether meditation led to increased health and happiness, you would give both groups a set of outcome measures—perhaps a combination of survey measures and a doctor's examination. If you found differences between these groups on the dependent measures, you could be fairly confident that meditation caused them to happen. For example, you might find lower blood pressure in the experimental group; this would suggest that meditation causes a drop in blood pressure.

Kraig Scarbinsky/Thinkstock

Testing the hypothesis that meditation improves health requires an experimental group and a control group.

Research: Making an Impact

Helping Behaviors

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

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

This research can be seen in action and has influenced safety measures in today's society. For example, when witnessing an emergency, no longer does it suffice to simply yell to the group, "Call 9-1-1!" Because of the bystander effect, we know that most people will believe someone else will do it, and the call will not be made. Instead, it is necessary to point to a specific person to designate them as the person to make the call. In fact, part of modern-day CPR training involves making individuals aware of the bystander effect and best practices for getting people to help and be accountable.

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

Choosing a Research Design

The choice of a research design is guided first and foremost by your research topic and research question, and then adjusted depending on practical and ethical concerns. At this point, there may be a nagging question in the back of your mind: If experiments are the most powerful type of design, why not use them every time? Why would you ever give up the chance to make causal statements? One reason is that we are often interested in variables that cannot be manipulated, for ethical or practical reasons, and that therefore have to be studied as they occur naturally. In one example, Matthias Mehl and Jamie Pennebaker happened to start a weeklong study of college students' social lives on September 10, 2001. Following the terrorist attacks on the morning of September 11, Mehl and Pennebaker were able to track changes in people's social connections and use this to understand how groups respond to traumatic events (Mehl & Pennebaker, 2003). Of course, it would have been unthinkable to experimentally manipulate a terrorist attack for this study, but since it occurred naturally, the researchers were able to conduct a correlational study of coping.

Another reason to use qualitative, descriptive, and correlational designs is that these are useful in the early stages of research. For example, before you start to think about the causes of binge drinking among college students, it is important to understand the experiences of binge drinkers and how common this phenomenon is. Before you design a time- and cost-intensive experiment on the effects of meditation, it is a good idea to conduct a correlational study to test whether meditation even predicts health. In fact, this example comes from a series of real research studies conducted by psychiatrist Sara Lazar and her colleagues at Massachusetts General Hospital. This research team first discovered that experienced practitioners of mindfulness meditation had more development in brain areas associated with attention and emotion. But this study was correlational at best; perhaps meditation causes changes in brain structure or perhaps people who are better at integrating emotions are drawn to meditation. In a follow-up study, they randomly assigned people to either meditate or complete stretching exercises for 2 months. These experimental findings confirmed that mindfulness meditation actually caused structural changes to the brain (Hölzel et al., 2011). In addition, this is a fantastic example of how research can progress from correlational to experimental designs. Table 2.2 summarizes the main advantages and disadvantages of our four types of designs.

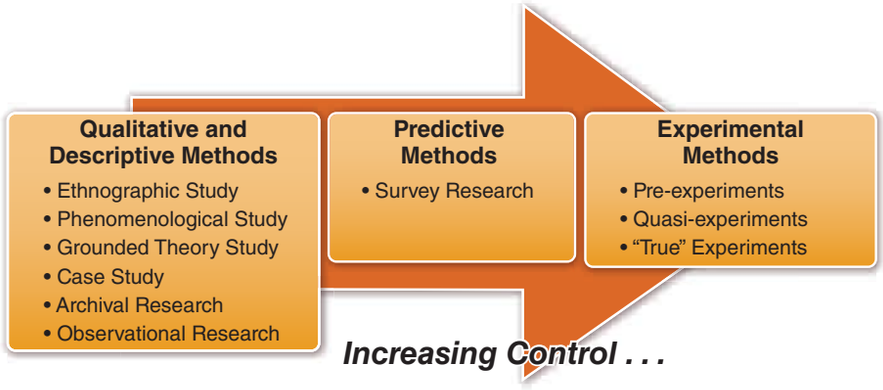
Table 2.2: Summary of research designs

Research Design	Goal	Advantages	Disadvantages
Qualitative	*Obtain insight and detailed descriptions* into people's attitudes, behaviors, value systems, concerns, motivations, aspirations, culture, or lifestyles	Does not require a strict design plan before the study begins; Uncovers in-depth and rich information about people's experiences in a natural setting; focuses on people's *individual* experiences	Does not assess relationships; difficult to make comparisons; difficult to make assumptions beyond the sample being studied; very time-consuming; high level of researcher involvement could skew results
Descriptive	*Describe* characteristics of an existing phenomenon	Provides a complete picture of what is occurring at a given time	Does not assess relationships; no explanation for phenomenon
Correlational	*Predict* behavior; assess strength of relationship between variables	Allows testing of expected relationships; enables predictions	Cannot draw inferences about causal relationships
Experimental	*Explain* behavior; assess impact of independent variable and dependent variable	Allows conclusions to be drawn about causal relationships	Many important variables cannot be manipulated

Designs on the Continuum of Control

Before we leave our design overview behind, a few words on how these designs relate to one another. The best way to think about the differences between the designs is in terms of the amount of control you have as a researcher. That is, experimental designs are the most powerful because the researcher controls everything from the hypothesis to the environment in which the data are collected. Correlational designs are less powerful because the researcher is restricted to measuring variables as they occur naturally. However, with correlational designs, the researcher does maintain control over several aspects of data collection, including the setting and the choice of measures. Descriptive designs and qualitative designs are the least powerful because it is difficult to control outside influences on data collection. For example, when people answer opinion polls over the phone, they might be sitting quietly and pondering the questions or they might be watching television, eating dinner, and dealing with a fussy toddler. In the case of unstructured, qualitative interviews, the researcher generally exerts little control over the direction of the interview and might obtain different information from various participants, making it difficult to make comparisons across the data. (We will discuss qualitative methods and interviews further in Chapter 3.) As a result, a researcher is more limited in the conclusions he or she can draw from these data. Figure 2.2 shows an overview of research designs in order of increasing control, from qualitative and descriptive, to predictive, and to experimental. As we progress through Chapters 3, 4, and 5, we will cover variations on these designs in more detail.

Figure 2.2: Research designs on the continuum of control



2.2 Reliability and Validity

Before beginning this section and the rest of this chapter, it should be noted that qualitative research and qualitative descriptive designs do not test for hypotheses. Rather, they seek to answer research questions in order to understand and describe behaviors, experiences, or phenomena and to potentially form hypotheses after the study has been conducted. In addition, reliability and validity are thought about quite differently in qualitative research designs and utilize different concepts, such as credibility, transferability, confirmability, and dependability. As a result, qualitative designs are not discussed in the following sections of this chapter but will be discussed further in Chapters 3 and 5.

Each of the three quantitative designs described in this chapter (descriptive-quantitative, correlational, and experimental) have the same basic goal: to take a hypothesis about some phenomenon and translate it into measurable and testable terms. That is, whether we use a descriptive, correlational, or experimental design to test our predictions about income and happiness, we still need to translate (or operationalize) the concepts of income and happiness into measures that will be useful for the study. The sad truth is that our measurements will always be influenced by factors other than the conceptual variable of interest. Answers to any set of questions about happiness will depend both on actual levels of happiness and the ways people interpret the questions. Our meditation experiment may have different effects depending on people's experience with meditation. Even describing the percentage of Republicans voting for independent candidates will vary depending on characteristics of a particular candidate.

These additional sources of influence can be grouped into two categories: random and systematic errors. **Random error** involves chance fluctuations in measurements, such as when a few people misunderstand the question or the experimenter enters the wrong values into a statistical spreadsheet. Although random errors can influence measurement, they generally cancel out over the span of an entire sample. That is, some people may overreact to a question while others underreact. The experimenter may accidentally type a 6 instead of a 5 but then later type a 5 instead of a 6 when entering the data. While both of these examples would add error to our dataset, they would cancel each other out in a sufficiently large sample.

Systematic errors, in contrast, are those that systematically increase or decrease along with values on our measured variable. For example, people who have more experience with meditation may show consistently more improvement in our meditation experiment than those with less experience. Or, people with higher self-esteem may score higher on our measure of happiness than those with lower self-esteem. In this case, our happiness scale will end up assessing a combination of happiness and self-esteem. These types of errors can cause more serious trouble for our hypothesis tests because they interfere with our attempts to understand the link between two variables.

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

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

For example:

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

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

Reliability

The consistency of time measurement by watches, cell phones, and clocks reflects a high degree of reliability. We think of a watch as reliable when it keeps track of the time consistently. Likewise, our scale is reliable when it gives the same value for our weight in back-to-back measurements.

Reliability is technically defined as the extent to which a measured variable is free from random errors. As we discussed above, our measures are never perfect, and reliability is threatened by five main sources of random error:

- *Transient states, or temporary fluctuations in participants' cognitive or mental state*; for example, some participants may complete your study after an exhausting midterm or in a bad mood after a fight with their significant others.
- *Stable individual differences among participants*; for example, some participants are habitually more motivated, or happier, than other participants.
- *Situational factors in the administration of the study*; for example, running your experiment in the early morning may make everyone tired or grumpy.
- *Bad measures that add ambiguity or confusion to the measurement*; for example, participants may respond differently to a question about "the kinds of drugs you are taking." Some may take this to mean illegal drugs, whereas others interpret it as prescription or over-the-counter drugs.
- *Mistakes in coding responses during data entry*; for example, a handwritten 7 could be mistaken for a 4.

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

Test–retest reliability refers to the consistency of our measure over time—much like our examples of a reliable watch and a reliable scale. A fair number of research questions in the social and behavioral sciences involve measuring stable qualities. For example, if you were to design a measure of intelligence or personality, both of these characteristics should be relatively stable over time. Your score on an intelligence test today should be roughly the same as your score when you take it again in 5 years. Your level of extraversion today should correlate highly with your level of extraversion in 20 years. The test–retest reliability of these measures is quantified by simply correlating measures at two time points. The higher these correlations are, the higher the reliability will be. This makes conceptual sense as well; if our measured scores reflect the true score more than they reflect random error, then this will result in increased stability of the measurements.

Interitem reliability refers to the internal consistency among different items on our measure. If you think back to the last time you completed a survey, you may have noticed that it seemed to ask the same questions more than once (more on this technique in Chapter 4 (4.1). This is done because a single item is more likely to contain measurement error than is the average of several items—remember that small random errors tend to cancel out. Consider the following items from Sheldon Cohen's *Perceived Stress Scale* (Cohen, Kamarck, & Mermelstein, 1983):

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

Each of these items taps into the concept of "stressed out," or overwhelmed by the demands of one's life. One standard way to evaluate a measure like this is by computing the average correlation between each pair of items, a statistic referred to as **Cronbach's alpha**. The more these items tap into a central, consistent construct, the higher the value of this statistic is. Conceptually, a higher alpha means that variation in responses to the different items reflects variation in the "true" variable being assessed by the scale items.

Interrater reliability refers to the consistency among judges observing participants' behavior. The previous two forms of reliability were relevant in dealing with self-report scales; interrater reliability is more applicable when research involves behavioral measures. Imagine you are studying the effects of alcohol consumption on aggressive behavior. You would most likely want a group of judges to observe participants in order to make ratings of their levels of aggression. In the same way that using multiple scale items helps to cancel out the small errors of individual items, using multiple judges cancels out the variations in each individual's ratings. In this case, people could have different ideas and thresholds for what constitutes aggression. Much like the process of evaluating multiple scale items, we can evaluate the judges' ratings by calculating the average correlation among the ratings. The higher our alpha values, the more the judges agree in their ratings of aggressive behavior. Conceptually, a higher alpha value means that variation in the judges' ratings reflects real variation in levels of aggression.

Validity

Let's return to our watch and scale examples. Perhaps you are the type of person who sets your watch 10 minutes ahead to avoid being late. Or perhaps you have adjusted your scale by 5 pounds to boost your motivation or your self-esteem. In these cases, your watch and your scale may produce consistent measurements, but the measurements are not accurate. It turns out that the reliability of a measure is a necessary but not sufficient basis for evaluating it. Put bluntly, our measures can be (and have to be) consistent but might still be garbage. The additional piece of the puzzle is the *validity* of our measures, or the extent to which they accurately measure what they are designed to measure.

Whereas reliability is threatened more by random error, validity is threatened more by systematic error. If the measured scores on our happiness scale reflect, say, self-esteem more than they reflect happiness, this would threaten the validity of our scale. We discussed in the previous section that a test designed to measure intelligence ought to be consistent over time. And in fact, these tests do show very high degrees of reliability. However, several researchers have cast serious doubts on the validity of intelligence testing, arguing that even scores on an official IQ test are influenced by a person's cultural background, socioeconomic status (SES), and experience with the process of test taking (for discussion of these critiques, see Daniels et al., 1997; Gould, 1996). For example, children growing up

in higher SES households tend to have more books in the home, spend more time interacting with one or both parents, and attend schools that have more time and resources available—all of which are correlated with scores on IQ tests. Thus, all of these factors amount to systematic error in the measure of intelligence and, therefore, threaten the validity of a measured score on an intelligence test.

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

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

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

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

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

On the less subjective end, we can assess the validity of our constructs by examining their empirical connections to both related and unrelated measures. Imagine that you wanted to develop a new measure of narcissism, usually defined as an intense desire to be liked and admired by other people. Narcissists tend to be self-absorbed but also very attuned to the feedback they receive from other people—at least as it pertains to the extent to which people admire them. Narcissism is somewhat similar to self-esteem but different enough; it is perhaps best viewed as high and unstable self-esteem. So, given these facts, we might assess the **discriminant validity** of our measure by making sure it did not overlap too closely with measures of self-esteem or self-confidence. This would establish that our measure stands apart from these different constructs. We might then assess the **convergent validity** of our measure by making sure that it did correlate with things like sensitivity to rejection and need for approval. These correlations would place our measure into a broader theoretical context and help to establish it as a valid measure of the construct of narcissism.

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

That said, criterion validity is also a useful way to supplement validation of a new questionnaire. For example, a questionnaire about generosity should be able to predict people's annual giving to charities, and a questionnaire about hostility ought to predict hostile behaviors. To supplement the construct validity of our narcissism measure, we might examine its ability to predict the ways people respond to rejection and approval. Based on the definition of our construct, we might hypothesize that narcissists would become hostile following rejection and perhaps become eager to please following approval. If these predictions were supported, we would end up with further validation that our measure was capturing the concept of narcissism.

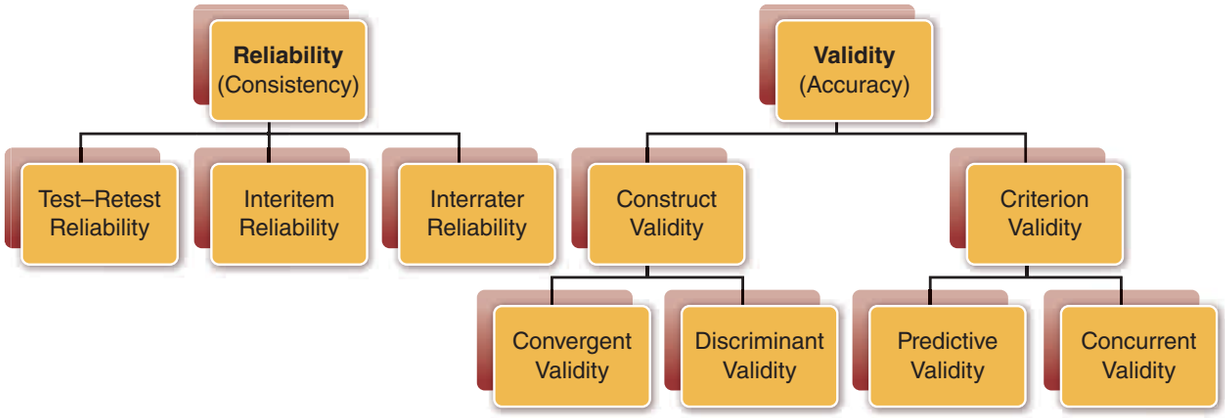
Criterion validity falls into one of two categories, depending on whether the researcher is interested in the present or the future. **Predictive validity** involves attempting to predict a future behavioral outcome based on the measure, as in our examples of the management style and career placement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

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

Summary: Comparing Reliability and Validity

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

Figure 2.3: Types of reliability and validity



We learned earlier that reliability is necessary but not sufficient to evaluate measured variables. That is, reliability has to come first and is an essential requirement for any variable—you would not trust a watch that was sometimes 5 minutes fast and other times 10 minutes slow. If we cannot establish that a measure is reliable, then there is really no chance of establishing its construct validity because every measurement might be a reflection of random error. However, just because a measure is consistent does not make it accurate. Your watch might consistently be 10 minutes fast; your scale might always be 5 pounds under your actual weight. For that matter, your test of intelligence might result in consistent scores but actually be capturing respondents' cultural background. Reliability tells us the extent to which a measure is free from random error. Validity takes the second step of telling us the extent to which the measure is also free from systematic error.

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

2.3 Scales and Types of Measurement

As you may remember from prior statistics classes, not all measures are created equal. One of the easiest ways to decrease error variance, and thereby increase our reliability and validity, is to make smart choices when we design and select our measures. Throughout this book, we will discuss guidelines for each type of research design and ways to ensure that our measures are as accurate and unbiased as

possible. In this section, we examine some basic rules that apply across all three types of design. We first review the four scales of measurement and discuss the proper use of each one; we then turn our attention to three types of measurement used in psychological research studies.

Scales of Measurement

Whenever we go through the process of translating our conceptual variables into measurable variables (i.e., operationalization; see Chapter 1, section 1.2), it is important to ensure that our measurements accurately represent the underlying concepts. We have covered this process already; in our discussion of validity, you learned that this accuracy is a critical piece of hypothesis testing. For example, if we develop a scale to measure job satisfaction, then we need to verify that this is actually what the scale is measuring. But there is an additional, subtler dimension to measurement accuracy: We also need to be sure that our chosen measurement accurately reflects the underlying mathematical properties of the concept. In many cases in the natural sciences, this process is automatically precise. When we measure the speed of a falling object or the temperature of a boiling object, the underlying concepts (speed and temperature) translate directly into scaled measurements. But in the social and behavioral sciences, this process is trickier; we have to decide carefully how best to represent abstract concepts such as happiness, aggression, and political attitudes. As we take the step of **scaling** our variables, or specifying the relationship between our conceptual variable and numbers on a quantitative measure, we have four different scales to choose from, presented below in order of increasing statistical power and flexibility.

Nominal Scales

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

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

Ordinal Scales

Ordinal scales are used to represent ranked orders of conceptual variables. For example, beauty contestants, horses, and Olympic athletes are all ranked by the order in which they finish—first, second, third, and so on. Likewise, movies, restaurants, and consumer goods are often rated using a system of stars (i.e., 1 star is not good; 5 stars is excellent) to represent their quality. In these examples, we can draw conclusions about the relative speed, beauty, or deliciousness of the rating target. But the numbers used to label these rankings do not necessarily map directly onto differences in the conceptual variable. The fourth-place finisher in a race is rarely twice as slow as the second-place finisher; the beauty contest winner is not three times as attractive as the third-place finisher; and the boost in quality between a four-star and a five-star restaurant is not the same as the boost between a two-star and three-star restaurant. Ordinal scales represent rank orders, but the numbers do not have any absolute value of their own. Thus, this type of scale is more powerful than a nominal scale but still limited in that we cannot perform mathematical operations. For example, if an Olympic athlete finished first in the 800-meter dash, third in the 400-meter hurdles, and second in the 400-meter relay, you might be tempted to calculate her average finish as being in second place. Unfortunately, the properties of ordinal scales prevent us from doing this sort of calculation because the distance between first, second, and third place would be different in each case. In order to perform any mathematical manipulation of our variables, we need one of the next two types of scale.

EMPCIS Sport/Associated Press

Olympic athletes are ranked using an ordinal scale.

Interval Scales

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

Ratio Scales

Ratio scales go one final step further, representing interval scales that also have a true zero point, that is, the potential for a complete absence of the conceptual variable. Ratio scales can be used in the case of physical measurements, such as length, weight, and time since

it is possible to have a complete absence of any of these. Ratio scales can also be used in measurement of behaviors since it is possible to have zero drinks per day, zero presses of a reward button, or zero symptoms of the flu. Temperature in degrees Kelvin is measured on a ratio scale because 0 Kelvin indicates an absence of molecular motion. (In contrast, 0 degrees Fahrenheit is only a center point on the temperature scale.) Contrast these measurements with many of the conceptual variables featured in psychology research—there is no such thing as zero happiness or zero self-esteem. The big advantage of having a true zero point is that it allows us to add, subtract, multiply, and divide scale values. When we measure weight, for example, it makes sense to say that a 300-pound adult weighs twice as much as a 150-pound adult. And, it makes sense to say that having two drinks per day is only ¼ as many as having eight drinks per day.

Summary—Choosing and Using Scales of Measurement

The take-home point from our discussion of these four scales of measurement is twofold. First, you should always use the most powerful and flexible scale possible for your conceptual variables. In many cases, there is no choice; time is measured on a ratio scale and gender is measured on a nominal scale. But in some cases, you have a bit of freedom in designing your study. For example, if you were interested in correlating weight with happiness, you could capture weight in a few different ways. One option would be to ask people their satisfaction with their current weight on a seven-point scale. However, the resulting data would be on an ordinal or interval scale (see discussion below), and the degree to which you could manipulate the scale values would be limited. Another, more powerful option would be to measure people's weight on a scale, resulting in ratio scale data. Thus, whenever possible, it is preferable to incorporate physical or behavioral measures. But the primary goal is also to represent your data accurately. Most variables in the social and behavioral sciences do not have a true zero point and must therefore be measured on nominal, ordinal, or interval scales.

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

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

Types of Measurement

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

Behavioral Measurement

Behavioral measures are those that involve direct and systematic recording of observable behaviors. If your research question involves the ways that married couples deal with conflict, you could include a behavioral measure by observing the way participants interact during an argument. Do they cut one another off? Listen attentively? Express hostility? Behaviors can be measured and quantified in one of four primary ways, as illustrated using the scenario of observing married couples during conflict situations:

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

John Gottman, a psychologist at the University of Washington, has been conducting research along these lines for several decades (Gottman & Levenson, 1992), observing body language and interaction styles among married couples as they discuss an unresolved issue in their relationship (you can read more about this research and its implications for therapy on Dr. Gottman's website, http://www.gottman.com/). What all of these behavioral measures provide is a nonreactive way to measure the health of a relationship. That is, the major strength of behavioral responses is that they are typically more honest and unfiltered than responses to questionnaires. As we will discuss in Chapter 4 (4.1), people are sometimes dishonest on questionnaires in order to convey a more positive (or less negative) impression.

This is a particular plus if you are interested in unpopular attitudes, such as prejudice and discrimination. If you were to ask people the extent to which they disliked members of other ethnic groups, they might not admit to these prejudices. Alternatively, you could adopt the approach used by Yale psychologist Jack Dovidio and colleagues and measure how close people sat to people of different ethnic and racial groups, using this distance as a subtle and effective behavioral measure of prejudice (see http://www.yale.edu/intergroup/ for more information). But you may have spotted the primary downside to using behavioral measures: We end up having to infer the reasons that people behave as they do. Let's say European-American participants, on average, sit farther away from

African-Americans than from other European-Americans. This could—and usually does—indicate prejudice; but, for the sake of argument, the farthest seat from the minority group member might also be the one closest to the window. In order to understand the reasons for behaviors, researchers have to supplement the behavioral measures with either physiological or self-report measurements.

Physiological Measurement

Physiological measures are those that involve quantifying bodily processes, including heart rate, brain activity, and facial muscle movements. If you were interested in the experience of test anxiety, you could measure heart rate as people completed a difficult math test. If you wanted to study emotional reactions to political speeches, you could measure heart rate, facial muscles, and brain activity as people viewed video clips. The big advantage of these types of measures is that they are the least subjective and controllable. It is incredibly difficult to control your heart rate or brain activity consciously, making these a great tool for assessing emotional reactions. However, as with behavioral measures, we always need some way to contextualize our physiological data.

The best example of this shortcoming is the use of the polygraph, or lie detector, to detect deception. The lie detector test involves connecting a variety of sensors to the body to measure heart rate, blood pressure, breathing rate, and sweating. All of these are physiological markers of the body's fight-or-flight stress response; so the goal is to observe whether you show signs of stress while being questioned. But here's the problem: It is also stressful to worry about being falsely accused. A trained polygraph examiner must place all of your physiological responses in the proper context. Are you stressed throughout the exam or only stressed when asked whether you pilfered money from the cash box? Are you stressed when asked about your relationship with your spouse because you killed him or because you were having an affair? The examiner has to be extremely careful to avoid false accusations based on misinterpretations of physiological responses.

Self-Report Measurement

Self-report measures are those that involve asking people to report on their own thoughts, feelings, and behaviors. If you were interested in the relationship between income and happiness, you could simply ask people to report their income and their level of happiness. If you wanted to know whether people were satisfied in their romantic relationships, you could simply ask them to rate their degree of satisfaction. The big advantage of these measures is that they provide access to internal processes. That is, if you want insight into why people voted for their favorite

Andy Sacks/Getty Images

A self-report measure might be used to determine why people voted for a particular political candidate.

political candidate, you could simply ask them. However, as we have suggested already, people may not necessarily be honest and forthright in their answers, especially when dealing with politically incorrect or unpopular attitudes. We will return to this balance again in Chapter 4 and discuss ways to increase the likelihood of honest self-reported answers.

There are two broad categories of self-report measures. One of the most common approaches is to ask for people's responses using a **fixed-format** scale, which asks them to indicate their opinion on a preexisting scale. For example, you might ask people, "How likely are you to vote for the Republican candidate for president?" on a scale from 1 (not likely) to 7 (very likely). The other broad approach is to ask for responses using a **free-response** format, which asks people to express their opinion in an open-ended format. For example, you might ask people to explain, "What are the factors you consider in choosing a political candidate?" The trade-off between these two categories is essentially a choice between data that are easy to code and analyze and data that are rich and complex. In general, fixed-format scales are used more in quantitative research while free-response formats are used more in qualitative research.

Research: Thinking Critically

Neuroscience and Addictive Behaviors

By Christian Nordqvist

Some people really are addicted to foods in a similar way others might be dependent on certain substances, like addictive illegal or prescription drugs, or alcohol, researchers from Yale University revealed in *Archives of General Psychiatry* (Gearhardt et al., 2011). Those with an addiction-like behavior seem to have more neural activity in specific parts of the brain in the same way substance-dependent people appear to have, the authors explained.

It's a bit like saying that if you dangle a tasty chocolate milkshake in front of a pathological eater, what goes on in that person's brain is similar to what would happen if you placed a bottle of scotch in front of an alcoholic.

The researchers wrote:

> One-third of American adults are now obese and obesity-related disease is the second leading cause of preventable death. Unfortunately, most obesity treatments do not result in lasting weight loss because most patients regain their lost weight within five years. Based on numerous parallels in neural functioning associated with substance dependence and obesity, theorists have proposed that addictive processes may be involved in the etiology of obesity. Food and drug use both result in dopamine release in mesolimbic regions and the degree of release correlates with subjective reward from both food and drug use.

The authors believe that no studies had so far looked into the neural correlates of addictive-like eating behavior. They explained that some studies had demonstrated that photos of nice food can get the brain's reward centers to become more active in much the same way that photos of alcoholic drinks might do for alcoholics. However, this latest study is the first to tell the food addicts from the just overeaters.

(continued)

Research: Thinking Critically *(continued)*

Ashley N. Gearhardt, M.S., M.Phil., and team looked at the relation between the symptoms of food addiction and neural activation. Food addiction was assessed by the Yale Food Addiction Scale, while neural activation was gauged via functional MRI (magnetic resonance imaging). Forty-eight study participants responded to cues that signaled the imminent arrival of very tasty food, such as a chocolate milkshake, compared to a control solution (something with no taste). They also compared what was going on while they consumed the milkshake compared to the tasteless solution.

The Yale Food Addiction Scale questionnaire identified 15 women with high scores for addiction-like eating behaviors. All the 48 study participants were young women, ranging in body mass index (BMI) from lean to obese. They were recruited from a healthy weight maintenance study.

The scientists discovered a correlation between food addiction and greater activity in the amygdala, the medial orbitofrontal cortex (OFC), and the anterior cingulated cortex (ACC) when tasty food delivery was known to arrive soon.

Those with high food addiction, the 15 women, showed greater activity in the dorsolateral prefrontal cortex compared to those with low addiction to foods. They also had reduced activity in the lateral orbitofrontal cortex while they were eating their nice food.

The authors explained:

> As predicted, elevated FA (food addiction) scores were associated with greater activation of regions that play a role in encoding the motivational value of stimuli in response to food cues. The ACC and medial OFC have both been implicated in motivation to feed and to consume drugs among individuals with substance dependence.
>
> In sum, these findings support the theory that compulsive food consumption may be driven in part by an enhanced anticipation of the rewarding properties of food. Similarly, addicted individuals are more likely to be physiologically, psychologically, and behaviorally reactive to substance-related cues.

They concluded:

> To our knowledge, this is the first study to link indicators of addictive eating behavior with a specific pattern of neural activation. The current study also provides evidence that objectively measured biological differences are related to variations in YFAS (Yale Food Addiction Scale) scores, thus providing further support for the validity of the scale. Further, if certain foods are addictive, this may partially explain the difficulty people experience in achieving sustainable weight loss. If food cues take on enhanced motivational properties in a manner analogous to drug cues, efforts to change the current food environment may be critical to successful weight loss and prevention efforts. Ubiquitous food advertising and the availability of inexpensive palatable foods may make it extremely difficult to adhere to healthier food choices because the omnipresent food cues trigger the reward system. Finally,

(continued)

Research: Thinking Critically *(continued)*

> if palatable food consumption is accompanied by disinhibition [loss of inhibition], the current emphasis on personal responsibility as the antidote to increasing obesity rates may have minimal effectiveness.

Nordqvist, C. (2011, April 5). Food addiction and substance dependence, similar brain activity going on. Medical News Today. *Retrieved from* http://www.medicalnewstoday.com/articles/221233.php

Think about it:

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

Choosing a Measurement Type

As you can see from these descriptions, each type of measurement has its strengths and flaws. So, how do you decide which one to use? This question has to be answered for every case, and the answer depends on three factors. First, and most obviously, the measure depends on the research question. If you are interested in effects of public speaking on stress levels, then the best measures will be physiological. If you are interested in attitudes toward capital punishment, these are better measured using self-reports. Second, the choice of measures is guided by previous research on the topic. If studies have assessed prejudice by using self-reports, then you could feel comfortable doing the same. If studies have measured fear responses using facial expressions, then let that be a starting point for your research. Finally, a mix of availability and convenience often guides the choice of measures. Measures of brain activity are a fantastic addition to any research program, but these measures also require a specialized piece of equipment that can run upwards of $2 million. As a result, many researchers interested in physiological measures opt for something less expensive like a measure of heart rate or movement of facial muscles, both of which can be measured using carefully placed sensors (i.e., on the chest or face).

In an ideal world, a program of research will use a wide variety of measures and designs. The term for this is **converging operations**, or the use of multiple research methods to solve a single problem. In essence, over the course of several studies—perhaps spanning several years—you would address your research question using different designs, different measures, and different levels of analysis. One good example of converging operations comes from the research of psychologist James Gross and his colleagues at Stanford University. Gross studies the ways that people regulate their emotional responses and has conducted this work using everything from questionnaires to brain scans (see http://spl.stanford.edu/projects.html).

One branch of Gross's research has examined the consequences of trying to either suppress emotions (pretend they're not happening) or reappraise them (think of them in a different light) (Gross, 1998; Butler et al., 2003). Suppression is studied by asking people to hold in their emotional reactions while watching a graphic medical video. Reappraisal is studied by asking people to watch the same video while trying to view it as a medical student, thus changing the meaning of what they see. When people try to suppress emotional responses, they experience a paradoxical increase in physiological and self-reported emotional responses, as well as deficits in cognitive and social functioning. Reappraising emotions, in contrast, actually works quite well. In another branch of the research, Gross and colleagues have examined the neural processes at work when people change their perspective on an emotional event (Goldin, McRae, Ramel, & Gross, 2008). In yet another branch of the research, they have examined individual differences in emotional responses, with the goal of understanding why some people are more capable of managing their emotions than others. Taken together, these studies all converge into a more comprehensive picture of the process of emotion regulation than would be possible from any single study or method.

2.4 Hypothesis Testing

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

The Logic of Hypothesis Testing

In Chapter 1 (Section 1.3, Research Problem and Questions), we discussed several criteria for identifying a "good" theory, one of which is that our theories have to be falsifiable. In other words, our research questions should have the ability to be proven wrong under the right set of conditions. Why is this so important? This will sound counterintuitive at first, but by the standards of logic, it is more meaningful when data run counter to our theory than when data support the theory.

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

Another way to think about this is from a statistical perspective. As we discussed earlier in this chapter, all measurements contain some amount of random error, which means that any pattern of data could be caused by random chance. This is the primary reason that research is never able to "prove" a theory. You'll also remember from your statistics class that at the end of any hypothesis test, we will calculate a ***p* value**, representing the *probability* that our results are due to random chance. Conceptually, this means we are calculating the probability that we're wrong rather than the probability that we're right in our predictions. And the bigger our effect, the smaller this probability will generally be. So, as strange as this seems, the ideal result of hypothesis testing is to have a small probability of being wrong.

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

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

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

Conceptual Hypothesis: People like cats better than dogs.

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

the "cat" mean is equal to the "dog" mean;

people like cats and dogs the same

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

the "cat" mean is not equal to the "dog" mean;

people like cats and dogs different amounts

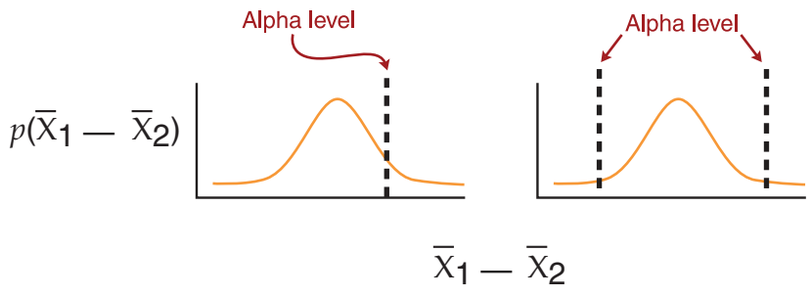
Directional Alternative Hypothesis: H_1: $\mu_{cat} > \mu_{dog}$

the "cat" mean is greater than the "dog" mean;

people like cats more than dogs

Why do we need to distinguish between directional and nondirectional hypotheses? As you'll see when we get to the statistical calculations, this decision has implications for our level of statistical significance. Because we always want to minimize the risk of coming to the wrong conclusion based on chance findings, we have to be more conservative with a nondirectional test. This idea is illustrated in Figure 2.4.

Figure 2.4: One-tailed vs. two-tailed hypothesis tests



These graphs represent the probability of obtaining a particular difference between our groups. The graph on the left represents a simple directional hypothesis—we will be comfortable rejecting the null hypothesis if our mean difference is above the alpha cutoff (figure: This figure shows the difference between one-tailed and two-tailed hypothesis tests. In a one-tailed test, we predict that our group difference will be above a cutoff score. In a two-tailed test, we predict that the difference will be either above or below a cutoff score, usually 5%). The graph on the right, however, represents a nondirectional hypothesis, which simply predicts that one group is higher or lower than the other. Because we are being less specific, we have to be more conservative. With a **directional hypothesis** (also called one-tailed), we predict that the group difference will fall on one extreme of the curve; with a **nondirectional hypothesis** (also called two-tailed), we predict that the group difference will fall on either extreme of the curve. The implication of a two-tailed hypothesis is that our 5% cutoff could become a 10% cutoff, with 5% on each side. Rather than double our chance of an error, we follow standard practice and use a 2.5% cutoff on each side of the curve.

Translation: We need bigger group differences to support our two-tailed, nondirectional hypotheses. In the cats-versus-dogs example, it would take a bigger difference in ratings to support the claim that people like cats and dogs different amounts than it would to support the claim that people like cats more than dogs. The goal of all this statistical and logical jargon is to place our hypothesis testing in the proper frame. The most important thing to remember is that hypothesis testing is designed to reject the null hypothesis, and our statistical tests tell us how confident to be in this rejection.

Five Steps to Hypothesis Testing

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

Step 1—State the Hypothesis

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

Step 2—Collect Data

The next step is to design and conduct a study that will test our hypothesis. We will elaborate on this process in great detail over the next three chapters, but the general idea is the same regardless of the design. In this case, the most appropriate design would be correlational because we want to predict happiness based on whether people have children. It would be impractical and unethical to randomly assign people to have children, so an experimental design is not possible in this case. One way to conduct our study would be to survey married couples about whether they had children and ask them to rate their current level of happiness with the marriage. Let's say we conduct this experiment and end up with the data in Table 2.3.

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

Table 2.3: Sample data for the "children and happiness" study	
No Children	Children
7	2
5	3
7	1
5	2
4	4
5	3
6	2
7	1
6	1
5	1
mean = 5.7	mean = 2.0
S = 1.06	S = 1.05
SE = .33	SE = .33

Step 3—Calculate Statistics

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

One of the most straightforward statistical tests to understand is **Student's *t*-test**, which is widely used to compare differences in the means of two groups. Because of its simplicity, it is also a great way to demonstrate the hypothesis-testing process. Conceptually, the *t*-test compares the difference between two group means with the overall variability in the data set. The end result is a test of whether our groups differ by a meaningful amount. Imagine you found a 10-point difference in intelligence test scores between Republicans

and Democrats. Before concluding that your favorite party was smarter, you would need to know how much scores varied on average. If your intelligence test were on a 100-point scale, with a standard deviation of 5, then your 10-point difference would be interesting and meaningful. But if you measured intelligence on a 1,000-point scale, with a standard deviation of 100, then 10 points probably wouldn't reflect a real difference.

So, conceptually, the *t*-test is a ratio of the mean difference to the average variability. Mathematically, the *t*-test is calculated like so:

$$t = \frac{x_1 - x_2}{SE_{pooled}}$$

Let's look at the pieces of this formula individually. First, the *x*s on top of the line are a common symbol for referring to the mean, or average, in our sample. Thus the terms x_1 and x_2 refer to the means for groups 1 and 2 in our sample, or the mean happiness for couples with children and no children. The term below the line, SE_{pooled}, represents our estimate of variability in the sample. You may remember this term from your statistics class, but let's walk through a quick review. One common estimate of variability is the **standard deviation**, which represents the average difference between individual scores and the mean of the group. It is calculated by subtracting each score from the mean, squaring the deviation, adding up these squared deviations, dividing by the sample size, and taking the square root of the result.

One problem with the standard deviation is that it generally underestimates the variability of the population, especially in small samples, because small samples are less likely to include the full range of population values. So, we need a way to correct our variability estimate in a small sample. Enter the **standard error**, which is computed by dividing the standard deviation by the square root of the sample size. (To save time, these values are already calculated and presented in Table 2.3.) The "pooled" standard error represents a combination of the standard errors from our two groups:

$$SE_{pooled} = \sqrt{SE_1 2 + SE_2 2} = \sqrt{(.33)^2 + (.33)^2} = \sqrt{.218} = .47$$

Our final step is to plug the appropriate numbers from our "children and happiness" data set into the *t*-test formula.

$$t = \frac{x_1 - x_2}{SE_{pooled}} = \frac{5.7 - 2}{.47} = \frac{3.7}{.47} = 7.87$$

If this all seems overwhelming, stop and think about what we've done in conceptual terms. The goal of our statistical test—the *t*-test—is to determine whether our groups differ by a meaningful and significant amount. The best way to do that is to examine the group difference as a ratio, relative to the overall variability in the sample. When we calculate this ratio, we get a value of 7.87, which certainly seems impressive, but there's one more step we need to take to interpret this number.

Step 4—Compare With a Critical Value

What does a 7.87 mean for our hypothesis test? To answer this question, we need to gather two more pieces of information and then look up our *t*-test value (i.e., 7.87) in a table. The first piece of information is the **alpha level**, representing the probability cutoff for our

hypothesis test. The standard alpha level to use is .05, meaning that we want to have less than a 5% chance of the result being due to chance. In some cases, you might elect to use an alpha level of .01, meaning that you would only be comfortable with a less than 1% chance of your results being due to chance.

The second piece of information we need is the **degrees of freedom** in the data set; this number represents the sample size and is calculated for a *t*-test via the formula $n - 2$, the number of couples in our sample minus 2. Think of it as a mathematical correction for the fact that we are estimating values in a sample rather than from the entire population. Another helpful way to think of degrees of freedom is as the number of values that are "free to vary." In our sample experiment, the no-children group has a mean of 5.7 while the children group has a mean of 2. Theoretically, the values for 9 of the couples in each group can be almost anything, but the 10th couple has to have a happiness score that will yield the correct overall group mean. Thus, of the 20 happiness scores in our experiment, 18 are free to vary, giving us 18 degrees of freedom (i.e., $n - 2$).

Armed with these two numbers—18 degrees of freedom and an alpha level of .05—we turn to a **critical value table**, which contains cutoff scores for our statistical tests. (You can find these values for a *t*-test at http://www.statstodo.com/TTest_Tab.php). The numbers in a critical value table represent the minimum value needed for the statistical test to be significant. In this case, with 18 degrees of freedom and an alpha level of .05, we would need a *t*-test value of 1.73 for a one-tailed (directional) hypothesis test and a *t*-test value of 2.10 for a two-tailed (nondirectional) hypothesis test. (Remember, we have to be more conservative for a nondirectional test.) In our children and happiness study, we had a clear directional/one-tailed hypothesis that children would make couples less happy, so we can legitimately use the one-tailed cutoff score of 1.73. Because our *t*-test value of 7.87 is unquestionably higher than 1.73, our statistical test is significant. In other words, there is less than a 5% chance that the difference in happiness ratings is due to chance.

Step 5—Make a Decision

Finally, we are able to draw a conclusion about our experiment. Based on the outcome of our statistical test (i.e., steps 3 and 4), we will make one of two decisions about our null hypothesis:

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

or

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

Because our *t*-test value was quite a bit higher than the required cutoff value, we can be confident in rejecting the null hypothesis. And, at long last, we can express our findings in plain English: Couples with children are less happy than couples without children!

Now that we have walked through this five-step process, it's time to let you in on a little secret. When it comes to analyzing your own data, to test your own hypotheses, you will actually rely on a computer program for part of this process—Steps 3 and 4 in particular.

In these modern times, it is rare to compute even a *t*-test by hand. Software programs such as SPSS (IBM), SAS/STAT (SAS), and Microsoft Excel can take a table of data, compute the mean difference, compare it to the variability, and calculate the probability that the results are due to chance. However, because these calculations happen behind the scenes, it is very important to understand the process. To draw conclusions about your hypotheses, you have to understand what a *p* value and a *t*-test value mean. By understanding how the software operates, you can reach informed conclusions about your research questions. Otherwise, you risk making one of two possible errors in your hypothesis test, discussed in the next section.

Errors in Hypothesis Testing

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

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

Fortunately, although this error sounds scary, we can generally compute the probability of making it. Our alpha level sets the bar for how extreme our data must be in order to reject the null hypothesis. At the end of the statistical calculation, we get a *p* value that tells us how extreme the data actually are. When we set an alpha level of, say, .05, we are attempting to avoid a Type I error; our results will only be statistically significant if the effect outweighs the random variability by a big-enough amount. If our *p* value falls below our predetermined alpha level, we decide that the risk of a Type I error is sufficiently small and can therefore reject the null hypothesis. If, however, our *p* value is greater than (or even equal to) our alpha cutoff, we decide that the risk of Type I error is too high to ignore and will therefore fail to reject the null hypothesis.

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

Fortunately, although this error sounds disappointing, there are some fairly easy ways to avoid or minimize it. The key factor in reducing Type II error is to maximize the power of the statistical test, or the probability of detecting a real difference. In fact, power is *inversely related* to the probability of a Type II error—the higher the power, the lower the chance of Type II error. Power is analogous to the sensitivity, or accuracy, of the hypothesis test; it is under the researcher's control in three main ways. First, as we discussed in the section "Reliability and Validity," it is important to make sure that your measures are capturing what you think they are. If your happiness scale actually captures something like narcissism, then this will cause problems for your hypothesis about the predictors of happiness. Second, it is important to be careful throughout the process of coding and analyzing your data. Small mistakes can occur at every step, from entering data, to calculating scale totals, to choosing an inappropriate analysis. And third, statistical tests generally have more power when there is a larger sample. We will discuss each of these factors in more detail as we move through the course.

Summary of Correct and Incorrect Decisions

In the real world, at the level of the entire population our null hypothesis is either true or false. That is, if we could test our hypothesis using every married couple in the world, we could say with 100% certainty whether or not the hypothesis was true. However, in each individual study, at the level of our sample, we have to either reject the null or fail to reject it. In the top left and bottom right cells, we make the right decision—either rejecting a null hypothesis that is false or failing to reject one that is true in the population. Table 2.4 summarizes the four possible outcomes of a decision about a hypothesis test. In the bottom left cell of the table, we make a Type I error, rejecting a null hypothesis that is actually true, and mistakenly thinking our hypothesis is supported (i.e., a false positive). In the top right cell of the table, we make a Type II error, failing to reject a null hypothesis that is actually false, and mistakenly thinking our hypothesis should be rejected (i.e., a false negative).

Table 2.4: Errors and correct decisions in hypothesis testing

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

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

Effect Size

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

There are several ways to calculate effect size, but in general, bigger values mean a stronger effect. One of these statistics, **Cohen's *d***, is calculated as the difference between two means divided by their pooled standard deviation. The resulting values can therefore be expressed in terms of standard deviations; a d of 1 indicates that the means are one standard deviation apart. How big should we expect our effects to be? Based on his analyses of typical effect sizes in the social sciences, Cohen suggests the following benchmarks: $d = .20$ is a small effect; $d = .40$ is a moderate effect; and $d = .60$ is a large effect. In other words, a large effect in social and behavioral sciences accounts for a little over half of a standard deviation. For comparison purposes, the effect of the polio vaccine on reducing polio symptoms was a $d = 2.72$ (almost three standard deviations; Oshinsky, 2006). In our children and happiness study, we get a $d = 3.82$, but fake data are always more impressive than real data.

Effect size is useful in two primary ways. First, at the end of an experiment, we can calculate the exact size of the effect in our particular sample. This is a useful supplement to our test of statistical significance because it is more independent of sample size. If we fail to reject the null hypothesis in a small sample, the effect size might tell us whether the effect is big enough to test again with a larger sample. And, if we support our research hypothesis, the effect size provides valuable information about the usefulness of our findings. Imagine you test two different diabetes drugs in two different studies. Let's say both show a statistically significant reduction in symptoms, but Drug A has an effect size of $d = .50$, and Drug B has an effect size of $d = 2.5$. This tells us that Drug B has a larger effect and could therefore have a bigger benefit for diabetes patients.

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

The best way to perform these calculations is by using any of the power calculators available over the Internet, such as the one found here: http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html. Try entering the values from our children and happiness study, plus the

pooled standard deviation of 1.25. This should result in the previously mentioned *d* of 3.82. According to this calculator, we would only need two people per group to detect this effect in a future study—much cheaper and easier than 10,000!

Summary

In this chapter, we have covered several basic principles of research design and emphasized the importance of ensuring that our study uses the best and most accurate measures available. We first examined the four main types of research design: qualitative, descriptive, correlational, and experimental. These designs increase the amount of control that a researcher has. Qualitative designs can lead to in-depth interpretations, verifications, and evaluations of phenomena, such as personal experiences, events, and behaviors. However, their drawbacks include difficulty making comparisons, difficulty making assumptions beyond the sample being studied, and being very time-consuming. Descriptive designs can provide rich descriptions of various phenomena, from brain tumors to voting preferences, but are unable to delve into why these things happen. Correlational designs allow us to predict variables from other variables but are still unable to identify a causal relationship. This limitation in correlational designs occurs for two reasons: We do not know the direction of the relationship; and it is always possible that a third variable is causing both of them. Finally, experimental designs allow us to state with some certainty that one variable causes another because these designs involve systematically testing the impact of variables while controlling the environment. The downside of experimental designs is that they often have to sacrifice some realism in order to establish control.

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

We next discussed the different options for scaling and measuring variables. In addition to ensuring the accuracy and consistency of measures, it is critical to use a scaling method that matches the mathematical properties of the variable. Nominal scales represent arbitrary labels for categories; ordinal scales represent rank ordering of values; interval scales represent scales with equal intervals; and ratio scales represent variables with true zero points. As a researcher, you should use the most powerful scale available—for example, by using behavioral counts rather than labels when possible. But you also have to be aware of the limitations of the scale that you choose. While ratio scale values can be added, subtracted, divided, and multiplied, ordinal scale values cannot be manipulated. We also discussed three primary types of measurement. Behavioral measures involve observation and systematic recording of behavior; self-report measures involve asking people to

report their own thoughts; and physiological measures involve measurements of bodily processes. Because each approach has advantages and disadvantages, many researchers use converging operations over the course of a research program, making use of all three in order to address a broad question.

Finally, this chapter discussed the process of hypothesis testing. Regardless of the question asked, the design used, and the way data are measured, all quantitative studies involve the same process of testing hypotheses using statistical results. We covered this process in five steps: (1) Lay out the null and alternative hypotheses; (2) collect data; (3) calculate the appropriate statistics; (4) compare statistical results to a critical value; and (5) make a decision about the original hypothesis. Despite our best efforts, a hypothesis test occasionally leads to incorrect conclusions. A Type I error occurs when the researcher rejects the null but shouldn't have; a Type II error occurs when the researcher fails to reject the null but could have under better conditions. As we will discuss in later chapters, you can reduce the odds of both errors through careful research design and analysis. In the next three chapters, we will cover the specifics of the three types of research design: descriptive (Chapter 3), correlational (Chapter 4), and experimental (Chapter 5).

Key Terms

alpha level Predetermined probability cutoff for a hypothesis test; usually set as $p < .05$.

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

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

Cronbach's alpha The average correlation between each pair of items on the measure; used to calculate an estimate of interitem reliability.

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

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

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

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

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

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

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

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

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

critical value table A table containing cutoff scores for statistical tests.

degrees of freedom A number representing sample size; calculated for a *t*-test via the formula $n - 2$ (the number of people in a sample minus 2 variables); the number of values that are "free to vary."

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

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

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

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

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

duration The length of time a behavior lasts.

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

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

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

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

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

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

frequency The number of times a behavior occurs.

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

intensity The strength or potency of a behavior.

interitem reliability The internal consistency among different questions on a questionnaire measure.

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

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

latency The length of delay before onset of a behavior.

negative correlation Relationship between two variables such that higher values of one variable predict lower values of the other variable.

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

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

null hypothesis The assumption that there is no real effect of variables in the overall population.

ordinal scale A scaling method used to represent ranked order of conceptual variables.

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

positive correlation Relationship between two variables such that higher values of one variable predict higher values of the other variable.

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

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

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

random error Chance fluctuations in the measurements.

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

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

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

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

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

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

standard error An estimate of variability that is computed by dividing the standard deviation by the square root of the sample size.

Student's *t*-test An inferential statistic used to compare differences in the means of two groups; calculated as a ratio of mean difference to average variability.

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

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

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

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

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

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

Apply Your Knowledge

1. For each of the following research questions, tell whether the most appropriate strategy involves descriptive, correlational, or experimental research.
   a. Are students more likely to cheat on exams in their first or last year of college?
   b. Does writing about a traumatic experience result in better health?
   c. What personality variables predict success in school?

2. Dr. Blutarsky is interested in predicting the link between poor parenting and teen alcohol abuse. To investigate this, he has parents fill out questionnaires about their parenting styles and then waits to see how likely their children are to abuse alcohol.
   a. Identify the independent and dependent variables in this study.
      independent:
      dependent:
   b. What type of research design is Dr. Blutarsky using?
   c. Give an operational definition of "poor parenting" and "alcohol abuse"
      poor parenting:
      alcohol abuse:

3. For each of the following, identify the scale of measurement:
   a. placing children in gifted and special needs programs based on ability
   b. an "attitudes toward the president" scale, measured from 1 to 7
   c. height measured in inches
   d. the number of drinks consumed per day

4. For each of the following abstract concepts, suggest a way to measure it using a behavioral and self-report measure:

	Behavioral	Self-Report
Conformity		
Enjoyment of Reading		
Leadership Ability		
Paranoia		
Independence		

Critical Thinking & Discussion Questions

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