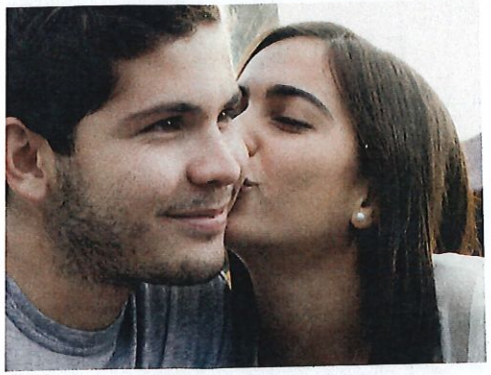
s the Happiest
ry in the World?
019



tude Is for Lovers
GOOD, 2013

Can Money Buy You Happiness
WALL STREET JOURNAL, 2014

5

Identifying Good Measurement

WHETHER STUDYING THE NUMBER of polar bears left in the Arctic Circle, the infection rate of the COVID-19 virus, the number of steps people take each day, or the level of human happiness, every scientist faces the challenge of measurement. When researchers test theories, they have to systematically measure phenomena by collecting data. These measurements must be good ones—or else they are useless.

Measurement in psychological research can be particularly challenging. Many of the phenomena psychologists are interested in—motivation, emotion, thinking, reasoning—are difficult to measure directly. Happiness, the topic of much research, is a good example of a construct that could be hard to assess. Is it really possible to quantify how happy people are? Are the measurements accurate? Before testing, for example, which country is the happiest, we might ask whether we can really measure happiness. Maybe people misrepresent their level of well-being, or maybe people aren't aware of how happy they are. How do we evaluate who is truly happy and who isn't? This chapter explains how to ask questions about the quality of a study's measures—the construct validity of quantifications of things like happiness, gratitude, or wealth. Construct validity, remember, refers to how well a study's variables are measured or manipulated.

Construct validity is a crucial piece of any psychological research study—for frequency, association, or causal claims. This chapter focuses on the construct validity of *measured variables*. You will learn, first, about different ways researchers operationalize

LEARNING OBJECTIVES

A year from now, you should still be able to:

1. Interrogate the construct validity of a study's measured variables.
2. Describe the kinds of evidence that support the construct validity of a measured variable.

» ew of measured ilated variables, ter 3, pp. 56–57.

measured variables. Then you'll learn how you can evaluate the reliability and validity of those measurements. The construct validity of *manipulated variables* is covered in Chapter 10.

WAYS TO MEASURE VARIABLES

The process of measuring variables involves some key decisions. As researchers decide how they should operationalize each variable in a study, they choose among three common types of measures: self-report, observational, and physiological. They also decide on the most appropriate scale of measurement for each variable they plan to investigate.

More About Conceptual and Operational Variables

In Chapter 3, you learned about operationalization, the process of turning a construct of interest into a measured or manipulated variable. Much psychological research requires two definitions of each variable. The **conceptual definition**, or construct, is the researcher's definition of the variable in question at a theoretical level. The operational definition represents a researcher's specific decision about how to measure or manipulate the conceptual variable.

OPERATIONALIZING "HAPPINESS"

Let's take the variable "happiness," for example. One research team, led by noted psychologist Ed Diener, began their study of happiness by developing a precise conceptual definition. Specifically, Diener's team reasoned that the word *happiness* might have a variety of meanings, so they explicitly limited their interest to "subjective well-being" (or well-being from a person's own perspective).

After defining happiness at the conceptual level, Diener and his colleagues developed an operational definition. Because they were interested in people's perspectives on their own well-being, they chose to operationalize subjective well-being, in part, by asking people to report on their own happiness in a questionnaire format. The researchers decided people should use their own criteria to describe what constitutes a "good life" (Pavot & Diener, 1993). They worded their questions so people could think about the interpretation of life satisfaction that was appropriate for them. These researchers operationally defined, or measured, subjective well-being by asking people to respond to five items about their satisfaction with life using a 7-point scale; 1 corresponded to "strongly disagree" and 7 corresponded to "strongly agree":

____ 1. In most ways my life is close to my ideal.
____ 2. The conditions of my life are excellent.
____ 3. I am satisfied with my life.
____ 4. So far, I have gotten the important things I want in life.
____ 5. If I could live my life over, I would change almost nothing.

The unhappiest people would get a total score of 5 on this self-report scale because they would answer "strongly disagree," or 1, to all five items (1 + 1 + 1 + 1 + 1 = 5). The happiest people would get a total score of 35 on this scale because they would answer "strongly agree," or 7, to all five items (7 + 7 + 7 + 7 + 7 = 35). Those at the neutral point would score 20—right in between satisfied and dissatisfied (4 + 4 + 4 + 4 + 4 = 20). Diener and Diener (1996) reported some data based on this scale, concluding that most people are happy, scoring above 20. For example, 63% of high school and college students scored above 20 in one study, and 72% of disabled adults scored above 20 in another study.

In choosing this operational definition of subjective well-being, the research team started with only one possible measure, even though there are many other ways to study this concept. Another way to measure happiness is to use a single question called the Ladder of Life (Cantril, 1965). The question goes like this:

> Imagine a ladder with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally stand at this time?

On this measure, participants respond by giving a value between 0 and 10. The Gallup polling organization uses the Ladder of Life scale in its daily Gallup-Healthways Well-Being Index.

Which one of these operational definitions do you think is a better measure of happiness? Diener's research team and Gallup have collected data on their measures of happiness and determined that they both do a good job of measuring the construct, as we'll see later in this chapter.

OPERATIONALIZING OTHER CONCEPTUAL VARIABLES

To study conceptual variables other than happiness, researchers follow a similar process: They start by stating a definition of their construct (the conceptual variable) and then create an operational definition. For example, to measure the association between wealth and happiness, researchers need to measure both happiness and wealth. They might operationally define wealth by asking about salary in dollars, by asking for bank account balances, or even by observing the kind of car people drive.

Consider another variable that has been studied in research on relationships: gratitude toward one's partner. Researchers who measure gratitude toward a relationship partner might operationalize it by asking people how often they thank their partner for something they did. Or they might ask people how appreciative they usually feel. Even a simple variable such as gender must be operationalized.

.1

es and Operational Definitions

	ONE POSSIBLE OPERATIONAL DEFINITION (OPERATIONALIZATION)	ANOTHER POSSIBLE OPERATIONAL DEFINITION
toward one's ıip partner	Asking people if they agree with the statement, "I appreciate my partner."	Watching couples interact and counting how many times they thank each other.
ıdentity	Asking people to report on a survey whether they identify as male, female, or nonbinary.	In phone interviews, a researcher guesses gender through the sound of the person's voice.
	Asking people to report their income within various ranges (less than $20,000, between $20,000 and 50,000, and more than $50,000).	Coding the value of a car from 1 (older, lower-status vehicle) to 5 (new, high-status vehicle in good condition).
ıce	An IQ test that includes problem-solving items, memory and vocabulary questions, and puzzles.	Recording brain activity while people solve difficult problems.
ing (happiness)	10-point Ladder of Life scale.	Diener's 5-item subjective well-being scale.

As **Table 5.1** shows, any conceptual variable can be operationalized in a number of ways. In fact, operationalizations are one place where creativity comes into the research process, as researchers work to develop new and better measures of their constructs.

Three Common Types of Measures

The types of measures psychological scientists typically use to operationalize variables generally fall into three categories: self-report, observational, and physiological.

SELF-REPORT MEASURES

A **self-report measure** operationalizes a variable by recording people's answers to questions about themselves in a questionnaire or interview. Diener's five-item scale and the Ladder of Life question are both examples of self-report measures about life satisfaction. Similarly, asking people how much they appreciate their partner and asking about gender identity are both self-report measures. If stress is the variable being studied, researchers might ask people to self-report on the frequency of specific events they've experienced in the past year, such as marriage, divorce, or moving (e.g., Holmes & Rahe, 1967).

In research on children, self-reports may be replaced with parent reports or teacher reports. These measures ask parents or teachers to respond to a series of questions, such as describing the child's recent life events, the words the child knows, or the child's typical classroom behaviors. (Chapter 6 discusses situations when self-report measures are likely to be accurate and when they might be biased.)

OBSERVATIONAL MEASURES

An **observational measure**, sometimes called a behavioral measure, operationalizes a variable by recording observable behaviors or physical traces of behaviors. For example, a researcher could operationalize happiness by observing how many times a person smiles. Intelligence tests can be considered observational measures, because the people who administer such tests in person are observing people's intelligent behaviors (such as being able to correctly solve a puzzle or quickly detect a pattern). Coding how much a person's car cost would be an observational measure of wealth (Piff et al., 2012).

Observational measures may record physical traces of behavior. Stress behaviors could be measured by counting the number of tooth marks left on a person's pencil, or a researcher could measure stressful events by consulting public legal records to document whether people have recently married, divorced, or moved. (Chapter 6 addresses how an observer's ratings of behavior might be accurate and how they might be biased.)

PHYSIOLOGICAL MEASURES

A **physiological measure** operationalizes a variable by recording biological data, such as brain activity, hormone levels, or heart rate. Physiological measures usually require the use of equipment to amplify, record, and analyze biological data. For example, moment-to-moment happiness has been measured using facial electromyography (EMG)—a way of electronically recording tiny movements in the muscles in the face. Facial EMG can be used to detect a happy facial expression because people who are smiling show particular patterns of muscle movement around the eyes and cheeks.

Other constructs might be measured using a brain scanning technique called functional magnetic resonance imaging, or fMRI. In a typical fMRI study, people engage in a carefully structured series of psychological tasks (such as looking at three types of photos or playing a series of rock-paper-scissors games) while lying in an MRI machine. The MRI equipment records and codes the relative changes in blood flow in particular regions of the brain, as shown in **Figure 5.1**. When more

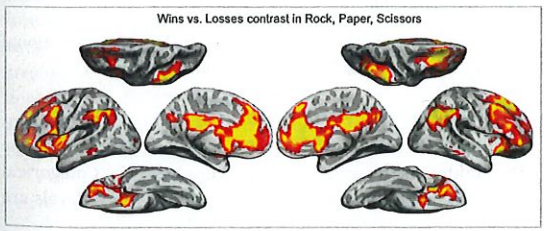


FIGURE 5.1
Images from fMRI scans showing brain activity.
In this study of how people respond to rewards and losses, the researchers tracked blood flow patterns in the brain when people had either won, lost, or tied a rock-paper-scissors game played with a computer. They found that many regions of the brain respond more to wins than to losses, as indicated by the highlighted regions. (Source: Vickery, Chun, & Lee, 2011.)

blood flows to a brain region while people perform a certain task, researchers conclude that brain area is activated because of the patterns on the scanned images.

Some research indicates a way fMRI might be used to measure intelligence in the future. Specifically, the brains of people with higher intelligence may be more efficient at solving complex problems, and their fMRI scans show relatively less brain activity for complex problems (Deary et al., 2010). Therefore, future researchers may be able to use the efficiency of brain activity as a physiological measure of intelligence. In contrast, head circumference, a physiological measure from a century ago, turned out to be flawed, because smarter brains are not necessarily stored inside larger skulls (Gould, 1996).

A physiological way to operationalize stress might be to measure the amount of the hormone cortisol released in saliva because people under stress show higher levels of cortisol (Carlson, 2009). Skin conductance, an electronic recording of the activity in the sweat glands of the hands or feet, is another way to measure stress physiologically. People under more stress have more activity in these glands. Another physiological measure detects electrical patterns in different brain regions near the scalp, using electroencephalography (EEG).

WHICH OPERATIONALIZATION IS BEST?

A single construct can be operationalized in several ways, from self-report to behavioral observation to physiological measures. Some people erroneously believe physiological measures are the most accurate. But even physiological results have to be validated by using other measures. For instance, as mentioned above, researchers used fMRI to learn that the brain works more efficiently relative to level of intelligence. But how was participant intelligence established in the first place? Before doing the fMRI scans, the researchers gave the participants an IQ test—an observational measure (Deary et al., 2010). Similarly, researchers might trust an fMRI pattern to indicate when a person is genuinely happy. However, the only way a researcher could know that some pattern of brain activity was associated with happiness would be by asking each person how happy they felt (a self-report measure) at the same time the brain scan was being done. As you'll learn later in this chapter, researchers normally expect self-report, observational, and physiological measures to show similar patterns of results.

Scales of Measurement

All variables must have at least two levels (see Chapter 3). The levels of operational variables, however, can be coded using different scales of measurement.

CATEGORICAL VS. QUANTITATIVE VARIABLES

Operational variables are primarily classified as categorical or quantitative. The levels of **categorical variables**, as the term suggests, are categories. (Categorical variables are also called *nominal variables*.) Examples are sex, whose levels are male and female, and species, whose levels in a study might be rhesus macaque, chimpanzee, and bonobo. A researcher might decide to assign numbers to the levels of a categorical variable (e.g., using "1" to represent rhesus macaques, "2" for chimps, and "3" for bonobos) during the data-entry process. However, the numbers do not have numerical meaning—a bonobo is different from a chimpanzee, but being a bonobo ("3") is not quantitatively "higher" than being a chimpanzee ("2").

In contrast, the levels of **quantitative variables** (also called continuous variables) are coded with *meaningful* numbers. Height and weight are quantitative because they are measured in numbers, such as 170 centimeters or 65 kilograms. Diener's scale of subjective well-being is quantitative too, because a score of 35 represents more happiness than a score of 7. IQ score, level of brain activity, and amount of salivary cortisol are also quantitative variables.

THREE TYPES OF QUANTITATIVE VARIABLES

For certain kinds of statistical tests, researchers need to further classify a quantitative variable in terms of ordinal, interval, or ratio scale.

An **ordinal scale** of measurement applies when the numerals of a quantitative variable represent a ranked order. For example, a bookstore's website might display the top 10 best-selling books. We know that the #1 book sold more than the #2 book, and that #2 sold more than #3, but we don't know whether the number of books that separates #1 and #2 is equal to the number of books that separates #2 and #3. In other words, the intervals may be unequal. Maybe the first two rankings are only 10 books apart, and the second two rankings are 150,000 books apart. Similarly, a professor might use the order in which exams were turned in to operationalize how fast students worked. This represents ordinal data because the fastest exams are on the bottom of the pile—ranked 1. However, this variable has not quantified *how much* faster each exam was turned in, compared with the others.

An **interval scale** of measurement applies to the numerals of a quantitative variable that meet two conditions: First, the numerals represent equal intervals (distances) between levels, and second, there is no "true zero" (a person can get a score of 0, but the 0 does not literally mean "nothing"). An IQ test is an interval scale—the distance between IQ scores of 100 and 105 represents the same as the distance between IQ scores of 105 and 110. However, a score of 0 on an IQ test does not mean a person has "no intelligence." Body temperature in degrees Celsius is another example of an interval scale—the intervals between levels are equal; however, a temperature of 0 degrees does not mean a person has "no temperature." Most researchers assume questionnaire scales like Diener's (scored from 1 = *strongly disagree* to 7 = *strongly agree*) are interval scales. They do not have a true zero, but we assume the distances between numerals, from 1 to 7, are equivalent. Because they do not have a true zero, interval scales cannot allow a researcher to say things like "twice as hot" or "three times happier."

Finally, a **ratio scale** of measurement applies when the numerals of a quantitative variable have equal intervals and when the value of 0 truly means "none"

2

ement Scales for Operational Variables

ARIABLE	CHARACTERISTICS	EXAMPLES
al	Levels are categories.	Nationality. Type of music. Kind of phone people use.
ive	Levels are coded with meaningful numbers.	
	A quantitative variable in which numerals represent a rank order. Distance between subsequent numerals may not be equal.	Order of finishers in a swimming race. Ranking of 10 shows from most to least favorite.
	A quantitative variable in which subsequent numerals represent equal distances, but there is no true zero.	IQ score. Shoe size. Degree of agreement on a 1–7 scale.
	A quantitative variable in which numerals represent equal distances and zero represents "none" of the variable being measured.	Number of exam questions answered correctly. How many episodes of a show watched. Height in cm.

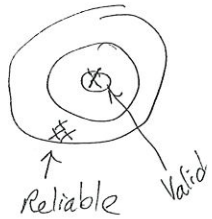
of the variable being measured. On a knowledge test, a researcher might measure how many items people answer correctly. On this scale, 0 truly represents "nothing correct" (0 answers correct). Even if nobody scores a 0, this value would still be meaningful. If a researcher measures how frequently people blink their eyes in a stressful situation, the number of eyeblinks is a ratio scale because 0 would represent zero eyeblinks. Because ratio scales do have a meaningful zero, a researcher can say something like "Alek answered twice as many problems as Hugo." **Table 5.2** summarizes all the above variations.

CHECK YOUR UNDERSTANDING

1. Explain why a variable usually has only one conceptual definition but can have multiple operational definitions.
2. Name the three common ways in which researchers operationalize their variables.
3. In your own words, describe the difference between categorical and quantitative variables. Come up with new examples of variables that would fit the definition of ordinal, interval, and ratio scales.

1. See pp. 118–120. 2. See pp. 120–122. 3. See pp. 122–124.

RELIABILITY OF MEASUREMENT: ARE THE SCORES CONSISTENT?



Once the variables in a study have been operationalized, we can ask the important construct validity question: How do we know that the operationalizations are appropriate? The construct validity of a measure has two aspects. **Reliability** refers to how consistent the results of a measure are, and **validity** refers to whether the operationalization is measuring what it is supposed to measure.

Introducing Three Types of Reliability

Researchers use data to decide which measures to use in a study, because establishing reliability is an empirical question. A measure's reliability is just what the word suggests: whether or not researchers can rely on a particular score. If an operationalization is reliable, it will yield a consistent pattern of scores every time.

Reliability can be assessed in three ways, depending on how a variable was operationalized, and all three involve consistency in measurement. When a measure has **test-retest reliability**, a study participant will get pretty much the same score each time they are measured with it. With **interrater reliability**, consistent scores are obtained no matter who measures the variable. With **internal reliability** (also called *internal consistency*), a study participant gives a consistent pattern of answers, no matter how the researchers phrase the question.

TEST-RETEST RELIABILITY

To illustrate test-retest reliability, let's suppose a sample of people took an IQ test today. When they take it again 1 month later, the pattern of scores should be consistent: People who scored the highest at Time 1 should also score the highest at Time 2. Even if all the scores from Time 2 have increased since Time 1 (due to practice or schooling), the pattern should be consistent: The highest-scoring Time 1 people should still be the highest scoring people at Time 2. Test-retest reliability can apply whether the operationalization is self-report, observational, or physiological, but it's most relevant when researchers are measuring constructs (such as intelligence, personality, or gratitude) that are theoretically stable. Happy mood, for example, may reasonably fluctuate from month to month or year to year for a particular person, so less consistency would be expected in this variable.

INTERRATER RELIABILITY

With interrater reliability, two or more independent observers will come up with consistent (or very similar) findings. Interrater reliability is most relevant for observational measures. Suppose you are assigned to observe the number of times each child smiles in 1 hour at a childcare playground. Your lab partner is assigned

to sit on the other side of the playground and make their own count of the same children's smiles. If, for one child, you record 12 smiles during the first hour and your lab partner also records 12 smiles in that hour for the same child, there is interrater reliability. Any two observers watching the same children at the same time should agree about which child has smiled the most and which child has smiled the least.

INTERNAL RELIABILITY

The third kind of reliability, internal reliability, applies to measures that combine multiple items. Suppose a sample of people take Diener's five-item subjective well-being scale. The questions on his scale are worded differently, but each item is intended to measure the same construct. Therefore, people who agree with the first item on the scale should also agree with the second item (as well as with Items 3, 4, and 5). Similarly, people who disagree with the first item should also disagree with Items 2, 3, 4, and 5. If the pattern is consistent across items in this way, the scale has internal reliability.

Using a Scatterplot to Quantify Reliability

Before using a particular measure in a study they are planning, researchers may collect data to see if it is reliable. They may also rely on data collected on the measure by other researchers. Two statistical devices researchers can use for data analysis are scatterplots (see Chapter 3) and the correlation coefficient *r* (discussed below). In fact, evidence for reliability is a special example of an association claim—the association between an earlier time and a later time, between one coder and another, or between one version of the measure and another.

Here's an example of how correlations are used to document reliability. Years ago, when some people thought smarter people had larger heads, they used head circumference as an operationalization of intelligence. Would this measure be reliable? Probably. Suppose you record the head circumference, in centimeters, for everyone in a classroom, using an ordinary tape measure. To see whether the measurements were reliable, you could measure all the heads twice (test-retest reliability) or you could measure them first, and then have someone else measure the same set (interrater reliability).

Figure 5.2 shows how the results of such a measurement might look, in the form of a data table and a scatterplot. In the scatterplot, the first measurements of head circumference for four students are plotted on the *y*-axis. The circumferences as measured the second time—whether by you again (test-retest) or by a second observer (interrater)—are plotted on the *x*-axis. In this scatterplot, each dot represents a person measured twice.

We would expect the two measurements of head circumference to be about the same for each person. They are, so the dots on the scatterplot all fall almost exactly on the sloping line that would indicate perfect agreement. The two measures won't always be exactly the same because there is likely to be some measurement error

A

	Head circumference (cm) 1	2
Taylor	50	52
Kendra	61	59
Mateo	65	64
Kurt	75	80

B

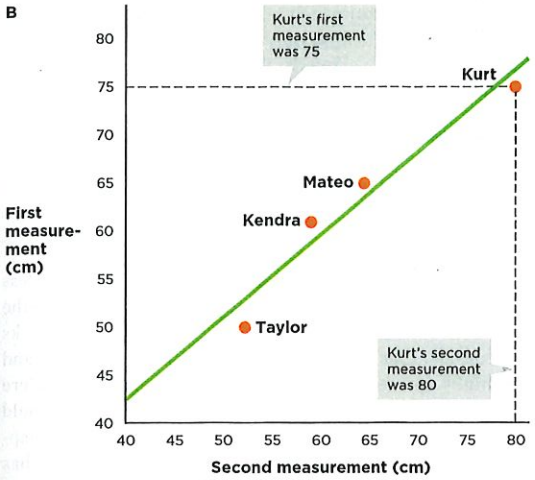


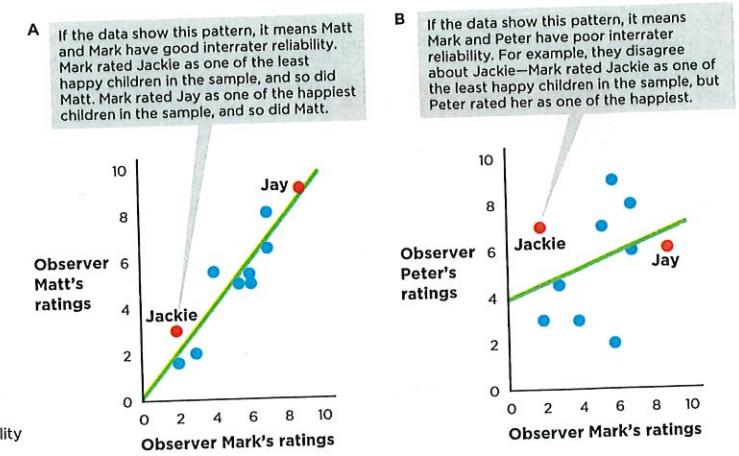
FIGURE 5.2
Two measurements of head circumference.
(A) The data for four participants in table form. **(B)** The same data presented in a scatterplot.

that will lead to slightly different scores even for the same person (such as variations in the tape measure placement for each person).

SCATTERPLOTS CAN SHOW INTERRATER AGREEMENT OR DISAGREEMENT

In a different scenario, suppose ten young children are being observed at a playground. Two independent observers, Mark and Matt, rate how happy each child appears to be, on a scale of 1 to 10. They later compare notes to see how well their ratings agree. From these notes, they create a scatterplot, plotting Observer Mark's ratings on the *x*-axis and Observer Matt's ratings on the *y*-axis.

If the data looked like those in **Figure 5.3A**, the ratings would have high interrater reliability. Both Mark and Matt rate Jay's happiness as 9—one of the happiest kids on the playground. Observer Mark rates Jackie a 2—one of the least happy kids. Observer Matt agreed because he rates her 3, and so on. The two observers do not show perfect agreement, but there are no great disagreements about the happiest and least happy kids. Again, the points are a bit scattered, but they cluster close to the sloping line that would indicate perfect agreement.



i.3
ır reliability.
ıter reliability is
ıterrater reliability

In contrast, suppose the data looked like **Figure 5.3B**, which shows much less agreement. Here, the two observers are Mark and Peter, and they are watching the same children at the same time, but Mark gives Jay a rating of 9 and Peter thinks he rates only a 6. Mark considers Jackie's behavior to be shy and withdrawn and rates her a 2, but Peter thinks she seems calm and content and rates her a 7. Here the interrater reliability would be considered unacceptably low. One reason could be that the observers did not have a clear enough operational definition of "happiness" to work with. Another reason could be that one or both of the coders has not been trained well enough yet.

A scatterplot can thus be a helpful tool for visualizing the agreement between two administrations of the same measurement (test-retest reliability) or between two coders (interrater reliability). Using a scatterplot, you can see whether the two ratings agree (if the dots are close to a straight line drawn through them) or whether they disagree (if the dots scatter widely from a straight line drawn through them).

Using the Correlation Coefficient *r* to Quantify Reliability

Scatterplots can provide a picture of a measure's reliability. However, a more common and efficient way to see if a measure is reliable is to use the correlation coefficient. Researchers can use a single number, called a **correlation coefficient**, or *r*, to indicate how close the dots, or points, on a scatterplot are to a line drawn through them.

Notice that the scatterplots in **Figure 5.4** differ in two important ways. One difference is that the scattered clouds of points slope in different directions. In Figure 5.4A and Figure 5.4B the points slope upward from left to right, in Figure 5.4C they slope downward, and in Figure 5.4D they do not slope up or down at all. This

»
more on the slope of a
terplot, see Chapter 3,
pp. 62–63.

slope is referred to as the direction of the relationship, and the **slope direction** can be positive, negative, or zero—that is, sloping up, sloping down, or not sloping at all.

The other way the scatterplots differ is that in some, the dots are close to a straight, sloping line; in others, the dots are more spread out. This spread corresponds to the **strength** of the relationship. In general, the relationship is strong when dots are close to the line; it is weak when dots are spread out.

The numbers below the scatterplots are the correlation coefficients, or *r*. The *r* indicates the same two things as the scatterplot: the direction of the relationship and the strength of the relationship, both of which psychologists use in evaluating reliability evidence. Notice that when the slope is positive, *r* is positive; when the slope is negative, *r* is negative. The value of *r* can fall only between 1.0 and –1.0. When

«
For more on how to compute *r*, see Statistics Review: Descriptive Statistics, pp. 480–484.

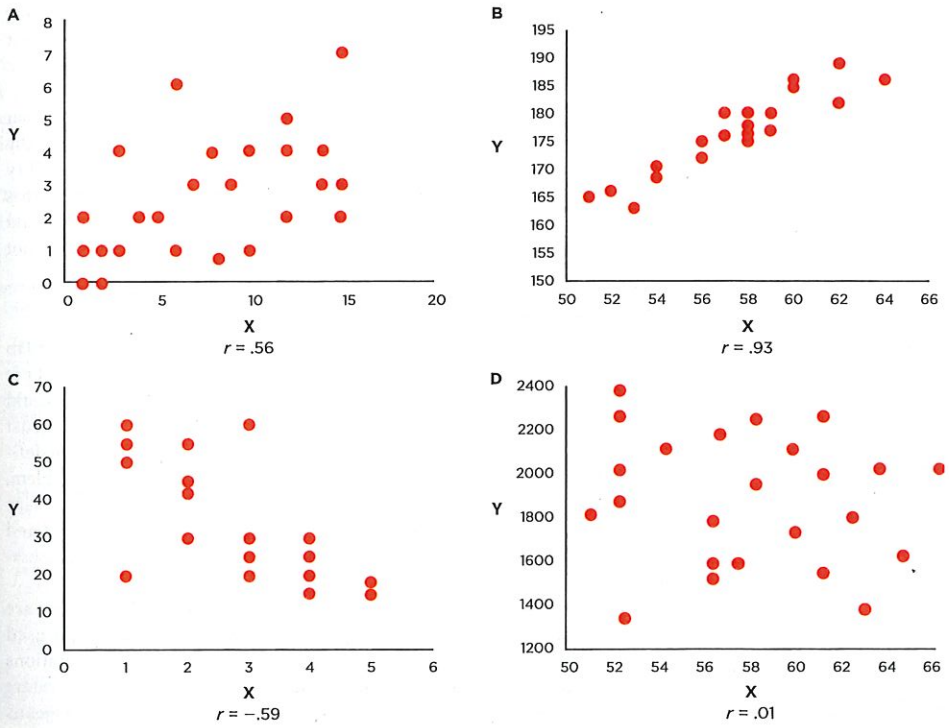


FIGURE 5.4
Correlation coefficients.
Notice the differences in the correlation coefficients (*r*) in these scatterplots. The correlation coefficient describes both the direction and the strength of the association between the two variables, regardless of the scale on which the variables are measured.

the relationship is strong, r is close to either 1 or –1; when the relationship is weak, r is closer to zero. An r of 1.0 represents the strongest possible positive relationship, and an r of –1.0 represents the strongest possible negative relationship. If there is no relationship between two variables, r will be .00 or close to .00 (i.e., .02 or –.04).

Those are the basics. How do psychologists use the strength and direction of r to evaluate reliability evidence?

TEST-RETEST RELIABILITY

To assess the test-retest reliability of some measure, we would assess the same set of participants on that measure at least twice. First, we'd give the set of participants the measure at Time 1. Then we'd wait a while (say, 2 months) and contact the same set of people again, at Time 2. After recording each person's score at Time 1 and Time 2, we could compute r. If r turns out to be positive and strong (for test-retest, we might expect .5 or above), we would have very good test-retest reliability. If r is positive but weak, we would know that participants' scores on the test changed from Time 1 to Time 2.

A low r would be a sign of poor reliability if we are measuring something that should stay the same over time. For example, a trait like intelligence is not usually expected to change over a few months, so if we assess the test-retest reliability of an IQ test and obtain a low r, we would be doubtful about the reliability of this test. In contrast, if we were measuring flu symptoms or seasonal stress, we would expect test-retest reliabilities to be low, simply because these constructs do not stay the same over time.

INTERRATER RELIABILITY

To test the interrater reliability of some measure, we might ask two observers to rate the same participants at the same time, and then we would compute r. If r is positive and strong (according to many researchers, r = .70 or higher), we would have very good interrater reliability. If r is positive but weak, we could not trust the observers' ratings. We would retrain the coders or refine our operational definition so it can be more reliably coded. A negative r would indicate a big problem. In the playground example, that would mean Observer Mark considered Jay very happy but Observer Peter considered Jay very unhappy, Observer Mark considered Jackie unhappy but Peter considered Jackie happy, and so on. When we're assessing reliability, a negative correlation is rare and undesirable.

Although r can be used to evaluate interrater reliability when the observers are rating a quantitative variable, a more appropriate statistic, called *kappa*, is used when the observers are rating a categorical variable. Although the computations are beyond the scope of this book, kappa measures the extent to which two raters place participants into the same categories. As with r, a kappa close to 1.0 means that the two raters agreed.

INTERNAL RELIABILITY

Internal reliability is relevant for measures that use multiple items or observations to get at the same construct. On self-report scales such as Diener's five-item subjective well-being scale, people answer the same question worded in multiple ways. Researchers usually plan to sum all the items to create a single composite score. Before combining the items, researchers assess the scale's internal reliability to evaluate whether people responded consistently to each item, despite the different wordings.

Let's consider the following adaptation of Diener's well-being scale. Would a group of people give consistent responses to all five items? Would people who agree with Item 1 also agree with Items 2, 3, 4, and 5?

_____ 1. In most ways my life is close to my ideal.
_____ 2. The conditions of my life are excellent.
_____ 3. I am satisfied with my life.
_____ 4. I am fond of polka dots.
_____ 5. If I could live my life over, I would change almost nothing.

You can see that only some of these items go together. Items 1 and 2 are probably correlated, since they are similar to each other, but Items 1 and 4 are probably not correlated, since people can like polka dots whether or not they are living their ideal lives. How do we quantify these intuitions?

First, researchers ask a large sample of participants to answer all of the items. Then they compute the correlations between every item and every other item (see **Table 5.3**). Next, they compute the **average inter-item correlation (AIC)**,

TABLE 5.3

Internal Reliability

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	ITEM 5
Item 1. "In most ways my life is close to my ideal."					
Item 2. "The conditions of my life are excellent."	.21				
Item 3. "I am satisfied with my life."	.22	.25			
Item 4. "I am fond of polka dots."	.01	–.05	.02		
Item 5. "If I could live my life over, I would change almost nothing."	.32	.40	.23	–.01	

Note: This matrix shows the correlations among the five items listed on page 131. To assess internal reliability, we start by computing the correlation of each item with every other item. To find the correlation (r) between two items, locate where one item's column and another item's row intersect (for example, the correlation between Item 1 and Item 3 is r = .22). You can see that Items 1, 2, 3, and 5 correlate with one another, but Item 4 (the one about polka dots) does not correlate with any of the other items. The pattern suggests that Item 4 is problematic and this set of items is not internally reliable. (Data are fabricated for illustration purposes.)

which is the average of all these correlations (for example, all 10 correlations in Table 5.3). An AIC between .15 and .50 means that the items go reasonably well together (Clark & Watson, 2019). After that, researchers might compute **Cronbach's alpha** (or *coefficient alpha*), which mathematically combines the AIC and the number of items in the scale. The closer Cronbach's alpha is to 1.0, the better the scale's reliability. For self-report measures, researchers are looking for Cronbach's alpha of .80 or higher (Clark & Watson, 2019). If the AIC or Cronbach's alpha is acceptable, the researchers determine that there is good internal reliability and they can sum all the items together. If the AIC or Cronbach's alpha is unacceptable, the researchers look carefully at the items, perhaps revising or omitting some.

Reading About Reliability in Journal Articles

Authors of empirical journal articles usually present reliability information for the measures they are using. One example of such evidence is shown in **Figure 5.5**, which comes from an actual journal article. According to the table, the subjective well-being scale called Satisfaction with Life (SWL) had been used in six studies. The table shows the internal reliability (labeled as coefficient alpha) from each of these studies as well as test-retest reliability for each one. The table did not present interrater reliability because the scale is a self-report measure, and interrater reliability is relevant only when two or more observers are rating something. Based on the evidence in this table, we can conclude that the subjective well-being scale has acceptable internal reliability and acceptable test-retest reliability. You'll see another example of how reliability is discussed in a journal article in the Working It Through section at the end of this chapter.

STRAIGHT FROM THE SOURCE

Table 2
Estimates of Internal Consistency and Temporal Reliability for the Satisfaction with Life Scale

Sample	Coefficient alpha	Test–retest	Temporal interval
Alfonso & Allison (1992a)	.89	.83	2 weeks
Pavot et al. (1991)	.85	.84	1 month
Blais et al. (1989)	.79–.84	.64	2 months
Diener et al. (1985)	.87	.82	2 months
Yardley & Rice (1991)	.80, .86	.50	10 weeks
Magnus, Diener, Fujita, & Pavot (1992)	.87	.54	4 years

Authors of study using SWL scale.

Coefficient (Cronbach's) alpha above .80 means SWL scale has good internal reliability.

High correlation of $r = .83$ for retesting 2 weeks apart means scale has good test-retest reliability.

5.5
ty of the well-being

rchers created this table
ow six studies supported
al reliability and test-retest
of their SWL scale.
Pavot & Diener, 1993, Table 2.)

CHECK YOUR UNDERSTANDING

1. Reliability is about consistency. Define the three kinds of reliability, using the word *consistent* in each of your definitions.
2. For each of the three common types of operationalizations—self-report, observational, and physiological—indicate which type(s) of reliability would be relevant.
3. Which of the following correlations is the strongest: $r = .25$, $r = -.65$, $r = -.01$, or $r = .43$?

1. See pp. 125–126. 2. Self-report: test-retest and internal may be relevant; observational: interrater and test-retest may be relevant; physiological: interrater and test-retest may be relevant. 3. $r = -.65$.

VALIDITY OF MEASUREMENT: DOES IT MEASURE WHAT IT'S INTENDED TO MEASURE?

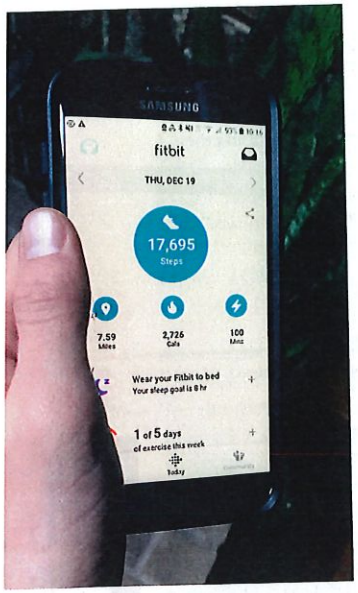
As they prepare to use some measure, researchers not only check to be sure it is reliable; they also want to be sure it measures the conceptual variables it was intended for. That's construct validity. You might ask how well the five-item well-being scale reflects Diener's theoretical definition of subjective well-being. Or you might ask whether a self-report measure of gratitude really reflects how thankful people are. You might ask whether recording the value of the car a person drives reflects that person's wealth.

Measurement reliability and measurement validity are separate steps in establishing construct validity. To demonstrate the difference between them, consider the example of head circumference as an operationalization of intelligence. Head size measurements are usually reliable because circumference is easy to measure. However, head circumference is not related to intelligence (Gould, 1996). Therefore, like a bathroom scale that always reads too light (**Figure 5.6**), the head circumference test may be reliable, but it is not valid as an intelligence test: It does not adequately capture the construct of intelligence.

FIGURE 5.6
Reliability is not the same as validity.
This person's bathroom scale may report that he weighs 50 pounds (22.7 kg) every time he steps on it. The scale is certainly reliable, but it is not valid.

Measurement Validity of Abstract Constructs

Does anyone you know use an activity monitor? Your friends may feel proud when they reach a daily steps goal or boast about how many miles they've covered



.7
ity monitors valid?
ore a pedometer during a hike and recorded
es. What data could you collect to know
not it accurately counted his steps?

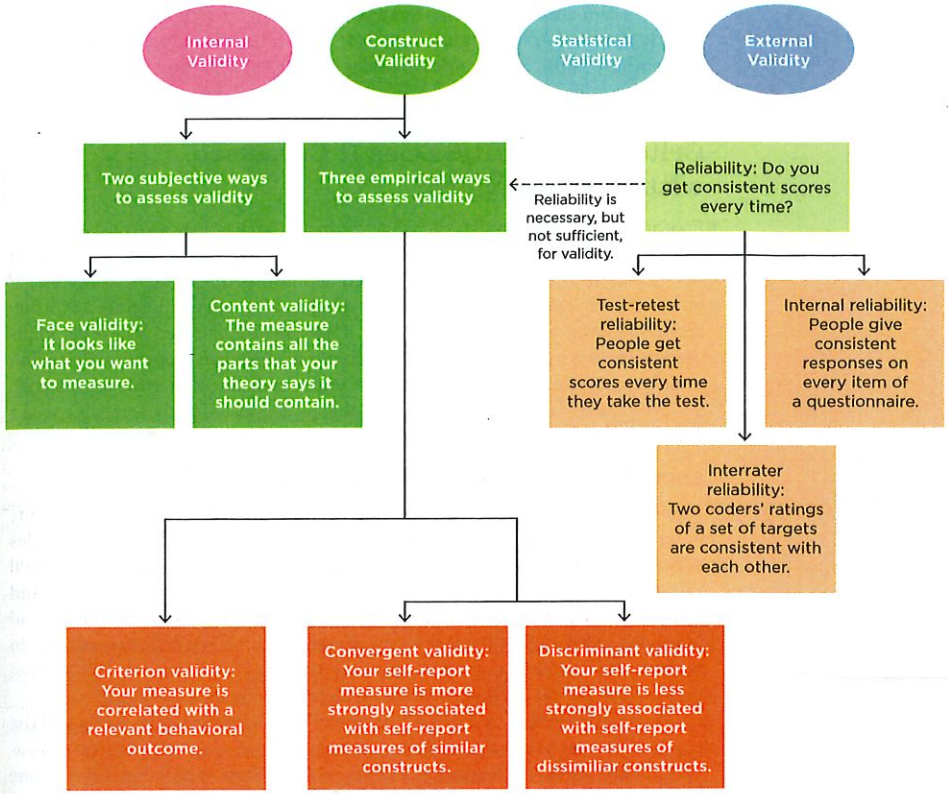
that day (**Figure 5.7**). How can you know for sure that these pedometers are accurate? Of course, it's straightforward to evaluate the validity of a pedometer: You'd simply walk around, counting your steps while wearing one, and then compare your own count to that of your device. If you're sure you walked 200 steps and your pedometer says you walked 200, then your device is probably valid. Similarly, if your pedometer recorded the correct distance after you've walked around a track or some other path with a known mileage, it's probably a valid monitor.

In the case of an activity monitor, we are lucky to have concrete, straightforward standards for accurate measurement. But psychological scientists often want to measure abstract constructs such as happiness, intelligence, stress, or self-esteem, which we can't simply count (Clark & Watson, 2019; Cronbach & Meehl, 1955; Smith, 2005a, 2005b). Construct validity is therefore especially important when a construct is not directly observable. Take happiness: We have no means of directly measuring how happy a person is. We could estimate it in a number of ways, such as scores on a well-being inventory, daily smile rate, blood pressure, stress hormone levels, or even the activity levels of certain brain regions. Yet each of these measures of happiness is indirect. And that is the challenge: How can we know if operationalizations are measuring our intended construct, happiness, and not something else?

We know by collecting a variety of data and evaluating it in light of our theory about the construct (Cronbach & Meehl, 1955). The evidence for construct validity is a matter of degree. Psychologists do not say a particular measure is or is not valid. Instead, they ask: What is the weight of evidence in favor of this measure's validity? Several kinds of evidence can convince a researcher, and we'll discuss them below. First, take a look at **Figure 5.8**, an overview of the reliability and validity concepts covered in this chapter.

Face Validity and Content Validity: Does It Look Like a Good Measure?

A measure has **face validity** if it is subjectively considered to be a plausible operationalization of the conceptual variable in question. Face-valid measures align well with the conceptual definition of a construct. Head circumference has high face validity as a measurement of hat size, but it has low face validity



as an operationalization of intelligence. In contrast, speed of problem solving, vocabulary size, and curiosity have higher face validity as operationalizations of intelligence. Researchers might check face validity by consulting experts. For example, we might assess the face validity of Diener's well-being scale by consulting people who are experts in the theoretical construct of well-being: Do they think the items are consistent with the construct?

Content validity also requires knowledge of the conceptual definition: A measure must capture all parts of a defined construct. For example, consider this conceptual definition of intelligence, which contains distinct elements, including the ability to "reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience" (Gottfredson, 1997, p. 13).

FIGURE 5.8
A concept map of measurement reliability and validity.
Construct validity is one of the four big validities, and it is supported by a variety of evidence.

To have adequate content validity, any operationalization of intelligence should include questions or items to assess each of these seven components. Indeed, most IQ tests have multiple categories of items, such as memory span, vocabulary, and problem-solving sections.

Criterion Validity: Does It Correlate with Key Behaviors?

Face and content validity establish that the operationalizations are consistent with the conceptual definition. Other forms of validity require data. We collect data to test that the measurement is associated with something it theoretically *should* be associated with. In some cases, such relationships can be illustrated by using scatterplots and correlation coefficients. They can be illustrated with other kinds of evidence too, such as comparisons of groups with known properties. **Criterion validity** evaluates whether the measure under consideration is associated with a concrete behavioral outcome that it should be associated with, according to the conceptual definition.

CORRELATIONAL EVIDENCE FOR CRITERION VALIDITY

Suppose you work for a company that wants to measure how well job applicants will perform as salespeople. Of the several commercially available tests of sales aptitude, which one should the company use? You have two choices, which we'll call Aptitude Test A and Aptitude Test B. Both look good in terms of face and content validity—their items ask about a candidate's motivation, optimism, and interest in sales. But do the test scores correlate with a key behavior: success in selling? It's an empirical question. Your company can collect data to tell them how well each of the two aptitude tests is correlated with success in selling.

To assess criterion validity, your company could give both sales tests to all the current sales representatives and then find out each person's sales figures—a measure of their selling performance. You would then compute two correlations: one between Aptitude Test A and sales figures, and the other between Aptitude Test B and sales figures. **Figure 5.9A** shows scatterplot results for Test A. The score on Aptitude Test A is plotted on the *x*-axis, and actual sales figures are plotted on the *y*-axis. (Alex scored 39 on the test and brought in $38,000 in sales, whereas Irina scored 98 and brought in $100,000.) **Figure 5.9B**, in contrast, shows the association of sales performance with Aptitude Test B.

Looking at these two scatterplots, we can see that the correlation in the first one is much stronger than in the second one. In other words, future sales performance is correlated more strongly with scores on Aptitude Test A than with scores on Aptitude Test B. If the data looked like this, the company would conclude that Aptitude Test A has better criterion validity as a measure of selling ability, so it seems better for hiring salespeople. In contrast, the other data show that scores on Aptitude Test B do not predict future sales performance; it has poor criterion validity as a measure of sales aptitude.

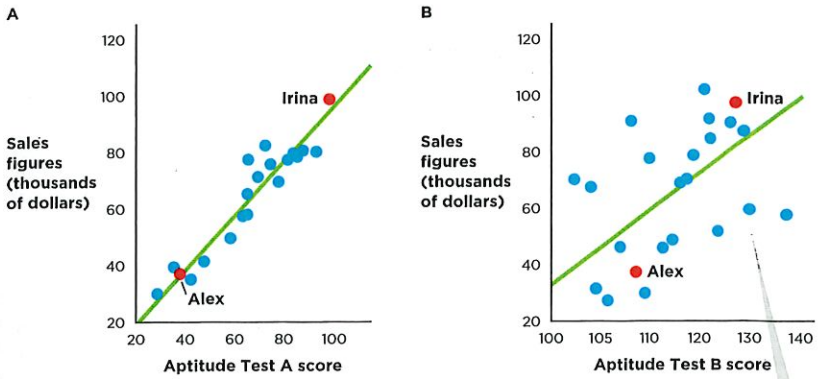


FIGURE 5.9
Correlational evidence for criterion validity.
(A) Aptitude Test A predicts sales performance, so criterion validity is high. **(B)** Aptitude Test B does not predict sales as well, so criterion validity is lower. A company would probably want to use Test A for identifying potential selling ability when selecting future salespeople.

Criterion validity is especially important for self-report measures because the correlation can indicate how well people's self-reports predict their actual behavior. Criterion validity provides some of the strongest evidence for a measure's construct validity. For example, Gallup presents criterion validity evidence for the 10-point Ladder of Life scale they use to measure happiness. They report that Ladder of Life scores correlate with key behavioral outcomes, such as becoming ill and missing work (Gallup, n.d.).

If an IQ test has criterion validity, it should be correlated with behaviors consistent with the construct of intelligence, such as how fast people can learn a complex set of symbols (an outcome that represents the conceptual definition of intelligence). Of course, the ability to learn quickly is only one component of that definition. Further criterion validity evidence could show that IQ scores are correlated with other behavioral outcomes that are theoretically related to intelligence, such as the ability to solve problems and indicators of life success (e.g., graduating from college, being employed in a high-level job, being curious and thoughtful).

KNOWN-GROUPS EVIDENCE FOR CRITERION VALIDITY

Another way to gather evidence for criterion validity is to use a **known-groups paradigm**, in which researchers see whether scores on the measure can discriminate among two or more groups whose behavior is already confirmed. For example, to validate the use of salivary cortisol as a measure of stress, a researcher could compare the salivary cortisol levels in two groups of people: those who are about to give a speech in front of a classroom and those who are in the audience. Public

speaking is recognized as being a stressful situation for almost everyone. Therefore, if salivary cortisol is a valid measure of stress, people in the speech group should have higher levels of cortisol than those in the audience group.

Lie detectors are another good example. These instruments record a set of physiological measures (such as skin conductance and heart rate) that are supposed to be different when a person is lying versus telling the truth. If skin conductance and heart rate are valid measures of lying, we could conduct a known-groups test in which we know in advance which of a person's statements are true and which are false. The physiological measures should be elevated only for the lies, not for the true statements. (For a review of the mixed evidence on lie detection, see Saxe, 1991.)

The known-groups method can also be used to validate self-report measures. Psychiatrist Aaron Beck and his colleagues developed the Beck Depression Inventory (BDI), a 21-item self-report scale with items that ask about major symptoms of depression (Beck et al., 1961). Participants circle one of four choices, such as the following:

0 I do not feel sad.
1 I feel sad.
2 I am sad all the time and I can't snap out of it.
3 I am so sad or unhappy that I can't stand it.

0 I have not lost interest in other people.
1 I am less interested in other people than I used to be.
2 I have lost most of my interest in other people.
3 I have lost all of my interest in other people.

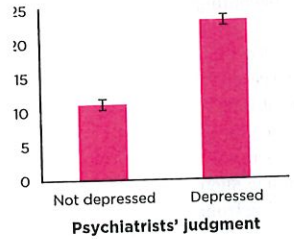
A clinical scientist adds up the scores on each of the 21 items for a total BDI score, which can range from a low of 0 (not at all depressed) to a high of 63.

To test the criterion validity of the BDI, Beck and his colleagues gave this self-report scale to two known groups of people. They knew one group was suffering from clinical depression and the other group was not because they had asked psychiatrists to conduct clinical interviews and diagnose each person. The researchers computed the mean BDI scores of the two groups and created a bar graph, shown in **Figure 5.10**. The evidence supports the criterion validity of the BDI. The graph shows the expected result: The average BDI score of the known group of depressed people was higher than the average score of the known group who were not depressed. Because its criterion validity was established in this way, the BDI is still widely used today when researchers need a quick and valid way to identify people who are vulnerable to depression.

Beck also used the known-groups paradigm to calibrate low, medium, and high scores on the BDI. When the psychiatrists interviewed the people in the sample, they evaluated not only whether they were depressed but also the level of depression in each person: none, mild, moderate, or severe. As expected, the BDI scores of the groups rose as their level of depression (assessed by psychiatrists) was more severe (**Figure 5.11**). This result was even clearer evidence that the BDI was a valid measure of depression. With the BDI, clinicians and researchers can confidently use specific ranges of BDI scores to categorize how severe a person's depression might be.

Diener's subjective well-being scale is another example of using the known-groups paradigm for criterion validity. In one review article, he and his colleague presented the subjective well-being scale averages from several studies. Each study had given the scale to different groups of people who could be expected to vary in happiness level (Pavot & Diener, 1993). For example, male prison inmates, a group that would be expected to report low subjective well-being, showed a lower mean score on the scale than Canadian college students, who averaged much higher—indicated by the *M* column in **Table 5.4**. Such known-groups patterns provide strong evidence for the criterion validity of the subjective well-being scale. Researchers can therefore have more confidence in the scale's validity.

What about the Ladder of Life scale, the measure of happiness used in the Gallup-Healthways Well-Being Index? This measure also has some known-groups evidence to support its criterion validity. National scores on this measure dropped, as you might expect, during the first months of the coronavirus pandemic (Witters & Harter, 2020). Scores also generally drop during economic recessions and rise during the summer months. These results fit what we would expect if the Ladder of Life is a valid measure of well-being.



E 5.10
ores of two known groups.
ttern of results provides evidence criterion validity of the BDI using own-groups method. Clients l to be more depressed by atrists also scored higher. Error bars e standard error of each mean. e: Adapted from Beck et al., 1961.)

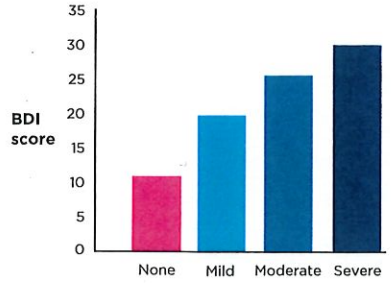


FIGURE 5.11
BDI scores of four known groups.
This pattern of results means it is valid to use BDI cutoff scores to decide if a person has mild, moderate, or severe depression. (Source: Adapted from Beck et al., 1961.)

TABLE 5.4

Subjective Well-Being Scores for Known Groups from Several Studies

SAMPLE CHARACTERISTICS	*N*	*M*	*SD*	STUDY REFERENCE
American college students	244	23.7	6.4	Pavot & Diener (1993)
French Canadian college students (male)	355	23.8	6.1	Blais et al. (1989)
Korean university students	413	19.8	5.8	Suh (1993)
Printing trade workers	304	24.2	6.0	George (1991)
Veterans Affairs hospital inpatients	52	11.8	5.6	Frisch (1991)
Abused women	70	20.7	7.4	Fisher (1991)
Male prison inmates	75	12.3	7.0	Joy (1990)

Note: N = Number of people in group. *M* = Group mean on subjective well-being. *SD* = Group standard deviation.
Source: Adapted from Pavot & Diener, 1993, Table 1.

Convergent Validity and Discriminant Validity: Does the Pattern Make Sense?

Criterion validity examines whether a measure correlates with key behavioral outcomes. Another form of validity evidence is whether

there is a meaningful pattern of similarities and differences among related measures. A self-report measure should correlate more strongly with measures of similar constructs than it does with those of dissimilar constructs. The patterns of correlations with measures of theoretically similar and dissimilar constructs are called **convergent validity** and **discriminant validity** (or *divergent validity*), respectively.

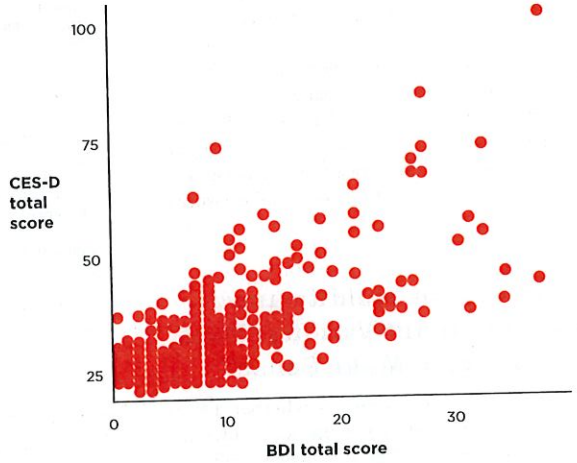
CONVERGENT VALIDITY

As an example of convergent validity, let's consider Beck's depression scale, the BDI, again. One team of researchers wanted to test the convergent and discriminant validity of the BDI (Segal et al., 2008). If the BDI really quantifies depression, the researchers reasoned, it should be correlated with (should converge with) other self-report measures of depression. Their sample of 376 adults completed the BDI and a number of other questionnaires, including a self-report instrument called the Center for Epidemiologic Studies Depression scale (CES-D).

As expected, BDI scores were positively, strongly correlated with CES-D scores (r = .68). People who scored as depressed on the BDI also scored as depressed on the CES-D; likewise, those who scored as not depressed on the BDI also scored as not depressed on the CES-D. **Figure 5.12** shows a scatterplot of the results. (Notice that most of the dots fall in the lower left portion of the scatterplot because most people in the sample are not depressed; they score low on both the BDI and the CES-D.) This correlation between similar self-report measures of the same construct (depression) provided evidence for the convergent validity of the BDI.

» on the strength of ns, see Chapter 8, 8.4, and Statistics eview: Descriptive tics, pp. 480–488.

This example of convergent validity is somewhat obvious: A measure of depression should correlate with a different measure of the same construct—depression.



5.12
e supporting the convergent of the BDI.
s strongly correlated with another of depression, the CES-D (r = .68), l evidence for convergent validity. Segal et al., 2008.)

But convergent validity evidence also includes *similar* constructs, not just the same one. The researchers showed, for instance, that the BDI scores were strongly correlated with a score quantifying psychological well-being (r = –.65). The strong negative correlation they observed made sense as a form of convergent validity because people who are depressed are also expected to have lower levels of well-being (Segal et al., 2008).

DISCRIMINANT VALIDITY

The BDI should *not* correlate strongly with measures of constructs that are very different from depression—it should show discriminant validity with them. For example, depression is not the same as a person's perception of their overall physical health. Although mental health problems, including depression, do overlap somewhat with physical health problems, we would not expect the BDI to be strongly correlated with a measure of perceived physical health problems. More important, we would expect the BDI to be more strongly correlated with the CES-D than it is with physical health problems. Sure enough, Segal and his colleagues found a correlation of only r = .16 between the BDI and a measure of perceived physical health problems. This weak correlation shows that the BDI is different from people's perceptions of their physical health, so we can say that the BDI has discriminant validity with physical health problems.

Figure 5.13 shows a scatterplot of the results. Notice that most of the dots fall in the lower left portion of the scatterplot because most people in the sample reported few health problems and were not depressed: They scored low on both the BDI and on the perceived physical health problems scale.

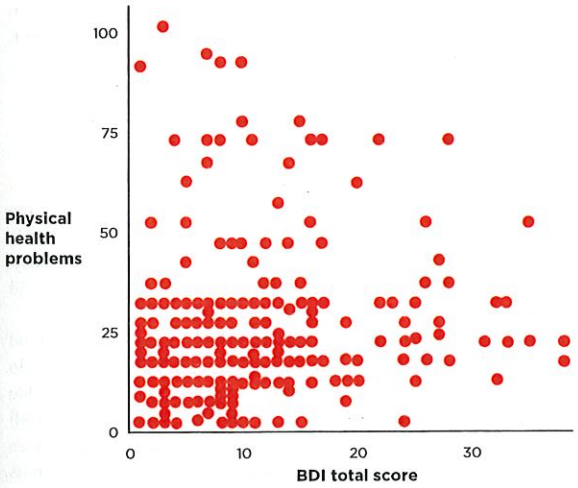


FIGURE 5.13
Evidence supporting the discriminant validity of the BDI.
As expected, the BDI is only weakly correlated with perceived health problems (r = .16), providing evidence for discriminant validity. (Source: Segal et al., 2008.)

As another example of discriminant validity, consider that many developmental disorders have similar symptoms, but diagnoses can vary. It might be important to specify, for instance, whether a child has autism or only a language delay. Therefore, a screening instrument for identifying autism should have discriminant validity; it should not accidentally diagnose that same child as having a language delay. Similarly, a scale that is supposed to diagnose learning disabilities should not be correlated with IQ because learning disabilities are not related to general intelligence.

It is usually not necessary to establish discriminant validity between a measure and something that is completely unrelated. Because depression is not likely to be associated with how many movies you watch or the number of siblings you have, we would not need to examine its discriminant validity with these variables. Instead, researchers focus on discriminant validity for "near neighbors": similar but different constructs. Does the BDI measure depression or perceived health problems? Does Diener's subjective well-being scale measure enduring happiness or just temporary mood? Does a screening technique identify autism or language delay?

Convergent validity and discriminant validity are usually evaluated together, as a pattern of correlations among self-report measures. A measurement should have higher correlations (higher *r* values) with similar traits (convergent validity) than it does with dissimilar traits (discriminant validity). There are no strict rules for what the correlations should be. Instead, the overall pattern of convergent and discriminant validity helps researchers decide whether their operationalization really measures the construct they want it to measure.

Finally, recall that construct validity requires evaluating the *weight* of the evidence; no single definitive outcome will establish validity of a measure (Smith, 2005a). A set of convergent and discriminant validity correlations (such as those in Figures 5.12 and 5.13) is helpful, but researchers must also consider face, content, and criterion validity evidence as well.

The Relationship Between Reliability and Validity

One essential point is worth reiterating: The validity of a measure is not the same as its reliability. A person might boast that some operationalization of behavior is "a very reliable test." But to say that a measure is "reliable" is only part of the story. Recall that head circumference is an extremely reliable measure, but it is not valid for assessing intelligence.

Although a measure may be less valid than it is reliable, it cannot be more valid than it is reliable. Intuitively, this statement makes sense. Reliability has to do with how well a measure correlates with itself. For example, an IQ test is reliable if it is correlated with itself over time. Validity, however, has to do with how well a measure is associated with something else, such as a behavior that indicates intelligence. An IQ test is valid if it is associated with another variable, such as school grades or life success. If a measure does not even correlate with itself, then how can it be more strongly associated with some other variable?

As another example, suppose you used your pedometer to count how many steps you take in your daily walk from your parking spot to your building. If the pedometer reading is very different from day to day, then the measure is unreliable—and of course, it also cannot be valid because the true distance of your walk has not changed. Therefore, reliability is necessary (but not sufficient) for validity.

CHECK YOUR UNDERSTANDING

1. What do face validity and content validity have in common?
2. Many researchers believe criterion validity is more important than convergent and discriminant validity. Can you see why?
3. Which requires stronger correlations for its evidence: convergent validity or discriminant validity? Which requires weaker correlations?
4. Can a measure be reliable but not valid? Can it be valid but unreliable?

1. They both require an expert's judgment; see pp. 134–136. 2. Because only criterion validity establishes how well a measure correlates with a behavioral outcome, not simply with other self-report measures; see pp. 136–142. 3. Convergent validity; discriminant validity. 4. It can be reliable but not valid, but a measure cannot be valid if it is unreliable; see pp. 142–143.

REVIEW: INTERPRETING CONSTRUCT VALIDITY EVIDENCE

Before using a stopwatch in a track meet, a coach wants to be sure the stopwatch is working well. Before taking a patient's blood pressure, a nurse wants to be sure the cuff she's using is reliable and accurate. Similarly, before conducting a study, researchers want to be sure all the measures they plan to use are reliable and valid ones. When you read empirical articles, look for the validity evidence for the measures the researchers used. You'll usually find reliability and validity information in the Method section, where the authors describe their measures. You should ask: Did the researchers present evidence that the measures they used have construct validity?

It turns out that many researchers report only internal reliability (Cronbach's alpha); they do not provide any other convergent, discriminant, or criterion validity evidence. That's a problem (Flake et al., 2017). However, some researchers get it right. The Working It Through section presents a good example of how validity information was described in a study by Gordon and colleagues (2012).

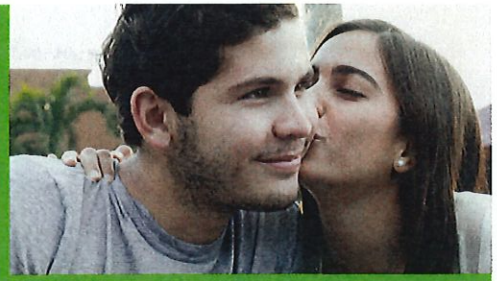
Item
1. I tell my partner often that s/he is the best.
2. I often tell my partner how much I appreciate her/him.
3. At times I take my partner for granted. (reverse scored item)
4. I appreciate my partner.
5. Sometimes I don't really acknowledge or treat my partner like s/he is someone special. (reverse scored item)
6. I make sure my partner feels appreciated.
7. My partner sometimes says that I fail to notice the nice things that s/he does for me. (reverse scored item)
8. I acknowledge the things that my partner does for me, even the really small things.
9. I am sometimes struck with a sense of awe and wonder when I think about my partner being in my life.

FIGURE 5.14
Items in the Appreciation in Relationships (AIR) Scale.
These items were used by the researchers to measure how much people appreciate their relationship partner. Do you think these items have face validity as a measure of appreciation? (Source: Gordon et al., 2012.)

The evidence reported by Gordon et al. (2012) supports the Appreciation in Relationships (AIR) scale as a reliable and valid scale (**Figure 5.14**). It has internal and test-retest reliability, and there is evidence of its convergent, discriminant, and criterion validity. The researchers were confident they could use AIR when they later tested their hypothesis that more appreciative couples would have healthier relationships. Many of their hypotheses about gratitude (operationalized by the AIR scale) were supported. One of the more dramatic results was from a study that followed couples over time. The authors reported: "We found that people who were more appreciative of their partners were significantly more likely to still be in their relationships at the 9-month follow-up" (Gordon et al., 2012, p. 268).

This empirical journal article illustrates how researchers use data to establish the construct validity of the measure they plan to use. Their research was picked up by the popular press and given the headline "Gratitude is for lovers."

WORKING IT THROUGH



How Well Can We Measure the Amount of Gratitude Couples Express to Each Other?

What do partners bring to a healthy romantic relationship? One research team proposed that gratitude toward one's partner would be important (Gordon et al., 2012). They predicted that when people are appreciative of their partners, close relationships are happier and last longer. In an empirical journal article, the researchers reported how they tested this hypothesis. But before they could study how the concept of gratitude contributes to relationship health, they needed to be able to measure the variable "gratitude" in a reliable and valid way. They created and tested the Appreciation in Relationships, or AIR, scale. We will work through the ways this example illustrates concepts from Chapter 5.

QUESTIONS TO ASK	CLAIMS, QUOTES, OR DATA	INTERPRETATION AND EVALUATION
Conceptual and Operational Definitions		
How did they operationalize the conceptual variable "gratitude"?	"In the first step, we created an initial pool of items based on lay knowledge, theory, and previous measures of appreciation and gratitude. . . . These items were designed to capture a broad conceptualization of appreciation by including items that assess the extent to which people recognize and value their partner as a person as well as the extent to which they are grateful for a partner's kind deeds" (p. 260).	This quoted passage describes how Gordon and her colleagues developed and selected the AIR items. Notice how they wrote items to capture their conceptual definition of gratitude (see Figure 5.14).

(Continued)

ONS TO ASK	CLAIMS, QUOTES, OR DATA	INTERPRETATION AND EVALUATION
AIR scale reliable? Did the scale give consistent scores?		
Reliability elf-report scale ple items, it ve good internal . Did the AIR e good internal ?	"In the initial online survey, participants completed a questionnaire with basic demographic information. Participants completed the AIR scale . . . α = .87" (p. 266).	In this passage, the authors report the internal reliability of the AIR scale. The value α = .87 indicates that people in the Gordon study answered all the AIR items consistently. A Cronbach's alpha above .80 is considered good internal reliability.
test Reliability t expect the e to have test- iability because e should be stable e. Were the AIR table over time?	"The AIR scale had strong test-retest reliability from baseline to the 9-month follow-up (. . . *r* = .61, *p* = .001)" (p. 267).	This passage reports the test-retest correlation, which was *r* = .61. Those who were the most appreciative at Time 1 were also the most appreciative at Time 2; similarly, those who were least appreciative at Time 1 were also least appreciative at Time 2.
ter Reliability the AIR scale report measure, archers do not report interrater y evidence.		
idence indicates the AIR scale has adequate internal and test-retest reliability. What evidence is for its validity? Does the scale really measure the concept of gratitude?		
rgent and minant Validity scores correlate rongly with es similar to de and less strongly easures dissimilar tude?	In a section on convergent and discriminant validity, the authors write: "As expected, . . . [the AIR scale was] positively correlated with the extent to which people had a grateful disposition [*r* = .25], as well as with people's gratitude in response to their partners' kind acts [*r* = .60]. In contrast, [the AIR scale was not] associated with people's feelings of indebtedness to their partners [*r* = .19]" (p. 262).	In this passage, the authors give convergent and discriminant validity evidence. The AIR scale has convergent validity with other measures of gratitude and discriminant validity with a measure of indebtedness. In other words, there was a pattern of higher correlations with gratitude than with indebtedness.
ion Validity he AIR scale predict nt behavioral nes?	The "final study allowed us to provide additional evidence for the validity of the AIR scale by examining cross-partner associations. . . . [P]eople who reported feeling more appreciative of their partners had partners who felt more appreciated by them, β = .50, t(66) = 5.87, *p* < .001, . . . suggesting that the AIR scale is capturing the interpersonal transmission of appreciation from one partner to the other" (p. 269).	In this passage, the authors present criterion validity evidence. If the AIR scale is a valid measure, you'd expect partners with higher AIR scores to also have partners who notice this appreciation. Because the results showed that AIR scores were associated with this relevant outcome, there is evidence for the AIR scale's criterion validity.

CHAPTER REVIEW

It's time to complete your study experience! Go to INQUIZITIVE to practice actively with this chapter's concepts and get personalized feedback along the way.

Summary

The construct validity of a study's measured variables is something you will interrogate for any type of claim.

WAYS TO MEASURE VARIABLES

- Psychological scientists measure variables in every study they conduct. Three common types of measures are self-report, in which people report on their own behaviors, beliefs, or attitudes; observational measures, in which raters record the visible behaviors of people or animals; and physiological measures, in which researchers measure biological data, such as heart rate, brain activity, and hormone levels.
- Depending on how they are operationalized, variables may be categorical or quantitative. The levels of categorical variables are categories. The levels of quantitative variables are meaningful numbers, in which higher numbers represent more of some variable.
- Quantitative variables can be further classified in terms of ordinal, interval, or ratio scales.

RELIABILITY OF MEASUREMENT: ARE THE SCORES CONSISTENT?

- Both measurement reliability and measurement validity are important for establishing a measure's construct validity.
- Researchers use scatterplots and correlation coefficients (among other methods) to evaluate evidence for a measure's reliability and validity.
- To establish a measure's reliability, researchers collect data to see whether the measure works consistently. There are three types of measurement reliability.
- Test-retest reliability establishes whether a sample gives a consistent pattern of scores at more than one testing.
- Interrater reliability establishes whether two observers give consistent ratings to a sample of targets.
- Internal reliability is established when people answer similarly worded items in a consistent way.
- Measurement reliability is necessary but not sufficient for establishing measurement validity.

VALIDITY OF MEASUREMENT: DOES IT MEASURE WHAT IT'S INTENDED TO MEASURE?

- Measurement validity can be established with subjective judgments (face validity and content validity) or with empirical data.
- Criterion validity requires collecting data that show a measure is correlated with expected behavioral outcomes.
- Convergent and discriminant validity require collecting data that show a measure is correlated more strongly with measures of similar constructs than with measures of dissimilar constructs.

REVIEW: INTERPRETING CONSTRUCT VALIDITY EVIDENCE

- Measurement reliability and validity evidence are reported in the Method section of empirical journal articles. Details may be provided in the text, as a table of results, or through cross-reference to a longer article that presents full reliability and validity evidence.

ms

lefinition, p. 118
ieasure, p. 120
l measure, p. 121
l measure, p. 121
ariable, p. 122
variable, p. 123
, p. 123
e, p. 123
. 123
125

validity, p. 125
test-retest reliability, p. 125
interrater reliability, p. 125
internal reliability, p. 125
correlation coefficient *r*, p. 128
slope direction, p. 129
strength, p. 129
average inter-item correlation, (AIC) p. 131
Cronbach's alpha, p. 132

face validity, p. 134
content validity, p. 135
criterion validity, p. 136
known-groups paradigm, p. 137
convergent validity, p. 140
discriminant validity, p. 140

see samples of chapter concepts in the popular media, sit www.everydayresearchmethods.com and click the box for Chapter 5.

Questions

each operational variable below as cate- or quantitative. If the variable is quantitative, classify it as ordinal, interval, or ratio.

ee of pupil dilation in a person's eyes in a study mantic couples (measured in millimeters).

iber of books a person owns.

ook's sales rank on Amazon.com.

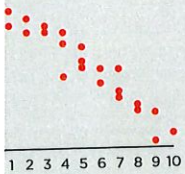
language a person speaks at home.

onality of the participants in a cross-cultural ly of Canadian, Ghanaian, and French students.

udent's grade in school.

of the following correlation coefficients best es the pictured scatterplot?

78
-.95
03
-.45



3. Classify each of the following results as an example of internal reliability, interrater reliability, or test-retest reliability.
   a. A researcher finds that people's scores on a measure of extroversion stay stable over 2 months.
   b. An infancy researcher wants to measure how long a 3-month-old baby looks at a stimulus on the right and left sides of a screen. Two undergraduates watch a tape of the eye movements of ten infants and time how long each baby looks to the right and to the left. The two sets of timings are correlated $r = .95$.
   c. A researcher asks a sample of 40 people a set of five items that all capture how extroverted they are. The Cronbach's alpha for the five items is found to be .85.
4. Classify each result below as an example of face validity, content validity, convergent and discriminant validity, or criterion validity.
   a. A professor gives a class of 40 people his five-item measure of conscientiousness (e.g., "I get chores done right away," "I follow a schedule," "I do not make a mess of things"). Average scores are correlated ($r = -.20$) with how many times each student has been late to class during the semester.
   b. A professor gives a class of 40 people his five-item measure of conscientiousness (e.g., "I get chores done right away," "I follow a schedule," "I do not make a mess of things"). Average scores are more highly correlated with a self-report measure of tidiness ($r = .50$) than with a measure of general knowledge ($r = .09$).
   c. The researcher e-mails his five-item measure of conscientiousness (e.g., "I get chores done right away," "I follow a schedule," "I do not make a mess of things") to 20 experts in personality psychology and asks them if they think his items are a good measure of conscientiousness.
   d. The researcher e-mails his five-item measure of conscientiousness (e.g., "I get chores done right away." "I follow a schedule," "I do not make a mess of things") to 20 experts in personality psychology and asks them if they think he has included all the important aspects of conscientiousness.

Learning Actively

1. For each measure below, indicate which kinds of reliability would need to be evaluated. Then draw a scatterplot indicating that the measure has good reliability and another one indicating the measure has poor reliability. (Pay special attention to how you label the axes of your scatterplots.)
   a. Researchers place unobtrusive video recording devices in the hallway of a local high school. Later, coders view tapes and code how many students are using cell phones in the 4-minute period between classes.
   b. Clinical psychologists have developed a seven-item self-report measure to quickly identify people who are at risk for panic disorder.
   c. Psychologists measure how long it takes a mouse to learn an eyeblink response. For 60 trials, they present a mouse with a distinctive blue light followed immediately by a puff of air. The 5th, 10th, and 15th trials are test trials, in which they present the blue light alone (without the air puff). The mouse is said to have learned the eyeblink response if observers record that it blinked its eyes in response to a test trial. The earlier in the 60 trials the mouse shows the eyeblink response, the faster it has learned the response.
   d. Educational psychologists use teacher ratings of classroom shyness (on a 9-point scale, where 1 = *not at all shy in class* and 9 = *very shy in class*) to measure children's temperament.
2. Consider how you might validate the 9-point classroom shyness rating example in Question 1d. First, what behaviors might be relevant for establishing this rating's criterion validity? Draw a scatterplot showing the results of a study in which the classroom shyness rating has good criterion validity (be careful how you label the axes). Second, come up with ways to evaluate the convergent and discriminant validity of this rating system. What traits should correlate strongly with shyness? What traits should correlate only weakly or not at all? Explain why you chose those traits. Draw a scatterplot showing the results of a study in which the shyness rating has good convergent or discriminant validity (be careful how you label the axes).
3. This chapter included the example of a sales ability test. Search online for "sales ability assessment" and see what commercially available tests you can find. Do the websites present reliability or validity evidence for the measures? If so, what form does the evidence take? If not, what kind of evidence would you like to see? You might frame your predictions in this form: "If this sales ability test has convergent validity, I would expect it to be correlated with . . . ."; "If this sales ability test has discriminant validity, I would expect it *not* to be correlated with . . . ."; "If this sales ability test has criterion validity, I would expect it to be correlated with . . . ."