



This is “Psychological Measurement”, chapter 5 from the book [Psychology Research Methods: Core Skills and Concepts \(index.html\)](#) (v. 1.0).

This book is licensed under a [Creative Commons by-nc-sa 3.0](http://creativecommons.org/licenses/by-nc-sa/3.0/) license. See the license for more details, but that basically means you can share this book as long as you credit the author (but see below), don't make money from it, and do make it available to everyone else under the same terms.

This content was accessible as of December 29, 2012, and it was downloaded then by [Andy Schmitz](#) (<http://lardbucket.org>) in an effort to preserve the availability of this book.

Normally, the author and publisher would be credited here. However, the publisher has asked for the customary Creative Commons attribution to the original publisher, authors, title, and book URI to be removed. Additionally, per the publisher's request, their name has been removed in some passages. More information is available on this project's [attribution page](http://2012books.lardbucket.org/attribution.html?utm_source=header).

For more information on the source of this book, or why it is available for free, please see [the project's home page](#) (<http://2012books.lardbucket.org/>). You can browse or download additional books there.

Chapter 5

Psychological Measurement

Researchers Tara MacDonald and Alanna Martineau were interested in the effect of female college students' moods on their intentions to have unprotected sexual intercourse (MacDonald & Martineau, 2002). MacDonald, T. K., & Martineau, A. M. (2002). Self-esteem, mood, and intentions to use condoms: When does low self-esteem lead to risky health behaviors? *Journal of Experimental Social Psychology*, 38, 299–306. In a carefully designed empirical study, they found that being in a negative mood increased intentions to have unprotected sex—but only for students who were low in self-esteem. Although there are many challenges involved in conducting a study like this, one of the primary ones is the measurement of the relevant variables. In this study, the researchers needed to know whether each of their participants had high or low self-esteem, which of course required measuring their self-esteem. They also needed to be sure that their attempt to put people into a negative mood (by having them think negative thoughts) was successful, which required measuring their moods. Finally, they needed to see whether self-esteem and mood were related to participants' intentions to have unprotected sexual intercourse, which required measuring these intentions.

To students who are just getting started in psychological research, the challenge of measuring such variables might seem insurmountable. Is it really possible to measure things as intangible as self-esteem, mood, or an intention to do something? The answer is a resounding yes, and in this chapter we look closely at the nature of the variables that psychologists study and how they can be measured. We also look at some practical issues in psychological measurement.

Do You Feel You Are a Person of Worth?

The Rosenberg Self-Esteem Scale (Rosenberg, 1989) Rosenberg, M. (1989). *Society and the adolescent self-image* (rev. ed.). Middletown, CT: Wesleyan University Press. is one of the most common measures of self-esteem and the one that MacDonald and Martineau used in their study. Participants respond to each of the 10 items that follow with a rating on a 4-point scale: *Strongly Agree*, *Agree*, *Disagree*, *Strongly Disagree*. Score Items 1, 2, 4, 6, and 7 by assigning 3 points for each *Strongly Agree* response, 2 for each *Agree*, 1 for each *Disagree*, and 0 for each *Strongly Disagree*. Reverse the scoring for Items 3, 5, 8, 9, and 10 by assigning 0 points for each *Strongly Agree*, 1 point for each *Agree*, and so on. The overall score is the total number of points.

1. I feel that I'm a person of worth, at least on an equal plane with others.
2. I feel that I have a number of good qualities.
3. All in all, I am inclined to feel that I am a failure.
4. I am able to do things as well as most other people.
5. I feel I do not have much to be proud of.
6. I take a positive attitude toward myself.
7. On the whole, I am satisfied with myself.
8. I wish I could have more respect for myself.
9. I certainly feel useless at times.
10. At times I think I am no good at all.

5.1 Understanding Psychological Measurement

LEARNING OBJECTIVES

1. Define measurement and give several examples of measurement in psychology.
2. Explain what a psychological construct is and give several examples.
3. Distinguish conceptual from operational definitions, give examples of each, and create simple operational definitions.
4. Distinguish the four levels of measurement, give examples of each, and explain why this distinction is important.

What Is Measurement?

Measurement¹ is the assignment of scores to individuals so that the scores represent some characteristic of the individuals. This very general definition is consistent with the kinds of measurement that everyone is familiar with—for example, weighing oneself by stepping onto a bathroom scale, or checking the internal temperature of a roasting turkey by inserting a meat thermometer. It is also consistent with measurement throughout the sciences. In physics, for example, one might measure the potential energy of an object in Earth's gravitational field by finding its mass and height (which of course requires measuring *those* variables) and then multiplying them together along with the gravitational acceleration of Earth (9.8 m/s^2). The result of this procedure is a score that represents the object's potential energy.

Of course this general definition of measurement is consistent with measurement in psychology too. (Psychological measurement is often referred to as **psychometrics**².) Imagine, for example, that a cognitive psychologist wants to measure a person's working memory capacity—his or her ability to hold in mind and think about several pieces of information all at the same time. To do this, she might use a backward digit span task, where she reads a list of two digits to the person and asks him or her to repeat them in reverse order. She then repeats this several times, increasing the length of the list by one digit each time, until the person makes an error. The length of the longest list for which the person responds correctly is the score and represents his or her working memory capacity. Or imagine a clinical psychologist who is interested in how depressed a person is. He administers the Beck Depression Inventory, which is a 21-item self-report questionnaire in which the person rates the extent to which he or she has felt sad, lost energy, and experienced other symptoms of depression over the past 2 weeks.

1. The assignment of scores to individuals so that the scores represent some characteristic of the individuals.
2. The measurement of psychological variables and constructs.

The sum of these 21 ratings is the score and represents his or her current level of depression.

The important point here is that measurement does not require any particular instruments or procedures. It does not require placing individuals or objects on bathroom scales, holding rulers up to them, or inserting thermometers into them. What it *does* require is *some* systematic procedure for assigning scores to individuals or objects so that those scores represent the characteristic of interest.

Psychological Constructs

Many variables studied by psychologists are straightforward and simple to measure. These include sex, age, height, weight, and birth order. You can almost always tell whether someone is male or female just by looking. You can ask people how old they are and be reasonably sure that they know and will tell you. Although people might not know or want to tell you how much they weigh, you can have them step onto a bathroom scale. Other variables studied by psychologists—perhaps the majority—are not so straightforward or simple to measure. We cannot accurately assess people’s level of intelligence by looking at them, and we certainly cannot put their self-esteem on a bathroom scale. These kinds of variables are called **constructs**³ (pronounced *CON-structs*) and include personality traits (e.g., extroversion), emotional states (e.g., fear), attitudes (e.g., toward taxes), and abilities (e.g., athleticism).

Psychological constructs cannot be observed directly. One reason is that they often represent *tendencies* to think, feel, or act in certain ways. For example, to say that a particular college student is highly extroverted (see [Note 5.6 "The Big Five"](#)) does not necessarily mean that she is behaving in an extroverted way right now. In fact, she might be sitting quietly by herself, reading a book. Instead, it means that she has a general tendency to behave in extroverted ways (talking, laughing, etc.) across a variety of situations. Another reason psychological constructs cannot be observed directly is that they often involve internal processes. Fear, for example, involves the activation of certain central and peripheral nervous system structures, along with certain kinds of thoughts, feelings, and behaviors—none of which is necessarily obvious to an outside observer. Notice also that neither extroversion nor fear “reduces to” any particular thought, feeling, act, or physiological structure or process. Instead, each is a kind of summary of a complex set of behaviors and internal processes.

3. A variable that cannot be observed directly because it represents a tendency to behave in certain ways or a complex pattern of behavior and internal processes. These include personality traits, emotional states, attitudes, and abilities.

The Big Five

The Big Five is a set of five broad dimensions that capture much of the variation in human personality. Each of the Big Five can even be defined in terms of six more specific constructs called “facets” (Costa & McCrae, 1992). Costa, P. T., Jr., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, 4, 5–13.

Figure 5.1
The Big Five Personality Dimensions

Big Five Dimension	Facets					
Openness to Experience	Fantasy	Aesthetics	Feelings	Actions	Ideas	Values
Conscientiousness	Competence	Order	Dutifulness	Achievement Striving	Self-Discipline	Deliberation
Extraversion	Warmth	Gregariousness	Assertiveness	Activity	Excitement Seeking	Positive Emotions
Agreeableness	Trust	Straight-forwardness	Altruism	Compliance	Modesty	Tender-Mindedness
Neuroticism	Worry	Anger	Discouragement	Self-Consciousness	Impulsivity	Vulnerability

The **conceptual definition**⁴ of a psychological construct describes the behaviors and internal processes that make up that construct, along with how it relates to other variables. For example, a conceptual definition of neuroticism (another one of the Big Five) would be that it is people’s tendency to experience negative emotions such as anxiety, anger, and sadness across a variety of situations. This definition might also include that it has a strong genetic component, remains fairly stable over time, and is positively correlated with the tendency to experience pain and other physical symptoms.

Students sometimes wonder why, when researchers want to understand a construct like self-esteem or neuroticism, they do not simply look it up in the dictionary. One reason is that many scientific constructs do not have counterparts in everyday language (e.g., working memory capacity). More important, researchers are in the business of developing definitions that are more detailed and precise—and that more accurately describe the way the world is—than the informal definitions in the dictionary. As we will see, they do this by proposing conceptual definitions, testing them empirically, and revising them as necessary. Sometimes they throw them out altogether. This is why the research literature often includes different conceptual definitions of the same construct. In some cases, an older conceptual definition has

4. A description of a variable or construct in terms of the behaviors and internal processes that are involved, along with how that construct relates to other variables.

been replaced by a newer one that works better. In others, researchers are still in the process of deciding which of various conceptual definitions is the best.

Operational Definitions

An **operational definition**⁵ is a definition of a variable in terms of precisely how it is to be measured. These measures generally fall into one of three broad categories.

Self-report measures⁶ are those in which participants report on their own thoughts, feelings, and actions, as with the Rosenberg Self-Esteem Scale.

Behavioral measures⁷ are those in which some other aspect of participants' behavior is observed and recorded. This is an extremely broad category that includes the observation of people's behavior both in highly structured laboratory tasks and in more natural settings. A good example of the former would be measuring working memory capacity using the backward digit span task. A good example of the latter is a famous operational definition of physical aggression from researcher Albert Bandura and his colleagues (Bandura, Ross, & Ross, 1961). Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology*, 63, 575–582. They let each of several children play for 20 minutes in a room that contained a clown-shaped punching bag called a Bobo doll. They filmed each child and counted the number of acts of physical aggression he or she committed. These included hitting the doll with a mallet, punching it, and kicking it. Their operational definition, then, was the number of these specifically defined acts that the child committed in the 20-minute period. Finally, **physiological measures**⁸ are those that involve recording any of a wide variety of physiological processes, including heart rate and blood pressure, galvanic skin response, hormone levels, and electrical activity and blood flow in the brain.

5. A definition of a variable or construct in terms of precisely how it will be measured.
6. A measure in which participants report on their own thoughts, feelings, and behaviors. Compare with behavioral measure and physiological measure.
7. A measure in which the researcher observes and records some aspect of participants' behavior. Compare with self-report measure and physiological measure.
8. A measure that involves recording a physiological variable. Compare with self-report measure and behavioral measure.

Figure 5.2



In addition to self-report and behavioral measures, researchers in psychology use physiological measures. An electroencephalograph (EEG) records electrical activity from the brain.

Source: Photo courtesy of James McCue

For any given variable or construct, there will be multiple operational definitions. Stress is a good example. A rough conceptual definition is that stress is an adaptive response to a perceived danger or threat that involves physiological, cognitive, affective, and behavioral components. But researchers have operationally defined it in several ways. The Social Readjustment Rating Scale is a self-report questionnaire on which people identify stressful events that they have experienced in the past year and assigns points for each one depending on its severity. For example, a man who has been divorced (73 points), changed jobs (36 points), and had a change in sleeping habits (16 points) in the past year would have a total score of 125. The Daily Hassles and Uplifts Scale is similar but focuses on everyday stressors like misplacing things and being concerned about one's weight. The Perceived Stress Scale is another self-report measure that focuses on people's feelings of stress (e.g., "How often have you felt nervous and stressed?"). Researchers have also

operationally defined stress in terms of several physiological variables including blood pressure and levels of the stress hormone cortisol.

When psychologists use multiple operational definitions of the same construct—either within a study or across studies—they are using **converging operations**⁹. The idea is that the various operational definitions are “converging” on the same construct. When scores based on several different operational definitions are closely related to each other and produce similar patterns of results, this constitutes good evidence that the construct is being measured effectively and that it is useful. The various measures of stress, for example, are all correlated with each other and have all been shown to be correlated with other variables such as immune system functioning (also measured in a variety of ways) (Segerstrom & Miller, 2004). Segerstrom, S. E., & Miller, G. E. (2004). Psychological stress and the human immune system: A meta-analytic study of 30 years of inquiry. *Psychological Bulletin*, 130, 601–630. This is what allows researchers eventually to draw useful general conclusions, such as “stress is negatively correlated with immune system functioning,” as opposed to more specific and less useful ones, such as “people’s scores on the Perceived Stress Scale are negatively correlated with their white blood counts.”

9. Multiple operational definitions of the same construct. When multiple operational definitions are closely related to each other and produce the same pattern of results, this constitutes evidence that the construct is being measured effectively and is a useful one.
10. Four different ways of assigning scores to individuals that provide increasing amounts of quantitative information about the characteristic being measured. The four levels are nominal, ordinal, interval, and ratio.
11. The level of measurement that involves assigning names or category labels to individuals. Scores at the nominal level indicate whether or not one individual is in the same category as another. They do not communicate any quantitative information.

Levels of Measurement

The psychologist S. S. Stevens suggested that scores can be assigned to individuals so that they communicate more or less quantitative information about the variable of interest (Stevens, 1946). Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680. For example, the officials at a 100-m race could simply rank order the runners as they crossed the finish line (first, second, etc.), or they could time each runner to the nearest tenth of a second using a stopwatch (11.5 s, 12.1 s, etc.). In either case, they would be measuring the runners’ times by systematically assigning scores to represent those times. But while the rank ordering procedure communicates the fact that the second-place runner took longer to finish than the first-place finisher, the stopwatch procedure also communicates *how much* longer the second-place finisher took. Stevens actually suggested four different **levels of measurement**¹⁰ (which he called “scales of measurement”) that correspond to four different levels of quantitative information that can be communicated by a set of scores.

The **nominal level**¹¹ of measurement is used for categorical variables and involves assigning scores that are category labels. Category labels communicate whether any two individuals are the same or different in terms of the variable being measured. For example, if you look at your research participants as they enter the room, decide whether each one is male or female, and type this information into a spreadsheet, you are engaged in nominal-level measurement. Or if you ask your

participants to indicate which of several ethnicities they identify themselves with, you are again engaged in nominal-level measurement.

The remaining three levels of measurement are used for quantitative variables. The **ordinal level**¹² of measurement involves assigning scores so that they represent the rank order of the individuals. Ranks communicate not only whether any two individuals are the same or different in terms of the variable being measured but also whether one individual is higher or lower on that variable. The **interval level**¹³ of measurement involves assigning scores so that they represent the precise magnitude of the difference between individuals, but a score of zero does not actually represent the complete absence of the characteristic. A classic example is the measurement of heat using the Celsius or Fahrenheit scale. The difference between temperatures of 20°C and 25°C is precisely 5°, but a temperature of 0°C does not mean that there is a complete absence of heat. In psychology, the intelligence quotient (IQ) is often considered to be measured at the interval level. Finally, the **ratio level**¹⁴ of measurement involves assigning scores in such a way that there is a true zero point that represents the complete absence of the quantity. Height measured in meters and weight measured in kilograms are good examples. So are counts of discrete objects or events such as the number of siblings one has or the number of questions a student answers correctly on an exam.

12. The level of measurement that involves rank ordering individuals. Scores at the ordinal level indicate whether one individual has more or less of the characteristic of interest, but they do not indicate how much more or less.

13. The level of measurement that involves assigning numerical scores so that a given difference between two scores always represents the same difference in the characteristic of interest but a score of zero does not literally represent none of the characteristic. Scores at the interval level indicate how much more or less of the characteristic one individual has than another. Ratios of one score to another are not meaningful at this level.

14. The level of measurement that involves assigning numerical scores so that a given difference between two scores always represents the same difference in the characteristic and a score of zero represents none of the characteristic. Ratios of one score to another are meaningful only at this level.

Stevens's levels of measurement are important for at least two reasons. First, they emphasize the generality of the concept of measurement. Although people do not normally think of categorizing or ranking individuals as measurement, in fact they are as long as they are done so that they represent some characteristic of the individuals. Second, the levels of measurement can serve as a rough guide to the statistical procedures that can be used with the data and the conclusions that can be drawn from them. With nominal-level measurement, for example, the only available measure of central tendency is the mode. Also, ratio-level measurement is the only level that allows meaningful statements about ratios of scores. One cannot say that someone with an IQ of 140 is twice as intelligent as someone with an IQ of 70 because IQ is measured at the interval level, but one can say that someone with six siblings has twice as many as someone with three because number of siblings is measured at the ratio level.

KEY TAKEAWAYS

- Measurement is the assignment of scores to individuals so that the scores represent some characteristic of the individuals. Psychological measurement can be achieved in a wide variety of ways, including self-report, behavioral, and physiological measures.
- Psychological constructs such as intelligence, self-esteem, and depression are variables that are not directly observable because they represent behavioral tendencies or complex patterns of behavior and internal processes. An important goal of scientific research is to conceptually define psychological constructs in ways that accurately describe them.
- For any conceptual definition of a construct, there will be many different operational definitions or ways of measuring it. The use of multiple operational definitions, or converging operations, is a common strategy in psychological research.
- Variables can be measured at four different levels—nominal, ordinal, interval, and ratio—that communicate increasing amounts of quantitative information. The level of measurement affects the kinds of statistics you can use and conclusions you can draw from your data.

EXERCISES

1. Practice: Complete the Rosenberg Self-Esteem Scale and compute your overall score.
2. Practice: Think of three operational definitions for sexual jealousy, decisiveness, and social anxiety. Consider the possibility of self-report, behavioral, and physiological measures. Be as precise as you can.
3. Practice: For each of the following variables, decide which level of measurement is being used.
 - a. A college instructor measures the time it takes his students to finish an exam by looking through the stack of exams at the end. He assigns the one on the bottom a score of 1, the one on top of that a 2, and so on.
 - b. A researcher accesses her participants' medical records and counts the number of times they have seen a doctor in the past year.
 - c. Participants in a research study are asked whether they are right-handed or left-handed.

5.2 Reliability and Validity of Measurement

LEARNING OBJECTIVES

1. Define reliability, including the different types and how they are assessed.
2. Define validity, including the different types and how they are assessed.
3. Describe the kinds of evidence that would be relevant to assessing the reliability and validity of a particular measure.

Again, measurement involves assigning scores to individuals so that they represent some characteristic of the individuals. But how do researchers know that the scores actually represent the characteristic, especially when it is a construct like intelligence, self-esteem, depression, or working memory capacity? The answer is that they conduct research using the measure to confirm that the scores make sense based on their understanding of the construct being measured. This is an extremely important point. Psychologists do not simply *assume* that their measures work. Instead, they collect data to *demonstrate* that they work. If their research does not demonstrate that a measure works, they stop using it.

As an informal example, imagine that you have been dieting for a month. Your clothes seem to be fitting more loosely, and several friends have asked if you have lost weight. If at this point your bathroom scale indicated that you had lost 10 pounds, this would make sense and you would continue to use the scale. But if it indicated that you had gained 10 pounds, you would rightly conclude that it was broken and either fix it or get rid of it. In evaluating a measurement method, psychologists consider two general dimensions: reliability and validity.

Reliability

15. The extent to which the scores on a measure are consistent across time, across multiple items on the same measure, and across researchers when a measure has an element of subjective judgment.

16. The extent to which scores on a measure are consistent across time for the same individuals.

Reliability¹⁵ refers to the consistency of a measure. Psychologists consider three types of consistency: over time (test-retest reliability), across items (internal consistency), and across different researchers (interrater reliability).

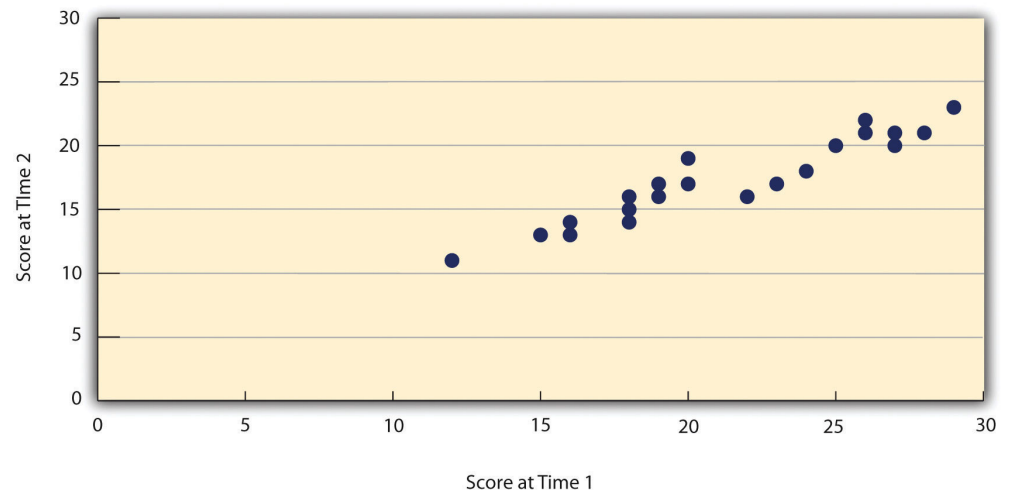
Test-Retest Reliability

When researchers measure a construct that they assume to be consistent across time, then the scores they obtain should also be consistent across time. **Test-retest reliability**¹⁶ is the extent to which this is actually the case. For example,

intelligence is generally thought to be consistent across time. A person who is highly intelligent today will be highly intelligent next week. This means that any good measure of intelligence should produce roughly the same scores for this individual next week as it does today. Clearly, a measure that produces highly inconsistent scores over time cannot be a very good measure of a construct that is supposed to be consistent.

Assessing test-retest reliability requires using the measure on a group of people at one time, using it again on the *same* group of people at a later time, and then looking at **test-retest correlation**¹⁷ between the two sets of scores. This is typically done by graphing the data in a scatterplot and computing Pearson's *r*. **Figure 5.3 "Test-Retest Correlation Between Two Sets of Scores of Several College Students on the Rosenberg Self-Esteem Scale, Given Two Times a Week Apart"** shows the correlation between two sets of scores of several college students on the Rosenberg Self-Esteem Scale, given two times a week apart. Pearson's *r* for these data is +.95. In general, a test-retest correlation of +.80 or greater is considered to indicate good reliability.

Figure 5.3 Test-Retest Correlation Between Two Sets of Scores of Several College Students on the Rosenberg Self-Esteem Scale, Given Two Times a Week Apart



Again, high test-retest correlations make sense when the construct being measured is assumed to be consistent over time, which is the case for intelligence, self-esteem, and the Big Five personality dimensions. But other constructs are not assumed to be stable over time. The very nature of mood, for example, is that it changes. So a measure of mood that produced a low test-retest correlation over a period of a month would not be a cause for concern.

17. The correlation between individuals' scores on a measure used at two different times.

Internal Consistency

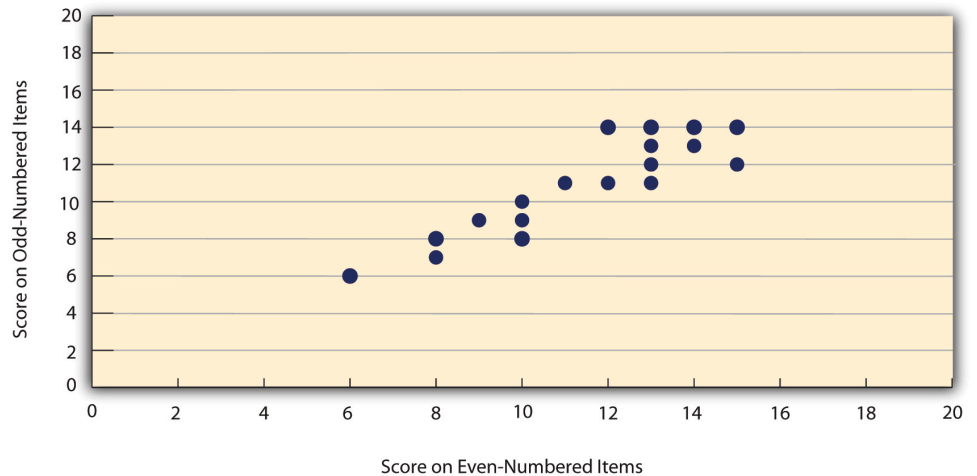
A second kind of reliability is **internal consistency**¹⁸, which is the consistency of people's responses across the items on a multiple-item measure. In general, all the items on such measures are supposed to reflect the same underlying construct, so people's scores on those items should be correlated with each other. On the Rosenberg Self-Esteem Scale, people who agree that they are a person of worth should tend to agree that they have a number of good qualities. If people's responses to the different items are not correlated with each other, then it would no longer make sense to claim that they are all measuring the same underlying construct. This is as true for behavioral and physiological measures as for self-report measures. For example, people might make a series of bets in a simulated game of roulette as a measure of their level of risk seeking. This measure would be internally consistent to the extent that individual participants' bets were consistently high or low across trials.

Like test-retest reliability, internal consistency can only be assessed by collecting and analyzing data. One approach is to look at a **split-half correlation**¹⁹. This involves splitting the items into two sets, such as the first and second halves of the items or the even- and odd-numbered items. Then a score is computed for each set of items, and the relationship between the two sets of scores is examined. For example, [Figure 5.4 "Split-Half Correlation Between Several College Students' Scores on the Even-Numbered Items and Their Scores on the Odd-Numbered Items of the Rosenberg Self-Esteem Scale"](#) shows the split-half correlation between several college students' scores on the even-numbered items and their scores on the odd-numbered items of the Rosenberg Self-Esteem Scale. Pearson's r for these data is $+.88$. A split-half correlation of $+.80$ or greater is generally considered good internal consistency.

18. The extent to which the items on a multiple-item measure are consistent with each other.

19. The correlation between scores based on one half of the items on a multiple-item measure and scores based on the other half of the items.

Figure 5.4 Split-Half Correlation Between Several College Students' Scores on the Even-Numbered Items and Their Scores on the Odd-Numbered Items of the Rosenberg Self-Esteem Scale



Perhaps the most common measure of internal consistency used by researchers in psychology is a statistic called **Cronbach's α** ²⁰ (the Greek letter alpha). Conceptually, α is the mean of all possible split-half correlations for a set of items. For example, there are 252 ways to split a set of 10 items into two sets of five. Cronbach's α would be the mean of the 252 split-half correlations. Note that this is not how α is actually computed, but it is a correct way of interpreting the meaning of this statistic. Again, a value of $+0.80$ or greater is generally taken to indicate good internal consistency.

Interrater Reliability

Many behavioral measures involve significant judgment on the part of an observer or a rater. **Interrater reliability**²¹ is the extent to which different observers are consistent in their judgments. For example, if you were interested in measuring college students' social skills, you could make video recordings of them as they interacted with another student whom they are meeting for the first time. Then you could have two or more observers watch the videos and rate each student's level of social skills. To the extent that each participant does in fact have some level of social skills that can be detected by an attentive observer, different observers' ratings should be highly correlated with each other. If they were not, then those ratings could not be an accurate representation of participants' social skills. Interrater reliability is often assessed using Cronbach's α when the judgments are quantitative or an analogous statistic called **Cohen's κ** ²² (the Greek letter kappa) when they are categorical.

20. A statistic used to assess the internal consistency of a multiple-item measure. It is conceptually equivalent to the mean of all possible split-half correlations.

21. When a measure involves human judgment, the extent to which different observers are consistent in their judgments.

22. A statistic used to assess interrater reliability when the observer judgments are categorical.

Validity

Validity²³ is the extent to which the scores from a measure represent the variable they are intended to. But how do researchers make this judgment? We have already considered one factor that they take into account—reliability. When a measure has good test-retest reliability and internal consistency, researchers should be more confident that the scores represent what they are supposed to. There has to be more to it, however, because a measure can be extremely reliable but have no validity whatsoever. As an absurd example, imagine someone who believes that people’s index finger length reflects their self-esteem and therefore tries to measure self-esteem by holding a ruler up to people’s index fingers. Although this measure would have extremely good test-retest reliability, it would have absolutely no validity. The fact that one person’s index finger is a centimeter longer than another’s would indicate nothing about which one had higher self-esteem.

Textbook presentations of validity usually divide it into several distinct “types.” But a good way to interpret these types is that they are other kinds of evidence—in addition to reliability—that should be taken into account when judging the validity of a measure. Here we consider four basic kinds: face validity, content validity, criterion validity, and discriminant validity.

Face Validity

Face validity²⁴ is the extent to which a measurement method appears “on its face” to measure the construct of interest. Most people would expect a self-esteem questionnaire to include items about whether they see themselves as a person of worth and whether they think they have good qualities. So a questionnaire that included these kinds of items would have good face validity. The finger-length method of measuring self-esteem, on the other hand, seems to have nothing to do with self-esteem and therefore has poor face validity. Although face validity can be assessed quantitatively—for example, by having a large sample of people rate a measure in terms of whether it appears to measure what it is intended to—it is usually assessed informally.

23. The extent to which scores on a measure represent the variable or construct they are intended to. Validity is a judgment based on the available evidence.

24. The extent to which a measure appears “on its face” to measure the variable or construct it is supposed to.

Face validity is at best a very weak kind of evidence that a measurement method is measuring what it is supposed to. One reason is that it is based on people’s intuitions about human behavior, which are frequently wrong. It is also the case that many established measures in psychology work quite well despite lacking face validity. The Minnesota Multiphasic Personality Inventory (MMPI) measures many personality characteristics and disorders by having people decide whether each of over 567 different statements applies to them—where many of the statements do not have any obvious relationship to the construct that they measure. Another

example is the Implicit Association Test, which measures prejudice in a way that is nonintuitive to most people (see [Note 5.31 "How Prejudiced Are You?"](#)).

How Prejudiced Are You?

The Implicit Association Test (IAT) is used to measure people's attitudes toward various social groups. The IAT is a behavioral measure designed to reveal negative attitudes that people might not admit to on a self-report measure. It focuses on how quickly people are able to categorize words and images representing two contrasting groups (e.g., gay and straight) along with other positive and negative stimuli (e.g., the words "wonderful" or "nasty"). The IAT has been used in dozens of published research studies, and there is strong evidence for both its reliability and its validity (Nosek, Greenwald, & Banaji, 2006). Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2006). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). London, England: Psychology Press. You can learn more about the IAT—and take several of them for yourself—at the following website: <https://implicit.harvard.edu/implicit>.

Content Validity

Content validity²⁵ is the extent to which a measure “covers” the construct of interest. For example, if a researcher conceptually defines test anxiety as involving both sympathetic nervous system activation (leading to nervous feelings) and negative thoughts, then his measure of test anxiety should include items about both nervous feelings and negative thoughts. Or consider that attitudes are usually defined as involving thoughts, feelings, and actions toward something. By this conceptual definition, a person has a positive attitude toward exercise to the extent that he or she thinks positive thoughts about exercising, feels good about exercising, and actually exercises. So to have good content validity, a measure of people's attitudes toward exercise would have to reflect all three of these aspects. Like face validity, content validity is not usually assessed quantitatively. Instead, it is assessed by carefully checking the measurement method against the conceptual definition of the construct.

Criterion Validity

Criterion validity²⁶ is the extent to which people's scores²⁶ on a measure are correlated with other variables (known as **criteria**²⁷) that one would expect them to

25. The extent to which a measure covers all aspects of the construct it is supposed to measure.

26. The extent to which scores on a measure are correlated with other variables and constructs that they are expected to be correlated with, given the conceptual definition of the construct being measured.

27. A variable or construct expected to be correlated with scores on a measure that is being evaluated. The plural is *criteria*.

be correlated with. For example, people's scores on a new measure of test anxiety should be negatively correlated with their performance on an important school exam. If it were found that people's scores were in fact negatively correlated with their exam performance, then this would be a piece of evidence that these scores really represent people's test anxiety. But if it were found that people scored equally well on the exam regardless of their test anxiety scores, then this would cast doubt on the validity of the measure.

A criterion can be any variable that one has reason to think should be correlated with the construct being measured, and there will usually be many of them. For example, one would expect test anxiety scores to be negatively correlated with exam performance and course grades and positively correlated with general anxiety and with blood pressure during an exam. Or imagine that a researcher develops a new measure of physical risk taking. People's scores on this measure should be correlated with their participation in "extreme" activities such as snowboarding and rock climbing, the number of speeding tickets they have received, and even the number of broken bones they have had over the years. Criteria can also include other measures of the same construct. For example, one would expect new measures of test anxiety or physical risk taking to be positively correlated with existing measures of the same constructs. So the use of converging operations is one way to examine criterion validity.

Assessing criterion validity requires collecting data using the measure. Researchers John Cacioppo and Richard Petty did this when they created their self-report Need for Cognition Scale to measure how much people value and engage in thinking (Cacioppo & Petty, 1982). Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42, 116–131. In a series of studies, they showed that college faculty scored higher than assembly-line workers, that people's scores were positively correlated with their scores on a standardized academic achievement test, and that their scores were negatively correlated with their scores on a measure of dogmatism (which represents a tendency toward obedience). In the years since it was created, the Need for Cognition Scale has been used in literally hundreds of studies and has been shown to be correlated with a wide variety of other variables, including the effectiveness of an advertisement, interest in politics, and juror decisions (Petty, Briñol, Loersch, & McCaslin, 2009). Petty, R. E, Briñol, P., Loersch, C., & McCaslin, M. J. (2009). The need for cognition. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 318–329). New York, NY: Guilford Press.

28. The extent to which scores on a measure are not correlated with other variables and constructs that are conceptually distinct.

Discriminant Validity

Discriminant validity²⁸ is the extent to which scores on a measure are not correlated with measures of variables that are conceptually distinct. For example,

self-esteem is a general attitude toward the self that is fairly stable over time. It is not the same as mood, which is how good or bad one happens to be feeling right now. So people's scores on a new measure of self-esteem should not be very highly correlated with their moods. If the new measure of self-esteem were highly correlated with a measure of mood, it could be argued that the new measure is not really measuring self-esteem; it is measuring mood instead.

When they created the Need for Cognition Scale, Cacioppo and Petty also provided evidence of discriminant validity by showing that people's scores were not correlated with certain other variables. For example, they found only a weak correlation between people's need for cognition and a measure of their cognitive style—the extent to which they tend to think analytically by breaking ideas into smaller parts or holistically in terms of “the big picture.” They also found no correlation between people's need for cognition and measures of their test anxiety and their tendency to respond in socially desirable ways. All these low correlations provide evidence that the measure is reflecting a conceptually distinct construct.

KEY TAKEAWAYS

- Psychological researchers do not simply assume that their measures work. Instead, they conduct research to show that they work. If they cannot show that they work, they stop using them.
- There are two distinct criteria by which researchers evaluate their measures: reliability and validity. Reliability is consistency across time (test-retest reliability), across items (internal consistency), and across researchers (interrater reliability). Validity is the extent to which the scores actually represent the variable they are intended to.
- Validity is a judgment based on various types of evidence. The relevant evidence includes the measure's reliability, whether it covers the construct of interest, and whether the scores it produces are correlated with other variables they are expected to be correlated with and not correlated with variables that are conceptually distinct.
- The reliability and validity of a measure is not established by any single study but by the pattern of results across multiple studies. The assessment of reliability and validity is an ongoing process.

EXERCISES

1. Practice: Ask several friends to complete the Rosenberg Self-Esteem Scale. Then assess its internal consistency by making a scatterplot to show the split-half correlation (even- vs. odd-numbered items). Compute Pearson's r too if you know how.
2. Discussion: Think back to the last college exam you took and think of the exam as a psychological measure. What construct do you think it was intended to measure? Comment on its face and content validity. What data could you collect to assess its reliability, criterion validity, and discriminant validity?
3. Practice: Take an Implicit Association Test and then list as many ways to assess its criterion validity as you can think of.

5.3 Practical Strategies for Psychological Measurement

LEARNING OBJECTIVES

1. Specify the four broad steps in the measurement process.
2. Explain how you would decide whether to use an existing measure or create your own.
3. Describe multiple strategies to identify and locate existing measures of psychological constructs.
4. Describe several general principles for creating new measures and for implementing existing and new measures.
5. Create a simple plan for assessing the reliability and validity of an existing or new measure.

So far in this chapter, we have considered several basic ideas about the nature of psychological constructs and their measurement. But now imagine that you are in the position of actually having to measure a psychological construct for a research project. How should you proceed? Broadly speaking, there are four steps in the measurement process: (a) conceptually defining the construct, (b) operationally defining the construct, (c) implementing the measure, and (d) evaluating the measure. In this section, we will look at each of these steps in turn.

Conceptually Defining the Construct

Having a clear and complete conceptual definition of a construct is a prerequisite for good measurement. For one thing, it allows you to make sound decisions about exactly how to measure the construct. If you had only a vague idea that you wanted to measure people's "memory," for example, you would have no way to choose whether you should have them remember a list of vocabulary words, a set of photographs, a newly learned skill, or an experience from long ago. Because psychologists now conceptualize memory as a set of semi-independent systems, you would have to be more precise about what you mean by "memory." If you are interested in long-term declarative memory (memory for facts), then having participants remember a list of words that they learned last week would make sense, but having them remember and execute a newly learned skill would not. In general, there is no substitute for reading the research literature on a construct and paying close attention to how others have defined it.

Deciding on an Operational Definition

Using an Existing Measure

It is usually a good idea to use an existing measure that has been used successfully in previous research. Among the advantages are that (a) you save the time and trouble of creating your own, (b) there is already some evidence that the measure is valid (if it has been used successfully), and (c) your results can more easily be compared with and combined with previous results. In fact, if there already exists a reliable and valid measure of a construct, other researchers might expect you to use it unless you have a good and clearly stated reason for not doing so.

If you choose to use an existing measure, you may still have to choose among several alternatives. You might choose the most common one, the one with the best evidence of reliability and validity, the one that best measures a particular aspect of a construct that you are interested in (e.g., a physiological measure of stress if you are most interested in its underlying physiology), or even the one that would be easiest to use. For example, the Ten-Item Personality Inventory (TIPI) is a self-report questionnaire that measures all the Big Five personality dimensions with just 10 items (Gosling, Rentfrow, & Swann, 2003). Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality*, 37, 504–528. It is not as reliable or valid as longer and more comprehensive measures, but a researcher might choose to use it when testing time is severely limited.

When an existing measure was created primarily for use in scientific research, it is usually described in detail in a published research article and is free to use in your own research—with a proper citation. You might find that later researchers who use the same measure describe it only briefly but provide a reference to the original article, in which case you would have to get the details from the original article. The American Psychological Association also publishes the *Directory of Unpublished Experimental Measures*, which is an extensive catalog of measures that have been used in previous research. Many existing measures—especially those that have applications in clinical psychology—are proprietary. This means that a publisher owns the rights to them and that you would have to purchase them. These include many standard intelligence tests, the Beck Depression Inventory, and the Minnesota Multiphasic Personality Inventory (MMPI). Details about many of these measures and how to obtain them can be found in other reference books, including *Tests in Print* and the *Mental Measurements Yearbook*. There is a good chance you can find these reference books in your college or university library.

Creating Your Own Measure

Instead of using an existing measure, you might want to create your own. Perhaps there is no existing measure of the construct you are interested in or existing ones are too difficult or time-consuming to use. Or perhaps you want to use a new measure specifically to see whether it works in the same way as existing measures—that is, to demonstrate converging operations. In this section, we consider some general issues in creating new measures that apply equally to self-report, behavioral, and physiological measures. More detailed guidelines for creating self-report measures are presented in [Chapter 9 "Survey Research"](#).

First, be aware that most new measures in psychology are really variations of existing measures, so you should still look to the research literature for ideas. Perhaps you can modify an existing questionnaire, create a paper-and-pencil version of a measure that is normally computerized (or vice versa), or adapt a measure that has traditionally been used for another purpose. For example, the famous Stroop task (Stroop, 1935) Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662.—in which people quickly name the colors that various color words are printed in—has been adapted for the study of social anxiety. Socially anxious people are slower at color naming when the words have negative social connotations such as “stupid” (Amir, Freshman, & Foa, 2002). Amir, N., Freshman, M., & Foa, E. (2002). Enhanced Stroop interference for threat in social phobia. *Journal of Anxiety Disorders*, 16, 1–9.

When you create a new measure, you should strive for simplicity. Remember that your participants are not as interested in your research as you are and that they will vary widely in their ability to understand and carry out whatever task you give them. You should create a set of clear instructions using simple language that you can present in writing or read aloud (or both). It is also a good idea to include one or more practice items so that participants can become familiar with the task, and to build in an opportunity for them to ask questions before continuing. It is also best to keep the measure brief to avoid boring or frustrating your participants to the point that their responses start to become less reliable and valid.

The need for brevity, however, needs to be weighed against the fact that it is nearly always better for a measure to include multiple items rather than a single item. There are two reasons for this. One is a matter of content validity. Multiple items are often required to cover a construct adequately. The other is a matter of reliability. People’s responses to single items can be influenced by all sorts of irrelevant factors—misunderstanding the particular item, a momentary distraction, or a simple error such as checking the wrong response option. But when several responses are summed or averaged, the effects of these irrelevant factors tend to cancel each other out to produce more reliable scores. Remember, however, that

multiple items must be structured in a way that allows them to be combined into a single overall score by summing or averaging. To measure “financial responsibility,” a student might ask people about their annual income, obtain their credit score, and have them rate how “thrifty” they are—but there is no obvious way to combine these responses into an overall score. To create a true multiple-item measure, the student might instead ask people to rate the degree to which 10 statements about financial responsibility describe them on the same five-point scale.

Finally, the very best way to assure yourself that your measure has clear instructions, includes sufficient practice, and is an appropriate length is to test several people. (Family and friends often serve this purpose nicely). Observe them as they complete the task, time them, and ask them afterward to comment on how easy or difficult it was, whether the instructions were clear, and anything else you might be wondering about. Obviously, it is better to discover problems with a measure before beginning any large-scale data collection.

Implementing the Measure

You will want to implement any measure in a way that maximizes its reliability and validity. In most cases, it is best to test everyone under similar conditions that, ideally, are quiet and free of distractions. Testing participants in groups is often done because it is efficient, but be aware that it can create distractions that reduce the reliability and validity of the measure. As always, it is good to use previous research as a guide. If others have successfully tested people in groups using a particular measure, then you should consider doing it too.

Be aware also that people can react in a variety of ways to being measured that reduce the reliability and validity of the scores. Although some disagreeable participants might intentionally respond in ways meant to “mess up” a study, participant **reactivity**²⁹ is more likely to take the opposite form. Agreeable participants might respond in ways they believe they are expected to. They might engage in **socially desirable responding**³⁰. For example, people with low self-esteem agree that they feel they are a person of worth not because they really feel this way but because they believe this is the socially appropriate response and do not want to look bad in the eyes of the researcher. Additionally, research studies can have built-in **demand characteristics**³¹: cues to how the researcher expects participants to behave. For example, a participant whose attitude toward exercise is measured immediately after she is asked to read a passage about the dangers of heart disease might reasonably conclude that the passage was meant to improve her attitude. As a result, she might respond more favorably because she believes she is expected to by the researcher. Finally, your own expectations can bias participants’ behaviors in unintended ways.

29. Participants’ reactions to the fact that they are being measured.

30. Participants’ responding in ways they believe to be socially appropriate rather than in ways that reflect their actual thoughts, feelings, and behavior.

31. Features of a study that cue participants as to how the researcher expects them to behave.

There are several precautions you can take to minimize these kinds of reactivity. One is to make the procedure as clear and brief as possible so that participants are not tempted to take out their frustrations on your results. Another is to guarantee participants' anonymity and make clear to them that you are doing so. If you are testing them in groups, be sure that they are seated far enough apart that they cannot see each other's responses. Give them all the same type of writing implement so that they cannot be identified by, for example, the pink glitter pen that they used. You can even allow them to seal completed questionnaires into individual envelopes or put them into a drop box where they immediately become mixed with others' questionnaires. Although informed consent requires telling participants what they will be doing, it does not require revealing your hypothesis or other information that might suggest to participants how you expect them to respond. A questionnaire designed to measure financial responsibility need not be titled "Are You Financially Responsible?" It could be titled "Money Questionnaire" or have no title at all. Finally, the effects of your expectations can be minimized by arranging to have the measure administered by a helper who is unaware of its intent or of any hypothesis being tested. Regardless of whether this is possible, you should standardize all interactions between researchers and participants—for example, by always reading the same set of instructions word for word.

Evaluating the Measure

Once you have used your measure on a sample of people and have a set of scores, you are in a position to evaluate it more thoroughly in terms of reliability and validity. Even if the measure has been used extensively by other researchers and has already shown evidence of reliability and validity, you should not assume that it worked as expected for your particular sample and under your particular testing conditions. Regardless, you now have additional evidence bearing on the reliability and validity of the measure, and it would make sense to add that evidence to the research literature.

In most research designs, it is not possible to assess test-retest reliability because participants are tested at only one time. For a new measure, you might design a study specifically to assess its test-retest reliability by testing the same set of participants at two times. In other cases, a study designed to answer a different question still allows for the assessment of test-retest reliability. For example, a psychology instructor might measure his students' attitude toward critical thinking using the same measure at the beginning and end of the semester to see if there is any change. Even if there is no change, he could still look at the correlation between students' scores at the two times to assess the measure's test-retest reliability. It is also customary to assess internal consistency for any multiple-item measure—usually by looking at a split-half correlation or Cronbach's alpha.

Criterion and discriminant validity can be assessed in various ways. For example, if your study included more than one measure of the same construct or measures of conceptually distinct constructs, then you should look at the correlations among these measures to be sure that they fit your expectations. Note also that a successful experimental manipulation also provides evidence of criterion validity. Recall that MacDonald and Martineau manipulated participant's moods by having them think either positive or negative thoughts, and after the manipulation their mood measure showed a distinct difference between the two groups. This simultaneously provided evidence that their mood manipulation worked *and* that their mood measure was valid.

But what if your newly collected data cast doubt on the reliability or validity of your measure? The short answer is that you have to ask why. It could be that there is something wrong with your measure or how you administered it. It could be that there is something wrong with your conceptual definition. It could be that your experimental manipulation failed. For example, if a mood measure showed no difference between people whom you instructed to think positive versus negative thoughts, maybe it is because the participants did not actually think the thoughts they were supposed to or that the thoughts did not actually affect their moods. In short, it is “back to the drawing board” to revise the measure, revise the conceptual definition, or try a new manipulation.

KEY TAKEAWAYS

- Good measurement begins with a clear conceptual definition of the construct to be measured. This is accomplished both by clear and detailed thinking and by a review of the research literature.
- You often have the option of using an existing measure or creating a new measure. You should make this decision based on the availability of existing measures and their adequacy for your purposes.
- Several simple steps can be taken in creating new measures and in implementing both existing and new measures that can help maximize reliability and validity.
- Once you have used a measure, you should reevaluate its reliability and validity based on your new data. Remember that the assessment of reliability and validity is an ongoing process.

EXERCISES

1. Practice: Write your own conceptual definition of self-confidence, irritability, and athleticism.
2. Practice: Choose a construct (sexual jealousy, self-confidence, etc.) and find two measures of that construct in the research literature. If you were conducting your own study, which one (if either) would you use and why?