LEARNING OBJECTIVES

Upon completion of this chapter, the reader should be able to:

2.1 Explain the hypothesis-testing approach and how investigators use experimental variables in their research.

2.2 Describe the case study method and some of its strengths and weaknesses.

2.3 Discuss the reasons why researchers use statistical analysis of data and why personality psychologists sometimes examine correlation coefficients.

2.4 Explain how researchers use indicators of reliability and validity to determine the usefulness of a test.

Not long ago, “Desperate in Dallas” wrote to a newspaper advice columnist about her husband’s 16-year-old cousin, who was living with them. The boy didn’t want to work, didn’t want to go to school, and generally was a very messy houseguest. What was she to do? The columnist explained to “Desperate” that the boy’s real problem was the rejection he had received from his parents earlier in his life, and these early childhood experiences were responsible for the boy’s lack of motivation. Within the next few weeks, the adviser also explained to “Wondering in Boston” that a 5-year-old boy became aggressive from watching too many violent programs on television. She told “Anonymous in Houston” that her 5-year-old daughter was going to be a leader, and “Intrigued in Norfolk” that, although some people are routinely incapacitated with minor aches and pains, others are capable of ignoring them.

Millions of readers seem to think this columnist has something to say about the causes and consequences of people’s behavior. But how does she know? Experience? Intelligence? A keen insight into human nature? Perhaps. In a way, advice columnists represent one avenue for understanding personality—through expert opinion. We could say the columnist is similar to some of the great personality theorists who studied the works of others, made their own observations, and then described what they saw as the causes of human behavior. As you will see in Chapter 3, Sigmund Freud proposed many groundbreaking ideas about personality. He read widely about what his contemporaries
were saying, worked and consulted with some of the great thinkers of the day, and care-
fully observed his patients. From information gathered through all these avenues, Freud
developed a theory of personality he spent the rest of his career promoting.

Although more scholarly than a columnist’s one-paragraph diagnosis, Freud’s writ-
ings often evoke a similar response: How does he know? Freud’s ideas are intriguing
and his arguments at times persuasive, but most personality psychologists want more
than an expert’s viewpoint before they accept the validity of a theory. They want empirical
research. They want studies examining key predictions from the theory. They want some
hard numbers to support those predictions. This is not because an expert’s ideas are of
no value. Quite the contrary, the views and observations of personality theorists form the
backbone of this book. But theories alone provide only part of the picture. Understanding
the nature of human personality also requires an examination of what psychologists
have learned from rigorous empirical investigations.

This chapter presents a brief introduction to psychological research with an empha-
sis on issues particularly relevant for personality. We begin with a description of some
basic concepts associated with the hypothesis-testing approach to research. Next, we
look at a procedure that has played a significant role in the history of personality psychol-
ogy—the case study method. We then touch briefly on what you will need to know about
statistical analysis of data. Finally, because personality psychologists often rely on per-
sonality assessment, we quickly review some of the concepts associated with measuring
individual differences in personality.

THE HYPOTHESIS-TESTING APPROACH

LO 2.1 Explain the hypothesis-testing approach and how investigators use experimental
variables in their research.

Each of us on occasion speculates about the nature of personality. You may have wondered
why you seem to be more self-conscious than other people, why a family member is depressed
so often, or why you have so much trouble making friends. In the latter case, you may have
observed the way popular students interact with the people they meet and compared their
behavior with the way you act around strangers. You may have even altered your behavior to be
more like theirs and then watched to see if this change affected how people react to you.

In essence, the difference between this process and that used by personality psychologists
lies only in the degree of sophistication. Like all of us, these researchers speculate about the
nature of personality. From observations, knowledge about previous theory and research, and
careful speculation, they generate hypotheses about why certain people behave the way they do.
Then, using experimental methods, they collect data to see if their explanations about human
behavior are correct. Like pieces in a large jigsaw puzzle, each study makes another contribution
to our understanding of personality. However, by the time you get to the end of this book, it
should be clear this is one puzzle that will never be finished.
Theories and Hypotheses

Most personality research begins with a theory—a general statement about the relationship between constructs or events. Theories differ in the range of events or phenomena they explain. Some, such as the major personality theories discussed in this book, are very broad. Psychologists have used Freud’s psychoanalytic theory to explain topics as diverse as what causes psychological disorders, why people turn to religion, and why certain jokes are funny. However, personality researchers typically work with theories considerably narrower in application. For example, they might speculate about the reasons some people are more motivated to achieve than others or about the relationship between a parent’s behavior and a child’s level of self-esteem. It might be useful to think of the larger theories, such as Freud’s, as collections of more specific theories that share certain assumptions about the nature of human personality.

A good theory possesses at least two characteristics. First, a good theory is parsimonious. Scientists generally operate under the “law of parsimony”—that is, the simplest theory that can explain the phenomenon is the best. As you will see throughout this book, several theories can be generated to explain any one behavior. Some can be quite extensive, including many concepts and assumptions, whereas others explain the phenomenon in relatively simple terms. Which theory is better? Although it sometimes seems that scientists enjoy wrapping their work in fancy terms and esoteric concepts, the truth is if two theories can account for an effect equally well, the simpler explanation is preferred.

Second, a good theory is useful. More specifically, unless a theory can generate testable hypotheses, it will be of little or no use to scientists. Ideas that cannot be tested are not necessarily incorrect. It’s just they do not lend themselves to scientific investigation. For example, throughout history some people have explained psychological disorders in terms of invisible demons taking over a person’s body. This may or may not be a correct statement about the cause of disorders. But unless this explanation is somehow testable, the theory cannot be examined through scientific methods and therefore holds little value for scientists.

However, theories themselves are never tested. Instead, investigators derive from the theory hypotheses that can then be tested in research. A hypothesis is a formal prediction about the relationship between two or more variables that is logically derived from the theory. For example, many psychologists are interested in individual differences in loneliness (Chapter 12). That is, they want to know why some people frequently suffer from feelings of loneliness, whereas others rarely feel lonely. One theory proposes that lonely people lack the social skills necessary to develop and maintain satisfying relationships. Because this is a useful theory, many predictions can be derived from it, as shown in Figure 2.1. For example, if the theory correctly describes a cause of loneliness, we might expect consistently lonely people to make fewer attempts to initiate conversations than those who are not lonely. Another prediction might be that these lonely people have a poor idea of how they are being perceived by others. Yet another prediction might maintain that lonely people make more socially inappropriate statements than nonlonely people during conversations.
Each of these predictions can be tested. For example, we might test the last prediction by recording conversations lonely and nonlonely people have with new acquaintances. Judges could evaluate the conversations in terms of number of appropriate and inappropriate responses each participant makes. If people who identify themselves as lonely make fewer appropriate responses during the conversation, the prediction is confirmed. We then say we have support for the theory. But notice the theory itself is not tested directly. In fact, theories are never proved or disproved. Rather, a theory is more or less supported by the research and therefore is more or less useful to scientists trying to understand the phenomenon. The more often research confirms a prediction derived from a theory, the more faith psychologists have that the theory accurately describes the nature of things. However, if empirical investigations consistently fail to confirm predictions, we are much less likely to accept the theory. In these cases, scientists typically generate a new theory or modify the old one to better account for the research findings.
Experimental Variables

Good research progresses from theory to hypotheses to experiments. The basic elements of an experiment are the experimental variables, which are divided into two types: independent variables and dependent variables. An independent variable determines how the groups in the experiment are divided. Often this is manipulated by the experimenter, such as when participants are randomly assigned to different experimental conditions. An independent variable might be the amount of a drug each group receives, how much anxiety is created in each group, or the type of story each group reads. For example, if level of anxiety is the independent variable, a researcher might tell Group A they will give a speech in front of a dozen critical people, Group B they will give a speech in front of a few supportive people, and Group C nothing about a speech. Because each of the groups created by the independent variable receives a slightly different treatment, researchers sometimes refer to the independent variable as the treatment variable.

A dependent variable is measured by the investigator and used to compare the experimental groups. In a well-designed study, differences among groups on the dependent variable can be attributed to the different levels of the independent variable. Returning to the anxiety example, suppose the researcher’s hypothesis was that people reduce anxiety about upcoming events by obtaining as much information about the situation as possible. The researcher might use level of anxiety as the independent variable, creating high-, moderate-, and low-anxiety conditions. The three groups might be compared on how many questions they ask the experimenter about the upcoming event. In this case, the number of questions is the dependent variable. The results of such an experiment might look like this:

<table>
<thead>
<tr>
<th></th>
<th>High Anxiety</th>
<th>Moderate Anxiety</th>
<th>Low Anxiety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of questions</td>
<td>5.44</td>
<td>3.12</td>
<td>1.88</td>
</tr>
</tbody>
</table>

If the experiment has been designed correctly, the investigator will attribute the difference in the dependent variable (the number of questions) to the different levels of the independent variable (anxiety). Because experimenters want to say that differences in the dependent variable are the result of the different treatment each of the groups received, some researchers refer to the dependent variable as the outcome variable.

However, most personality research is more elaborate than this example indicates. Researchers typically use more than one independent variable. In the information-seeking example, an experimenter might want to further divide participants into groups according to how extraverted or introverted they typically are. The researcher might predict that anxiety leads to a search for information, but only among people who are extraverted. Researchers in this hypothetical study might use two independent variables to divide participants into groups. They might randomly assign participants to either an anxiety (anticipates speech) or a no-anxiety group, and within each of these groups identify those who are extraverted and those who are introverted. If the dependent variable remains the number of questions asked of the experimenter, the results might turn out like those shown in Figure 2.2.
Manipulated Versus Nonmanipulated Independent Variables

Sometimes personality researchers randomly assign participants to conditions. However, other times they simply identify which group the participant already belongs to, such as whether the person is extraverted or introverted. The significance of this difference is illustrated in the following example.

Suppose you are interested in the effect violent television programs have on the amount of aggression children display in real life. You recruit two kinds of elementary school students—those who watch a lot of violent TV shows and those who watch relatively few. You then look at the school records to see how often each child was disciplined for aggressive acts, such as threatening other children, pushing, or fighting. Consistent with your hypothesis, you find that children who watch a lot of violent television are more aggressive than those who watch relatively little violent TV. You might be tempted to conclude that watching violent television programs causes children to be more aggressive. However, based on this study alone, your conclusion must be tempered. For example, it’s possible that some children watch violent TV shows precisely because they are aggressive. Perhaps they are more entertained by programs that include shootings, stabbings, and other violent acts. Thus, although consistent with the hypothesis, it is difficult to make statements about cause and effect from these findings.

This example illustrates the fundamental difference between research using manipulated independent variables and research using nonmanipulated independent variables. An investigator who uses a manipulated independent variable begins with a large number of participants and randomly assigns them to experimental groups. That is, each person has an equally likely chance of being assigned to Condition A as to Condition B (or C, or D, and so on). Investigators know all participants are not exactly alike at the beginning of the study. Some are naturally
more aggressive than others, some more anxious, some more intelligent. Each has different life experiences that might affect what they do in the study. However, by using many participants and randomly assigning them to conditions, researchers assume all these differences will even out. Although within any given condition we will find participants who are typically high or low in aggressiveness, each of the conditions should have the same average level of aggressiveness at the beginning of the experiment.

The researcher then introduces the independent variable. For example, one group might be shown 30 minutes of violent television programming, another group might watch a basketball game, and still another group might sit quietly and watch no television. Because we assume participants in each condition are nearly identical on average at the start of the study, any differences among the groups after watching the program can be attributed to the independent variable. That is, if participants who watched the violent TV shows are more aggressive afterward than those who watched the nonviolent shows or those who watched no TV, we can conclude with reasonable confidence that watching the violent TV shows likely caused the participants to act more aggressively.

This procedure contrasts with one that uses nonmanipulated variables. A nonmanipulated independent variable (sometimes referred to as a subject variable) exists without the researcher’s intervention. For example, researchers might divide people into high self-esteem and low self-esteem groups, or into first-born, middle-born, or last-born categories. In these cases, the investigator does not randomly assign participants to a condition. Returning to the earlier example, the researcher who compared frequent and infrequent television viewers did not manipulate participants into those two categories. Rather, each participant already belonged to one of the groups, and the researcher simply had to determine which group that was.

The problem researchers face when using nonmanipulated independent variables is they cannot assume their conditions are on average nearly identical at the beginning of the experiment. Children who watch little violent television might be different from those who watch a lot in many ways. They might differ in terms of self-esteem, parental involvement, and, most notably, their level of aggression prior to participating in the experiment. Children not watching television probably have more time for other activities, such as playing with friends and thereby learning how to get along with others. In short, any differences we find between the two groups could be caused by any of these differences and not necessarily by the number of violent TV shows each group watches.

Because it is difficult to determine cause-and-effect relationships with nonmanipulated independent variables, researchers generally prefer to manipulate variables. However, doing so is not always possible. Sometimes manipulating the variable is too expensive, too difficult, unethical, or simply impossible. This is not to say that research with nonmanipulated independent variables is useless. On the contrary, personality psychologists often find that relying on nonmanipulated variables is the only way to examine many of the topics that interest us. How else can we study differences between introverts and extraverts or differences between men and women? In fact, a survey of academic journals found the vast majority of personality research relies on nonmanipulated independent variables (Revelle & Oehlberg, 2008). Nonetheless, investigators who conduct this research must remain cautious when making statements about cause and effect.
Prediction Versus Hindsight

Which person is more impressive: the one who can explain after a football game why the winning team was victorious, or the one who accurately tells you before the game which team will win and why? Most of us are more impressed with the second individual. After all, anyone can come up with an explanation after the facts are in. But people who really understand the game can make reasonable guesses about what will happen when two teams meet.

In a similar manner, we expect researchers to make predictions about what will happen in a study before the data are collected. Remember, the purpose of research is to provide support for a hypothesis. Suppose a researcher examines the relationship between self-esteem and helping behavior, but the investigator has no clear prediction beforehand of what this relationship might be. If the study finds that high self-esteem people help more than low self-esteem people, the researcher might conclude this is because people who feel good about themselves know they can maintain that positive evaluation by doing good things. The explanation sounds reasonable, but do the data support the hypothesis? From a scientific standpoint, the answer is “No,” because the hypothesis was generated after the results were seen. With that sequence, there is no way the hypothesis would not be supported. If the study found that low self-esteem people help more, the same researcher might conclude this is because these individuals are trying to improve their self-image by doing good things. With no possibility that the hypothesis might not be supported, the hypothesis has not really been tested. This is not to say researchers should ignore findings they haven’t predicted. On the contrary, such findings are often the basis for future hypotheses and further research. But explaining everything after the results are in explains nothing.

Replication

When investigators conduct a well-designed study and uncover interesting results, they usually report the findings in a journal or perhaps at a professional conference. Sometimes the findings are cited in popular media as something psychology now knows about the topic. However, most psychologists are more cautious about drawing conclusions from a single investigation. Sometimes effects found in one study are not reproduced when researchers conduct a second study using similar procedures (Klein et al., 2018).

There are many reasons why a researcher might find an effect in a given study. There could be something peculiar about the people in the sample. There might be something special about the time the research was conducted—perhaps an unusual mood in the country or on campus, caused by an important event. Or the finding could be the result of some unknown and inadvertent aspect of the particular experimental procedure. There is also the ever-present possibility the findings were just a fluke; that is, the outcome occurred because of some random and unknown set of coincidences. Whatever the reason, it is dangerous to assume a significant finding from one study provides reliable evidence for an effect.

The way to deal with this problem is through replication. The more often an effect is found in research, the more confidence we have that it reflects a genuine relationship. Replications often examine participants who come from different populations than used in the original
research. This procedure helps to determine whether the effect applies to a larger number of people or is limited to the kind of individuals used in the initial study. Yet determining the strength of an effect by how often it is replicated is not always easy. One difficulty has been called the “File Drawer” problem. That is, investigators tend to publish and report research only when they find significant effects. When an attempt at replication fails, the researcher may decide something has gone wrong—perhaps the wrong materials were used, perhaps something was not done the way the original researcher did it, and so on. And so, the research is stored away in a file drawer and never reported. The result is that a well-known research finding may, in fact, be difficult to replicate. But because the failures at replication are stored away in file drawers, we might not realize the problem exists. To address this problem, some psychologists are now creating online archives and developing other methods to make failures to replicate more accessible.

THE CASE STUDY METHOD

LO 2.2 Describe the case study method and some of its strengths and weaknesses.

Like a carpenter or a physician, personality researchers must use many different tools to be effective in their job. Although most personality psychologists rely on empirical studies with large numbers of participants to test their ideas, there are other ways to examine individual differences and personality processes. One procedure occasionally used by personality researchers is the case study method, an in-depth evaluation of a single individual (or sometimes a few individuals). Often the focus of a case study is a particularly interesting psychotherapy client. Occasionally, researchers conduct an extensive psychological analysis of a unique individual they may have never met, a method known as psychobiography. Psychologists have written psychobiographies about Leonardo da Vinci, John Lennon, and George W. Bush, among others (Kasser, 2017; Kovary, 2011; Schultz & Lawrence, 2017).

Researchers conducting a case study record in great detail the person’s history, behavior, and changes in behavior over the course of the investigation, which sometimes lasts for years. Case study data are usually descriptive. That is, rather than reporting a lot of numbers and statistical analyses, investigators describe their impressions of what the person did and what the behavior means. Researchers occasionally include quantitative assessments, such as recording how many times the person washes their hands in a 24-hour period. However, data comparing the individual with another group or another person are rarely reported.

As you will see throughout this book, case studies have played an important role in the history of personality psychology. Sigmund Freud relied almost exclusively on his own in-depth analysis of patients when formulating ideas about personality. In fact, many of Freud’s initial insights into the functions of the human mind came from his observations of one early patient, Anna O., whose story is told in Chapter 3. Gordon Allport, the first psychologist to promote the concept of traits, argued that we cannot capture the essence of a whole personality without an in-depth analysis of a single individual. Humanistic theorists, most notably Carl Rogers, developed their unique concept of human nature through the extensive evaluation of psychotherapy clients. Behaviorists also sometimes rely on case studies to illustrate various aspects of their
theories and the effectiveness of their therapies. For example, we will review John B. Watson’s work with an infant named “Little Albert” in Chapter 13. This famous case study has been widely cited as evidence for the behaviorist explanation of abnormal behaviors.

### STATISTICAL ANALYSIS OF DATA

**L0 2.3 Discuss the reasons why researchers use statistical analysis of data and why personality psychologists sometimes examine correlation coefficients.**

Suppose a restaurant server wants to know, for obvious reasons, what kind of behavior elicits the largest tips from customers. Her hypothesis is smiling and acting in a friendly manner will result in better tips than acting in a more professional and reserved manner. She tests her hypothesis by alternating between the friendly and professional styles each night for 14 nights. At the end of each evening, she counts her tips and records the data. Suppose these are her findings:

<table>
<thead>
<tr>
<th>Friendly Approach</th>
<th>Professional Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>$51.50</td>
<td>$56.90</td>
</tr>
<tr>
<td>62.75</td>
<td>51.75</td>
</tr>
<tr>
<td>59.60</td>
<td>58.00</td>
</tr>
<tr>
<td>52.00</td>
<td>52.25</td>
</tr>
<tr>
<td>61.10</td>
<td>53.60</td>
</tr>
<tr>
<td>49.45</td>
<td>59.30</td>
</tr>
<tr>
<td>50.20</td>
<td>50.60</td>
</tr>
<tr>
<td><strong>$55.23 average</strong></td>
<td><strong>$54.63 average</strong></td>
</tr>
</tbody>
</table>

The server concludes the friendly approach indeed works best, and changes to a friendly style from then on. But is this conclusion justified? We can see from the numbers that the friendly style came up with a higher average tip than the professional style. But by now you probably have already wondered if an average of $55.23 is reliably different from an average of $54.63. Because of naturally occurring variation in the amount of money made each evening, we would not expect the averages to come out exactly the same, even if the server never changed her style. One condition in this study would almost always come out at least a little higher than the other. And so, the question becomes: How much higher must one of the averages be before we conclude the difference is not just a chance fluctuation but in fact represents a real difference between the two styles of serving customers? This is the question of statistical significance.

### Statistical Significance

How can researchers tell if different group averages represent real effects or just chance fluctuations? Fortunately, statisticians have developed formulas allowing us to estimate the likelihood that the difference between the averages could have occurred by chance alone. There are
many types of statistical tests, each appropriate for different types of data and different research designs. Some of the more common tests are an analysis of variance, a chi-square test, and a correlation coefficient.

Returning to the server example, if the two averages differed by an amount so small that it could have been caused by a chance fluctuation, we say the difference has not reached statistical significance. Conversely, if the difference is so large that in all likelihood it was not caused by chance but reflects a true difference between the two service styles, we say the difference is statistically significant. In the latter case, the conclusion would be that one style does seem to result in better tips than the other.

However, statistical tests do not really provide a yes or no answer to our question. All they tell us is the statistical likelihood that the difference between the groups was caused by chance. But this observation raises another question. Suppose you apply a statistical test to the server's data and find that a difference this large would occur by chance only one out of every three times you conducted the study. What would you conclude? That the different averages represent a real effect? It would be difficult to have much confidence in such a statement. The difference might be real, but there is a reasonable likelihood the finding is just a fluke. So, when can we say we have a real difference? Traditionally, psychologists use a significance level of .05 to answer this question. This means if the difference between the scores is so large that it would occur less than 5% of the time by chance, the difference is probably genuine. However, referring back to our earlier discussion, most psychologists are hesitant to put too much stock in a statistically significant finding until the effect is found to be replicable.

Also keep in mind statistically significant results are not necessarily “significant” in all ways. When researchers use a large number of participants, even small differences can be statistically significant. Whether the difference is large enough to be important is another question. In response to this concern, investigators often examine and report the size of the effect through statistical values known as effect size indicators.

**Correlation Coefficients**

The correlation coefficient is a favorite statistic among personality researchers and one that will pop up from time to time in this book. It is the appropriate statistical test when we want to understand the relationship between two measures. For example, suppose you are interested in the relationship between loneliness and depression. You might ask a large number of people to complete a loneliness scale as well as a depression inventory. If loneliness and depression are related, we would expect people who score high on loneliness to also score high on depression. Similarly, those who score low on loneliness should score low on depression.

Figure 2.3 presents three possible outcomes for this study. Each dot in the figure represents one participant’s scores on both scales. The first outcome indicates a person’s score on one scale is a fairly good predictor of that person’s score on the other scale. In this case, if we know someone is high on loneliness, we know that person is probably going to score high on depression as well. The second outcome indicates little or no relationship between the measures. Knowing a person’s score on one scale tells us nothing about what the other score will be. The third
outcome, like the first, indicates that knowing a person’s loneliness score will help predict the depression score but not in the way we might have anticipated. Here, a high score on one measure predicts a low score on the other.

**FIGURE 2.3 Three Possible Relations Between Loneliness and Depression**

![Diagram showing three possible relations between loneliness and depression](image)

After conducting the appropriate statistical test, we can reduce the data from our study to a single number called a correlation coefficient. This number can range from 1.00 to -1.00. Returning to the figure, the first outcome indicates a fairly strong relationship between loneliness and depression. The correlation coefficient for this figure might be .60. Because a high score on one measure indicates a high score on the other measure, this is a positive correlation. For the second outcome, the correlation coefficient is close to .00, indicating no relationship between the measures. The third outcome might yield a correlation of -.60, also a fairly strong relationship between the variables. This last outcome is referred to as a negative correlation, but this does not mean it is less important than a positive correlation of the same magnitude. For example, if we had compared scores on a loneliness scale with scores on a sociability measure, we probably would have anticipated that a high score on one test would predict a low score on the other.

**PERSONALITY ASSESSMENT**

**LO 2.4 Explain how researchers use indicators of reliability and validity to determine the usefulness of a test.**

Sometimes our culture seems obsessed with measuring personality. Popular magazines and web pages often promote short tests, or “quizzes,” to measure how good a roommate you are, what type of romantic partner you need, or the vacation spot that matches your personality. Although the creators rarely claim their tests are based on scientific investigations, the popularity of these tests suggests readers find them at least interesting if not believable. There is something about calculating a score that gives credibility to an untested 10-item quiz.

Psychologists take the task of measuring personality much more seriously. Valid and reliable assessment is a key component of personality research. If we are going to study achievement motivation, self-esteem, social anxiety, and so on, we need to measure these concepts as accurately as possible. Similarly, psychologists working in education, human resources, and
counseling often rely on psychological tests to determine whether a child should be placed in a special class, whether an employee should be promoted to a new position, or whether a client needs professional counseling.

In each case, it is the responsibility of the people using the test to see that it accurately measures the concept they are interested in. Unfortunately, not all personality tests are as good as most psychologists prefer, and even the best tests can be used inappropriately. So how can we tell a good test from a bad one or determine whether the test measures what we want to measure? Before using any standardized test, we need to examine its reliability and validity.

**Reliability**

Suppose you took a personality test today and it indicated, compared to others your age, you scored high on *independence*. That is, more than most people your age, you enjoy being on your own and making your own decisions. However, suppose next week you take the test again, and this time your score indicates you are relatively low on independence. Which of these scores reflects your true personality? Unfortunately, you have no way to know from this test whether you are an independent or a dependent person. The test suffers from poor reliability.

A test has good *reliability* when it measures consistently. One indication of a test’s reliability is how consistent the scores are over time. That is, a reliable personality test will produce roughly the same score each time you take it. Many factors can lead to small changes in test scores, but because we assume personality is relatively consistent over time, tests designed to measure personality should provide similar scores from one testing session to another.

A common way to determine a test’s consistency is to calculate a *test–retest reliability coefficient*. Researchers first administer the test to a large group of participants. At some later time, usually after a few weeks, the same people take the test again. The investigators then calculate the correlation coefficient between the two scores. That is, how well can we predict a person’s score on the second administration if we know the person’s score from the first administration? Recall that correlation coefficients can range from 1.00 to -1.00. A high positive correlation coefficient indicates good consistency over time.

Unfortunately, whether a test is reliable is not a simple yes-or-no question. On one hand, a test–retest coefficient of .90 is probably reliable enough to meet most people’s needs. On the other hand, a reliability coefficient of .20 is no doubt too low for most purposes. But what about something in between? Is a test with a reliability coefficient of .50 or .60 acceptable? The answer depends on the researcher’s needs and the availability of more reliable tests. Sometimes the nature of the concept being measured contributes to low reliability. For example, tests given to young children, who often fluctuate in mood and attention, frequently have lower than desirable levels of reliability.

**Validity**

Reliability data tell us whether a test is measuring something consistently. But they tell us nothing about *what* the test is measuring. That is why psychologists also examine data concerning the test’s validity. *Validity* refers to the extent to which a test measures what it is designed to
Personality measure. As with reliability, the question is not whether a test does or does not have validity. Rather, the question is how well the validity of the test has been demonstrated.

Unfortunately, demonstrating validity for most personality tests is not an easy task. Personality psychologists typically are interested in measuring hypothetical constructs, such as intelligence, masculinity, or social anxiety. Hypothetical constructs are useful inventions researchers employ to describe concepts that have no physical reality. That is, we cannot see a person’s intelligence. We can observe behavior that suggests high intellectual functioning, but intelligence remains a theoretical entity.

The problem for personality researchers, then, is how to demonstrate a test measures something that, in reality, is but a useful abstract invention. How do we know if a test measures self-esteem? Test takers who agree with the item, “I am not as competent in sports as most people,” might have low self-esteem. Then again, they might just have poor athletic ability, or they might be depressed. Fortunately, there are several ways personality researchers can determine the validity of their tests. They can look for evidence of a test’s face validity, congruent validity, discriminant validity, and behavioral validation.

**Face Validity**

Perhaps the most obvious way to decide whether a test measures what it says it measures is to look at the test items. Most of us would accept that a test asking people “Do you feel nervous interacting with others?” or “Are you uncomfortable meeting new people?” is probably measuring something like social anxiety. The test would have good face validity. That is, on the face of it, the test appears to be measuring social anxiety.

**Congruent Validity**

However, face validity alone provides only weak evidence for the validity of a personality test. A good test maker should also supply information about the test’s congruent validity. The congruent validity of a test, sometimes called convergent validity, is the extent to which scores from the test correlate with other measures of the same construct. If two tests measure the same thing, scores from the two tests should be highly correlated. Imagine you developed a new intelligence test that takes less time to administer than most commonly used tests. You’d probably want to see how scores people receive from your test compare with their scores on an established intelligence test. If the scores are highly correlated, you have evidence for your test’s congruent validity. But suppose you found a correlation of only .20 between the two test scores. In this case, a person could attain a high score on one intelligence test and a low score on the other, leaving you to wonder which is the true measure of intelligence. This is not to say the old scale is measuring intelligence and your scale is not, but the low correlation would mean the two tests are not measuring the same construct.

**Discriminant Validity**

In contrast to congruent validity, a test’s discriminant validity refers to the extent to which its scores do not correlate with the scores of theoretically unrelated measures. This time, let’s imagine you have designed a creativity test. It is important to show your test measures creativity
instead of something that only resembles creativity, such as intelligence. To establish discriminant validity, you might give both the creativity test and an intelligence test to a group of people. If the two test scores are highly correlated, someone could argue your creativity test does not measure creativity at all, but simply intelligence. However, if the correlation between the two tests is low, you have evidence that the two tests measure different constructs.

**Behavioral Validation**

Finally, we come to *behavioral validation*, that is, evidence test scores can predict behavior. To demonstrate the validity of an assertiveness scale, you might use scores from that test to predict how people respond when they receive poor service at a restaurant. Naturally, you would expect highly assertive individuals to complain about the service and people low in assertiveness to tolerate the inconvenience. But what if the test scores were completely unrelated to the behavior? What if people with low scores on the scale acted just as assertively as those with high scores? It is possible test takers respond to items on the scale by indicating how they wish they would act or how they believe they are supposed to act rather than how they probably would act. In other words, it is possible to demonstrate a test’s face validity, congruent validity, and discriminant validity, and yet still not have a valid measure. If test scores cannot predict behavior, the usefulness of the test must be questioned. On the other hand, if the test does a good job of predicting how people act in relevant situations, we have strong evidence for the test’s validity.

**SUMMARY**

**LO 2.1** Explain the hypothesis-testing approach and how investigators use experimental variables in their research.

Personality psychologists examine personality processes through scientific research. Most of this research is based on the hypothesis-testing approach in which hypotheses are derived from theories. Hypotheses are then tested in studies, and the theory either is or is not supported. A good theory is parsimonious and capable of generating many testable hypotheses.

The basic elements of an experimental research design are the independent and dependent variables. One important distinction in personality research concerns whether independent variables are manipulated by the researcher. When researchers examine nonmanipulated variables, they have less confidence when making statements about cause and effect. Predicted results are better than those explained in hindsight because the latter approach does not allow for hypothesis testing. Researchers need to replicate their findings, but obtaining reliable information about how often an effect is replicated is not always easy.

**LO 2.2** Describe the case study method and some of its strengths and weaknesses.

A case study usually describes a single individual’s history, behavior, and changes in behavior over the course of the investigation. Personality psychologists sometimes rely on case studies to get an in-depth understanding of an individual
and to generate hypotheses. A case study is particularly useful when examining a rare case, when psychologists can argue that the individual being studied is similar to all normal people on the dimension of interest, when illustrating a treatment, and when demonstrating possibilities. Weaknesses of the method include questionable generalizability from one individual to other people, difficulty determining cause–effect relationships, and the questionable accuracy of some observations.

LO 2.3 Discuss the reasons why researchers use statistical analysis of data and why personality psychologists sometimes examine correlation coefficients.

Researchers use statistical tests to determine whether the differences they find between groups are the result of chance fluctuations or whether they represent genuine effects. Personality researchers often use correlation coefficients when analyzing their data. A correlation coefficient identifies the direction and size of a relationship between two measures.

LO 2.4 Explain how researchers use indicators of reliability and validity to determine the usefulness of a test.

Psychologists frequently use personality tests in their work. To determine the usefulness of a test, researchers look at evidence for the test's reliability and validity. Reliability can be gauged through test–retest correlations. Validity is determined through face validity, congruent validity, discriminant validity, and behavioral validation. Researchers must make subjective judgments when deciding whether tests are reliable and valid enough for their needs.

**KEY TERMS**

- case study method (p. 27)
- correlation coefficient (p. 29)
- dependent variable (p. 23)
- hypothesis (p. 21)
- independent variable (p. 23)
- manipulated independent variable (p. 24)
- nonmanipulated independent variable (p. 25)
- reliability (p. 31)
- statistical significance (p. 29)
- theory (p. 21)
- validity (p. 31)

**ACTIVE LEARNING EXERCISES**

**Exercise 2.1: The Hypothesis-Testing Approach**

A researcher hypothesizes that depression is caused by low activity or low engagement with the world. Think of three predictions derived from this hypothesis. For one of your predictions, think of how you might test the hypothesis.
Exercise 2.2: Independent and Dependent Variables

Identify the independent and dependent variables in each of the following studies. Also indicate whether the independent variable is manipulated or nonmanipulated.

1. A researcher compares academic achievement in three groups of people: those who are first-born, middle-born, and last-born.
2. A researcher randomly assigns participants to two conditions. One group of participants is guided through a relaxing visual imagery exercise (going to the beach) for 10 minutes. The second group is told to sit for 10 minutes thinking about whatever they wish to think about. Next, all participants complete questionnaires to measure their level of stress.

Exercise 2.3: Reliability and Validity

For each scenario, indicate whether the researcher is trying to measure the reliability or the validity of a test.

1. A group of participants takes both a newly developed friendliness questionnaire and an established, commonly used friendliness questionnaire. The researcher correlates the two sets of scores and obtains a correlation coefficient of .68.
2. A researcher gives a newly developed measure of general well-being to a group of participants. The researcher gives the same measure to the same group of participants 6 months later. The correlation coefficient between the two sets of scores is .59.
3. A team of researchers is developing a measure of general mental health. They administer the measure to college students. The researchers also ask the students to indicate the number of days this past week in which they felt anxious. The two measures are negatively correlated, with a correlation coefficient of -.47.