

Chapter 5

Measurement

Operational Definitions

Numbers and Precision

Scales of Measurement

Nominal Scale

Ordinal Scale

Interval Scale

Ratio Scale

Validity of Measurement

Content Validity

Face Validity

Concurrent Validity

Predictive Validity

Construct Validity

Thinking Critically About Everyday Information

Reliability of Measurement

Test-Retest Reliability

Alternate Form Reliability

Split-Half Reliability

Factors That Affect Reliability

Case Analysis

General Summary

Detailed Summary

Key Terms

Review Questions/Exercises

Operational Definitions

An essential component of an operational definition is measurement. A simple and accurate definition of **measurement** is the assignment of numbers to a variable in which we are interested. These numbers will provide the raw material for our statistical analysis.

Measurement is so common and taken for granted that we seldom ask why we measure things or worry about the different forms that measurement may take. It is often not sufficient to describe a runner as “fast,” a basketball player as “tall,” a wrestler as “strong,” or a baseball hitter as “good.” If coaches recruited potential team members on the basis of these imprecise words, they would have difficulty holding down a job. Coaches want to know how fast the runner runs the 100-yard dash or the mile. They want to know exactly how tall the basketball player is, the strength of the wrestler, the batting average of the hitter. Measurement is a way of refining our ordinary observations so that we can assign numerical values to our observations. It allows us to go beyond simply describing the presence or absence of an event or thing to specifying how much, how long, or how intense it is. With measurement, our observations become more accurate and more reliable.

Precision is important in all areas of our lives, especially in the sciences and technologies, and we look for ways of increasing it. Here is an interesting classroom demonstration of the precision of numbers versus the precision of words. Ask the class members to write down on a piece of paper what number the word “several” represents to them. Gather the responses and then plot them on the board. You will be surprised at the wide range of numbers represented by the word (it usually ranges from 2 to 7).

How often have you been in an argument with a friend, only to find out after much debate that you are using key words in different ways? The argument is one of *semantics* rather than of issues. You defined the word one way, and your friend defined it a different way. This experience is more common among laypersons than among scientists, but it still occurs. Before the merits of an issue or a position can be discussed, there must be agreement about the meaning of the important terms. The same is true in science. If we are to avoid confusion and misinterpretation, we must be able to communicate unambiguously the meaning of such terms as *intelligence*, *anxiety*, *altruism*, *hostility*, *love*, *alienation*, *aggression*, *guilt*, *reinforcement*, *frustration*, *memory*, and *information*. These terms have all been used scientifically, in very precise ways. Each of these terms could be given a dictionary definition, usually referred to as a literary or conceptual definition. But dictionary definitions are not sufficiently precise for many scientific terms because they are too general and often too ambiguous. When a word is to be used scientifically or technically, its precise meaning must be conveyed—it must be clear and unambiguous. We achieve this clarity of meaning by operationally defining the term. To state the operations for a term means to make the term observable by pointing to how it is measured. An **operational definition**, then, makes the concept observable by stating what the scientist does to measure it.

For example, *anxiety* could be defined in dictionary terms as “a state of being uneasy, apprehensive, or worried.” An operational definition of the term could include observable measures such as sweating palms (observable as sweat gland activity), increased heart rate (observable with heartbeat recording), dilated pupils, and other observable physiological changes. It could also be a self-rating scale or a paper-and-pencil questionnaire. We could in each case specify the precise amounts of each measure necessary for our operational definition of anxiety.

As another example, consider the hypothesis that we proposed in the last chapter. We hypothesized that the effect of TV violence on older children’s aggressive behavior at school will be less if the characters are not human. Although this appears to be a clear statement, more specific operational definitions would be necessary before any research could be undertaken to test the hypothesis. The researcher must make several decisions. What is violence on TV? Certainly, one character killing another character would be considered violence. What about a shove or push? What about a verbal assault? What about when Wile E. Coyote falls off the cliff and is hit in the head with a rock? What constitutes a character that is not human? We could probably agree that Wiley Coyote fits this category. What about a computer-animated person? How will aggressive behavior at school be defined? Of course, getting into a fight would be aggressive behavior. What about profanity directed toward another student or teacher? What about little Johnny chasing Mary on the playground? Notice that there are no correct answers to these questions. However, the researcher must decide what is going to be meant by each of the variables in a particular study and be able to communicate those operational definitions to those who will be consumers of the research findings.

Table 5.1 contains both dictionary definitions and operational definitions of some common terms. Note that in each case, the operational definition refers to events that are observable or events that can easily be made observable. Note further that the definition is very specific rather than general.

Table 5.1 Dictionary and Operational Definitions of Several Terms Commonly Used by Psychologists

For each concept, other operational definitions are possible.		
TERM	DICTIONARY DEFINITION	OPERATIONAL DEFINITION
Punishment	Harsh or injurious treatment for an offense	Presentation of 3 milliamp shock for .5 second following certain (specified) behavior
Learning	Acquiring knowledge or skill	Change in behavior (specify kind of behavior) as a function of practice
Anxiety	State of being uneasy, apprehensive, or worried	Sweat gland activity (amount), heart rate (amount), physiological changes (specify), self-reported anxiety on a scale of 1 to 7
Intelligence	Ability to learn or understand from experience	Score on the Stanford–Binet Intelligence Test, score on the Wechsler Intelligence Scale for Children
Thirst	Distressful feeling caused by a desire or need for water	Eighteen hours (or other value) without access to water
Sleep	Recurring condition of rest, no conscious thought, eyes closed, etc.	Specific brain wave frequencies (EEG) for different sleep stages
Guilt	A painful feeling of self-reproach	Score on a personality inventory, self-reported guilt on a scale of 1 to 10

The feature that determines whether a particular definition is more useful than another is whether it allows us to discover meaningful laws about behavior. Some will, and some will not. Those definitions that are helpful to our understanding of behavior will be retained; those that do not will be discarded. The first step in the life of a concept is to define it in clearly unambiguous, observable terms. It then may or may not be useful. If the concept of intelligence were defined as “the distance between the ears,” or “the circumference of the head,” its meaning would be clear, but it is very doubtful that it would ever become useful.

Let’s look at one additional point before leaving the topic of definitions. An operational definition, or any other kind of definition, is not an explanation. When definitions are unintentionally used as explanations, we label them as **tautological** or **circular reasoning**. Circular reasoning has little value. A definition doesn’t explain behavior or provide you with information that will, in and of itself, help in understanding behavior. It is a necessary step in discovering lawful relations, but it is only one side of a two-sided law. To explain behavior, two independent (different) types of observation are necessary: one is observations that relate to the independent variable (variable manipulated by the experimenter or “cause”), and the second is observations that relate to the dependent variable (behavior of participant or “effect”). When the relationship between the independent and dependent variables is predictable, we say

that we have a lawful relationship. A circular argument uses only one side of the relationship—only one of these observations. For example, suppose we observe two children fighting with each other (body contact with intent to harm). We may be tempted to say they are fighting because they are hostile children, because hostility leads to fighting. To this point, we have not explained anything. All we have is an operational definition of hostility as fighting behavior. Our argument would be a tautology (circular) if we said that the children are fighting because they are hostile and then said that we know that they are hostile because they are fighting. To avoid circularity and to explain the behavior, we would have to define hostility and fighting independently and show that the operations for defining hostility do in fact give rise to fighting.

Tautological reasoning occurs with a higher frequency than it should. For example, it is not uncommon to hear the statement “Individuals who commit suicide are mentally ill.” To the question “How do you know they are mentally ill?” the response is often “Because they committed suicide.” Another common tautology refers to musical ability. For example, it is said “Individuals who play the piano well do so because they have musical ability.” To the question “How do you know they have musical ability?” the response is “Because they play the piano well.” Another example is “Individuals drink excessively because they are alcoholics. We know that they are alcoholics because they drink excessively.” We repeat, tautological arguments do not advance our knowledge. To avoid circularity in our last example, we would have to define what we mean by “drinks excessively” and then identify the factors that give rise to drinking excessively—for example, genetics, specific early experiences, or stressful events. We then would have an explanation for the drinking.

Numbers and Precision

As noted earlier, measurement scales are important because they allow us to transform or substitute precise numbers for imprecise words. We are restricted in what we can do with words but less so with numbers. Numbers permit us to perform certain activities and operations that words do not. In many instances, numbers permit us to add, multiply, divide, or subtract. They also permit the use of various statistical procedures. These statistics, in turn, result in greater precision and objectivity in describing behavior or other phenomena. At a minimum, we know that the numbers 1, 2, 3, 4, and so on, when applied to the frequency of occurrence of any event, mean that 4 instances are more than 3, which in turn are more than 2, and so on. Contrast numbers with words such as *frequently*, *often*, or *many times*. Does an event occurring *frequently* occur a greater or fewer number of times than an event occurring *often*? It may be true that a given individual uses the two terms *frequently* and *often* consistently across situations; another individual may also use the two terms consistently, but in reverse order. The result would be confusion.

The use of numbers rather than words increases our precision in communicating in other ways also. Finer distinctions (discriminations) can often be achieved with numbers if the distinctions can be made reliably. Instead of saying a certain behavior was either present or absent, or occurred with high, medium, or low frequency, numbers permit us to say, more precisely, how frequently the behavior occurred. Words are often too few in number to allow us to express finer distinctions.

Our number system is an abstract system of symbols that has little meaning in and of itself. It becomes meaningful when it becomes involved in measurement. As noted earlier, measurement is the process of assigning numbers to objects and events in accordance with a set of rules. To grasp the full impact of measurement, we need to understand the concept of a measurement scale. There are several different kinds of scales: nominal, ordinal, interval, and ratio. The distinction among scales becomes of particular importance when we conduct statistical analyses of data. Underlying statistical tests are various assumptions, including those relating to the scale of measurement. In other words, the scale of measurement for a variable can determine the most appropriate type of statistical analysis of the data.

Scales of Measurement

Nominal Scale

There has been some disagreement among experts whether a **nominal scale** should even be described as a scale. Most would agree that it should. The fact is that we do name things, and this naming permits us to do other things as a result. The word *nominal* is derived from the Latin word for *name*. With a nominal scale, numbers are assigned to objects or events simply for identification purposes. For example, participants in various sports have numbers on their jerseys that quickly allow spectators, referees, and commentators to identify them. This identification is the sole purpose of the numbers. Performing arithmetic operations on these numbers, such as addition, subtraction, multiplication, or division, would not make any sense. The numbers do not indicate more or less of any quantity. A baseball player with the number 7 on his back does not necessarily have more of something than a player identified by the number 1. Other examples include your social security number, your driver's license number, or your credit card number. Labeling or naming allows us to make qualitative distinctions or to categorize and then count the frequency of persons, objects, or things in each category. This activity can be very useful. For example, in any given voting year, we could label or name individuals as Democrat or Republican, Liberal or Conservative, and then count frequencies for the purpose of predicting voting outcomes. Other examples of nominal scales used for identifying and categorizing are male–female, violent show–nonviolent show, and punishment–reward. As you will see later, a chi-square statistic is appropriate for data derived from a categorical (nominal) scale.

Ordinal Scale

An **ordinal scale** allows us to rank-order events. Original numbers are assigned to the order, such as first, second, third, and so on. For example, we might determine that runners in a race finished in a particular order, and this order would provide us with useful information. We would know that the runner finishing first (assigned a value of 1) ran the distance faster than the runner finishing second (assigned a value of 2), that the second-place finisher ran faster than the third-place finisher (assigned a value of 3), and so on. However, we would not know how much faster the first runner was than the second-place runner, or the second compared with the third. The difference between the first- and second-place runners may have been a fraction of a second, or it could have been several seconds. Similarly, the difference between the second- and third-place runners could have been very small or very large. An ordinal scale does not convey precise quantitative information. With an ordinal scale, we know the rank order, but we do not have any idea of the distance or interval between the rankings. Some other examples of ordinal scales are grades such as “A,” “B,” “C,” “D,” and “F”; scores given in terms of high, medium, and low; birth order in terms of firstborn, second-born, or later-born; a list of examination scores from highest to lowest; a list of job candidates ranked from high to low; and a list of the ten best-dressed persons.

What about the common use of Likert-type scales in behavioral research? For example, a researcher may pose a question to a teacher as follows:

How aggressive has Johnny been in your classroom this week?

Not at all		Somewhat		Very
1	2	3	4	5

Although most psychological scales are probably ordinal, psychologists assume that many of the scales have equal intervals and act accordingly. In other words, the difference in level of aggression between a score of 1 and a score of 2 is about the same as the difference in level of aggression between a score of 2 and a score of 3, and so on. Many researchers believe that these scales do approximate equality of intervals reasonably well, and it is unlikely that this assumption will lead to serious difficulties in interpreting our findings.

Interval Scale

When we can specify both the order of events *and* the distance between events, we have an **interval scale**. The distance between any two intervals on this type of scale is equal throughout the scale. The central shortcoming of an interval scale is its lack of an absolute zero point—a location where the user

can say that there is a complete absence of the variable being measured. This type of scale often has an arbitrary zero point, sometimes called an *anchor point*. An example may make clear the difference between an arbitrary zero point and an absolute zero point. Scores on intelligence tests are considered to be on an interval scale. With intelligence test scores, the anchor point is set at a mean IQ value of 100 with a standard deviation (SD) of 15. A score of 115 is just as far above the mean (one SD) as a score of 85 is below the mean (one SD). Because we have a relative zero point and not an absolute one, we cannot say that a person with an IQ of 120 is twice as intelligent as a person with an IQ of 60. It is simply not meaningful to do so. Some additional examples of interval scales are both the centigrade and Fahrenheit scales of temperature, altitude (zero is sea level rather than the center of the earth), and scores on a depression scale or an anxiety scale. Students often confuse historical time. Is the year 2000 twice as old as the year 1000? The answer is no. Why?

Ratio Scale

A **ratio scale** has a number of properties that the others do not. With ratio scales, we can identify rank order, equal intervals, *and* equal ratios—two times as much, one-half as much. Ratios can be determined because the zero point is absolute, a true anchor—the complete absence of a property. Zero weight or height means the complete absence of weight or height. A 100-pound person has one-half the weight of a 200-pound person and twice the weight of a 50-pound person. We can say these things because we know that the starting points for these dimensions or measures is 0. It is important to notice that it is not necessary for any research participant to obtain a score of 0, only that it exists on the scale. Obviously no research participant would receive a weight score of 0!

A ratio scale is common when the researcher is counting the number of events. For example, you might measure a child's aggressive behavior by counting the number of times that the child inflicts physical harm on another person during a one-week observation period. Clearly, 10 incidents would be twice as many as 5, and 0 incidents would represent the absence of the variable you are measuring. Frequency counts that represent the number of times that a particular event occurred are a common example of measurement on a ratio scale. But be careful not to confuse this use of *frequency* with the use of *frequency* as a summary statistic for data measured on a nominal scale (how many times observations fit a particular category).

Table 5.2 provides additional examples of each scale of measurement.

Table 5.2 Examples of Variables for Each Scale of Measurement

Nominal Label or category	Type of disorder (schizophrenia, depression, anxiety) Religious affiliation (none, Catholic, Protestant, Jewish, other) Region of the country (Northeast, Midwest, Southwest, etc.) Eye color (blue, brown, hazel, green) Flavor of ice cream (vanilla, chocolate, strawberry) Type of reinforcer (food, water, money, compliment)
Ordinal Rank order	College classification (freshman, sophomore, junior, senior) Grade on a test (A, B, C, D, F) National ranking of a sports team (1st, 2nd, 3rd, 4th, 5th, etc.) Reaction time (fastest, 2nd fastest, 3rd fastest, 4th fastest, etc.) Intensity of light (very bright, bright, dim, none) Age classification (child, teen, young adult, adult, older adult)
Interval Rank order + Equal intervals	Difference between the mean test score and each student's score Score on the Beck Anxiety Scale (scores range from 0 to 44, but note that a score of 0 does not represent a complete absence of anxiety) Score on a Likert-type scale (1, 2, 3, 4, 5, 6, 7) Temperature measured in degrees Celsius or degrees Fahrenheit Weight measured on a scale not calibrated to zero
Ratio Rank order + Equal intervals + Absolute zero	Number of college credits completed Number of correct answers on a test Number of total points scored in a season by a sports team Reaction time measured in milliseconds Intensity of light measured in lumens A person's age measured in years

Determining the scale of measurement for a variable is often a challenging task for students learning about behavioral research. One technique that you may find useful is to remember three questions and use the decision tree shown in Figure 5.1.

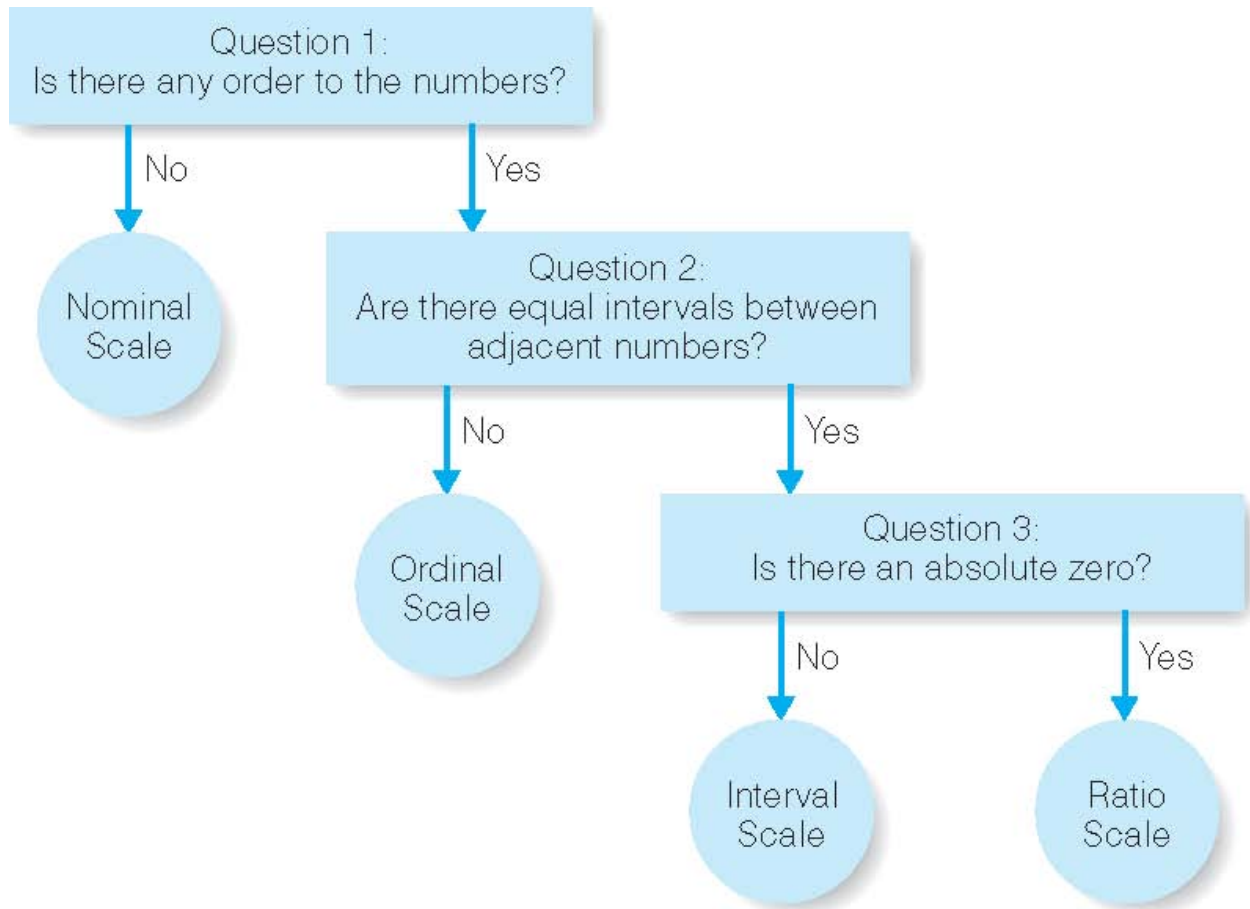


Figure 5.1 Decision tree to determine the appropriate scale of measurement.

Validity of Measurement

Two important concepts relate to the measuring instruments that we use: validity and reliability. In a general sense, **validity** refers to whether the measuring instrument does what it is intended to do, and **reliability** refers to how consistently it does so. It is critical that our measurements of variables be both valid and reliable. If they are not, then we cannot be sure of what we are measuring.

The validity and reliability of a test are established by evidence. Does the SAT (Scholastic Aptitude Test) measure students' ability to do academic work (question of validity), and if so, how consistently does it do so (question of reliability)? Do those students scoring high on the SAT perform predictably better in college than those scoring low (validity)? If a student took the SAT today and then again six months from today, would the score on the two performances be similar (reliability)? We should note that a test may be highly reliable, but if it does not relate to our behavior of interest (validity), it is useless for predicting that behavior. The length of the big toe can be measured with a high degree of consistency, but it is unlikely to be a valid predictor of academic success. We can measure many things very reliably, but

unless they enter into meaningful behavioral relationships (lawful relationships), they are not useful. Thus, we can achieve reliability without having validity. However, the reverse is not possible. In order to be valid, a measure must be reliable.

Let us look at five types of validity that are commonly distinguished: content validity, face validity, concurrent validity, predictive validity, and construct validity (see Figure 5.2). Content validity and face validity relate to tests of current knowledge and skills, whereas predictive validity relates to future performance on a job or task. Concurrent validity assesses similarity to other current measures, and construct validity deals with evaluating theoretical terms.

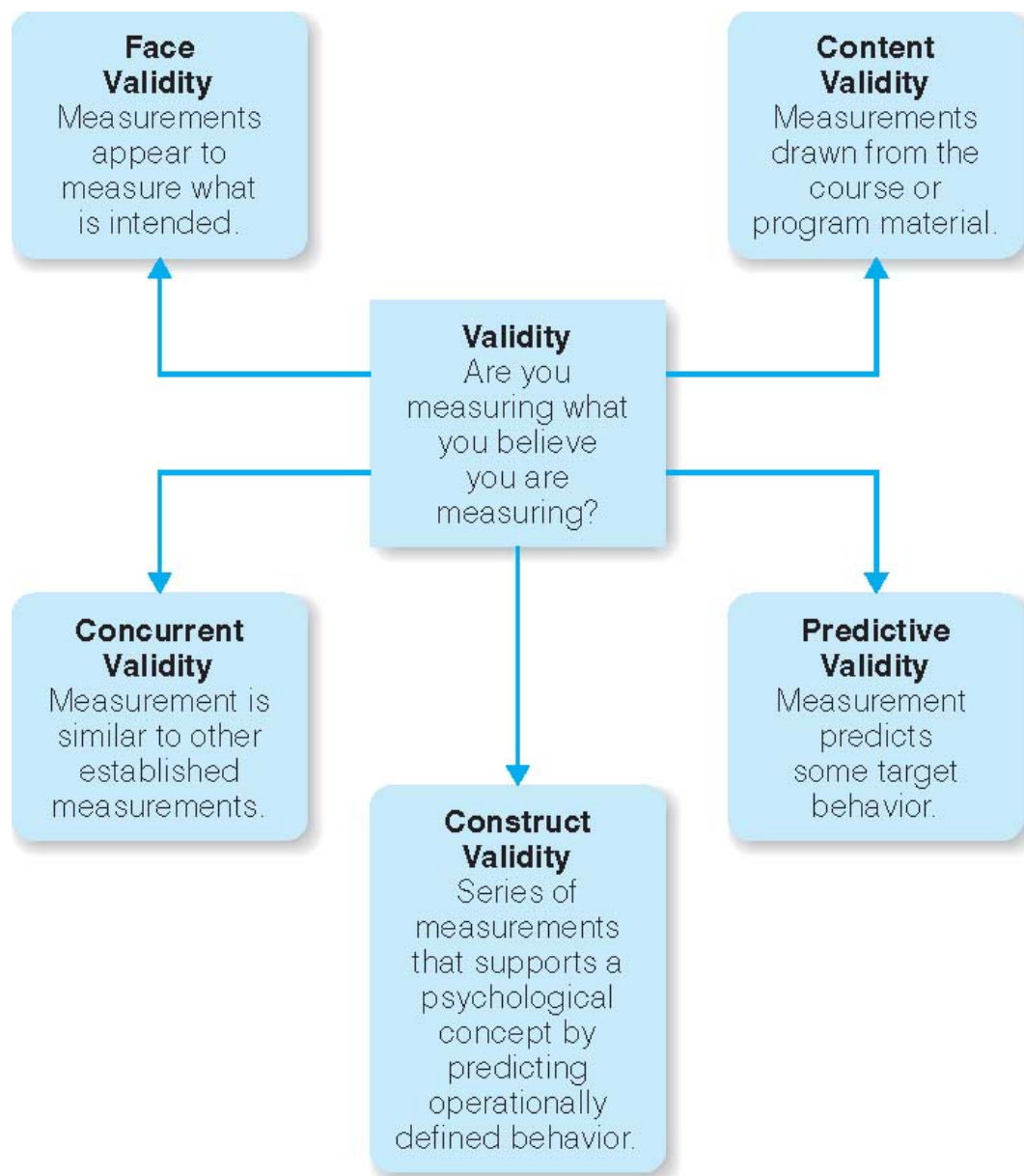


Figure 5.2 Ways to assess validity.

Content Validity

With **content validity**, we are interested in assessing current performance rather than predicting future performance. A test is constructed to measure whether participants in a program have mastered the contents of the program. Course examinations—whether midterm, finals, or some other kind—deal with content validity. Developers of social programs that require training of personnel are concerned about the effectiveness of the training. Thus, they frequently rely on content validity. *Content validity is achieved by constructing a test drawn directly from material in a program or course.* There can be disagreements regarding the representativeness of the test materials. For example, students taking an exam sometimes complain that it did not cover the material they were required to read. If this were true, the exam would be low in content validity. Although not always possible, we could assure representativeness by constructing a complete list of the content for the program or course and then selecting test items randomly from the material. Thus, content validity is based on prior considerations about what content will be included.

Face Validity

Face validity is similar to content validity, but it is determined *after* the test is constructed. We ask ourselves whether or not the test appears, on the face of it, to measure what we intend it to measure. If the test is known to have content validity, face validity can be assumed. However, it does not work in reverse direction; that is, face validity does not ensure content validity. The face validity of an exam in research methods would be high if the questions appeared to deal with research methods. However, without actual knowledge of the course materials, there would be no assurance that the questions were representative of the actual course material.

Concurrent Validity

Often, we measure psychological constructs for which there is a long history of investigation. Such constructs include intelligence, aggression, fear, attention, depression, anxiety, and many others. One way to assess the validity of our instrument is to determine whether it has **concurrent validity**—that is, whether it provides measurements that are similar to those provided by other instruments, previously established as valid, designed to measure the same construct. If our new measure of aggression results in scores that are the same as or similar to those found with a traditional measure of aggression, then we can be reasonably confident that we are measuring the same construct. For example, if the authors developed a new Lammers–Badia Depression Scale, we would want individuals to complete our instrument and also to complete an established instrument such as the Beck Depression Inventory (BDI). If our scale in fact

measures depression, then individuals who score high (depressed) on our scale ought to score high on the BDI as well.

Predictive Validity

Predictive validity, also called criterion validity, is often used in program evaluation studies and is very suitable for applied research. Essentially, a test is constructed and developed for the purpose of predicting some form of behavior. For example, score on the SAT (the predictor) is moderately correlated with academic performance in college (the criterion). Knowing this relationship and an individual's score gives us a better idea of how a student will perform in college than we would have in the absence of this information. The higher the correlation between the criterion and the predictor, the greater the predictive validity. Indeed, if the correlation is perfect (1.00), the prediction is also perfect. However, most of the time correlations are only modest, somewhere between .30 and .60. Nevertheless, they are considered useful. It is very difficult to construct a good predictor measure; it is also very difficult to obtain a good criterion measure.

Construct Validity

Construct validity is one of the most important and also the most difficult to understand. We can only touch upon it here. Construct validity is especially important for the construction of theories. Researchers create theoretical constructs to better understand, explain, and predict behavior. Many hypotheses that are tested in research are derived from constructs, or theories of behavior. The construct or theory predicts how people should behave under certain conditions. The degree to which the results support the theory is a measure of construct validity. Examples of some constructs are frustration, aggression, motivation, learning, fear, hopelessness, creativity, and intelligence. Constructs are created by describing the procedures for measuring them. If the instrument for measuring them assists us in understanding and predicting behavior, they are considered valid. If they do not serve these purposes, their validity is rejected. Notice how constructs are related to the notion of operational definitions discussed at the beginning of this chapter. Constructs are general terms that can be operationally defined in different ways. For example, the construct of aggression could be operationally defined as the number of behaviors that lead to physical harm, or it could be defined as the number of physical and verbal threats. Thus, a construct attempts to summarize or explain a set of operational definitions.

Construct validity asks whether a particular measure actually measures the construct that it is intended to measure. Establishing construct validity involves several steps. Generally, the researcher or theorist designs a test and specifies what it is intended to measure. The test is then tied to, or related to,

the relevant theory. Then predictions are made that relate behavior to the test scores. Finally, data are gathered to ascertain whether the predictions are confirmed.

Let's consider two related examples of psychological constructs and how construct validity was assessed. In the late 1960s, Martin Seligman conducted a series of studies with dogs that led him to develop the psychological construct of learned helplessness. He observed that dogs that had been exposed to situations in which they could not avoid an unpleasant event would, in later situations, not even attempt to remove or avoid an unpleasant event even when it was possible (Seligman & Maier, 1967). This finding led to the hypothesis (theory) that early lack of control over environmental events can be debilitating, both cognitively and physiologically. That is, when individuals are subjected to aversive environmental events over which they have no control, the result is learned helplessness, which will be reflected in impaired learning, reduced levels of motivation, poor performance, and physiological deterioration. An extensive series of studies in other animal species (including humans) since the late 1960s supports the notion that the construct of learned helplessness (predictor) does lead to behavioral evidences of impaired learning, motivation, and so on (criterion). This correlation between the predictor and the criterion provides evidence of construct validity. In more recent years, Seligman has developed the related construct of learned optimism (Seligman, 1998) and the Seligman Attributional Style Questionnaire (SASQ) to measure optimism. The construct validity of learned optimism has been demonstrated in numerous studies (reported to be over 500) that show a relationship between optimism scores on the SASQ (predictor) and measures of success, performance, motivation, and physical well-being (criterion variables). Again, such research evidence supports the notion that the psychological construct of optimism is useful in understanding, explaining, and predicting behavior.

What we are suggesting through these examples is that constructs reflect basic behavioral processes. If learned helplessness and learned optimism exist, then certain behaviors should follow. If our expectations are confirmed a number of times in a variety of settings, our construct is useful—that is, valid. Usually, many studies are necessary to demonstrate construct validity. With time and continued research, both the construct and the theory usually undergo a series of modifications and become more refined.

Before turning to the issue of reliability, take a look at the box “Thinking Critically About Everyday Information” and review several concepts presented thus far in the chapter.

Thinking Critically About Everyday Information: Understanding Sleep Apnea

What follows is an article that appeared in the *New York Times* newspaper.

John Holman of St. Paul is a very busy man, and he was not accustomed to having to “let things go” because he was too tired to do them. But tired he was, tired driving to and from work, tired during the day at the warehouse company he runs, too tired some days to play his beloved game of tennis after work, and really tired in the evening—in bed by 8:30 to arise at 6.

But Mr. Holman, an average-size 67-year-old, did nothing about his fatigue until his wife, Marna, threatened to move out of their bedroom because his snoring was keeping her awake. It was she who suggested that he might have sleep apnea. With it, breathing stops, often for a minute or longer, sometimes hundreds of times a night and resumes each time with a loud snort or snore. So at his wife’s urging, Mr. Holman spent a night in the sleep laboratory at Abbott Northwestern Hospital in Minneapolis. Her suspicions were confirmed. Mr. Holman now sleeps with a device that assists his breathing, and both he and his wife, who feared he would fall asleep while driving and kill himself or someone else, are feeling a lot better.

The National Institutes of Health estimates that as many as 18 million Americans, 6 percent of the nation, have sleep apnea. After the proliferation of sleep centers around the country and greater awareness of the condition, the diagnosis of sleep apnea increased twelvefold from 1990 to 1998.

Still, experts estimate that fewer than 10 percent of people who have it are aware of it. As a result, they risk their health and their lives and possibly the lives of others.

Sleep apnea is a disorder that occurs only in sleep and more in REM (rapid eye movement) sleep than in the other stages of sleep. When this disorder is present, the individual stops breathing during sleep for 10 seconds or longer but can breathe normally when awake. Cessation of breathing may occur hundreds of times during the night, with no memory of it in the morning. To determine whether the disorder is present requires the individual to spend a night at a sleep disorders clinic where brain waves (EEG) and respiratory activity can be monitored using a polygraph machine. The EEG pattern reveals when the individual is asleep, how long it takes the person to fall asleep, and the various sleep stages that he or she passes through during the night. Measurement of respiration reveals how frequently an individual has stopped breathing and for how long. Respiration is measured in two ways during the night. One is with sensors near the nose and mouth to measure airflow. The second way is with a respiratory belt attached to diaphragm/chest to measure breathing effort. People with sleep apnea wake up very tired in the morning and are very sleepy all day long. They frequently fall asleep while driving, watching television, reading, or sitting in a meeting. As mentioned, they are unaware that they have sleep apnea, even though they may have awakened hundreds of times during the night. The most prominent symptoms are daytime sleepiness, mood changes, and irritability. If the condition is left untreated, other health problems emerge.

Effective treatments for sleep apnea are available. Any effective treatment would have to address the symptoms noted. One way that sleepiness is measured is with a self-rating scale. Similar rating scales are used for mood changes and irritability. Consider the following questions:

- Is sleep apnea a quantitative or qualitative event?
- What kind of measurement scale would be involved in measuring severity of sleep apnea?
- What measurement scale would be involved if sleepiness, irritability, and mood were each rated on a self-report scale from 1 to 10 (1 = not at all sleepy, 10 = extremely sleepy)?
- In addition to a self-report assessment of sleepiness, can you think of another way to operationally define sleepiness that would be less subjective (perhaps on a ratio scale)?
- How might you show concurrent validity for the psychological construct of sleepiness?

- How might you show predictive validity for the psychological construct of sleepiness?

SOURCE: Jane E. Brody, "Sleep Apnea: A Noisy but Often Invisible Threat," *New York Times*, 17 September 2002, p. F7. Copyright 2002 The New York Times Company.

Reliability of Measurement

A measuring instrument is reliable if measurements recorded at different times give similar results. Obviously, we would not want to use a measuring instrument if it did not give similar results under similar circumstances. Consistency is imperative if we are to obtain useful data. As we have previously noted, reliability is a necessary but not a sufficient condition for validity. Information regarding test reliability should be available for every test or measure. This information should specify the extent of the reliability, including the type of reliability and the conditions under which it was obtained. Reliability estimates are expressed in terms of a correlation coefficient and, therefore, are not all-or-none scores. As we noted, correlations can vary from -1.0 to 1.0 , but most reliability coefficients remain in the 0.0 to 1.0 range.

The upper limits of a test of reliability are established by the similarity of items in terms of what they have in common (interitem correlation) and the number of items on the test. As long as the test items provide independent measures of the same construct, the larger the number of items on a test, the more reliable the test will be. This notion is reflected in the way many sports championships are determined (as in baseball, basketball, and hockey), using a series of several games rather than just one. On any given day, a weaker team may defeat a stronger team, but it is unlikely that a weaker team will consistently beat a stronger team. Similarly, witness the Olympic decathlon. It is a contest in which an athlete must take part in ten events. Consider another example. A course in psychology at the end of the term may have available 1,000 questions that could be used on the final exam. The final exam will be constructed by sampling these items. A sample of only 5 items would be too few for reliability purposes. As we added items, our reliability would increase rapidly, but beyond a certain point, reached rather quickly, adding more items would add very little to our reliability. Thus, a 50-item test would be considerably more reliable than a 5-item test.

For those interested in pursuing this notion further, an estimate of reliability that considers both interitem correlation and the number of items on a test is referred to as the *coefficient alpha* or the *Kuder-Richardson Method*. Among other things, this method will allow you to determine the number of items needed to achieve a given value of reliability.

Next, we discuss three specific methods to assess reliability (see Figure 5.3).

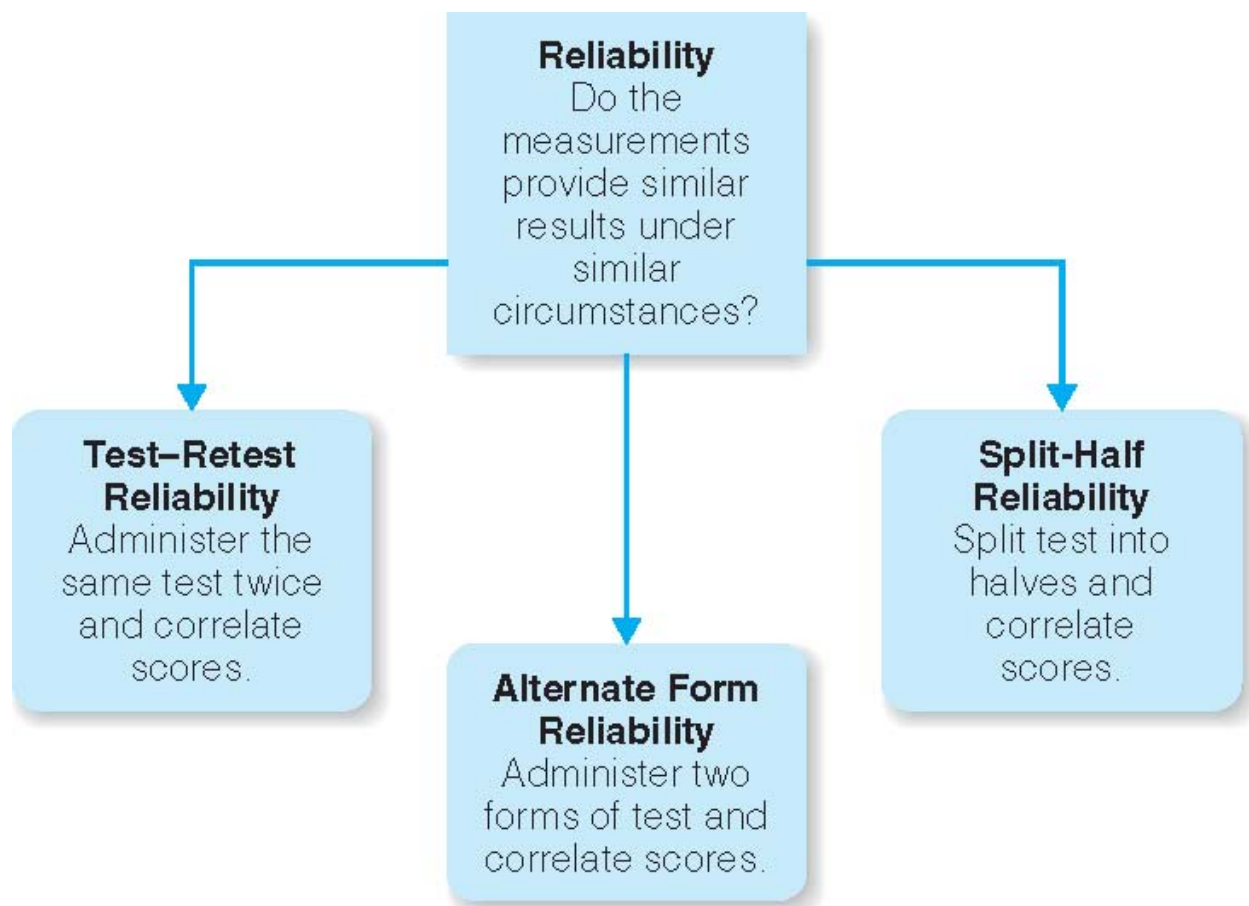


Figure 5.3 Ways to assess reliability.

Test-Retest Reliability

One obvious way to determine the reliability of a test is to administer the same test twice to the same individual. To establish **test-retest reliability**, we administer the identical test a second time, usually after a week or two has gone by. We then calculate a correlation coefficient for the first and second test scores. The time interval between the tests must be specified because retest correlations decrease as the time interval between tests increases. Some problems with this procedure relate to possible practice effects. Also, if the interval between the tests is very short, the individual may be able to remember previous answers given the first time, and the scores between the two sets of test scores would not be independent.

Take a look at the scores shown in Table 5.3. These scores represent a test-retest assessment for two different tests. Although high test-retest reliability ($r = 0.98$) suggests that both instruments are relatively free of measurement error, care should be taken in interpreting this coefficient. It is evident in Table 5.3 that high reliability does not mean that an individual's first and second scores are nearly the same; this may or may not be the case. Glancing over the pairs of scores for Test A, you do get the impression that

the individuals scored similarly on both test administrations. However, glancing over the pairs of scores for Test B, you can see that the similarity in scores is much less and that, overall, the scores on the second administration are higher. Because test–retest reliability is derived through a correlational analysis, the strength of the relationship depends upon the similarity in rank order on the first and second test distributions—that is, whether the individual was first, second, third, and so on, on the two distributions of test scores. In this regard, test–retest reliability is a relative measure. It is possible to have high test–retest reliability and yet have different absolute scores on the first and second testing. Although this situation is unusual, the occurrence of some significant event between the two test administrations could cause a shift in scores.

TEST TAKER	TEST A		TEST B	
	1st Administration	2nd Administration	1st Administration	2nd Administration
Ryan	76	78	76	87
Beth	84	83	84	92
Brandi	85	85	85	94
Chandler	67	64	67	73
John	86	89	86	98
Mary	92	91	92	100
Casey	71	75	71	84
Nina	58	55	58	64
Amy	63	60	63	69
	$r = 0.98$		$r = 0.98$	

We also want to comment on low test–retest reliability. Low reliability does not always suggest a faulty measuring instrument. If a significant event occurs between the two test administrations for some of the test takers but not all, then low test–retest reliability may result. For example, a therapy or training program for a reading disability may take place between the first and second testing periods. If those with the greatest disability showed the greatest improvement, then this would lower test–retest reliability. However, if no deliberate effort to change the condition of individuals was made and only a short period of time intervened between the first and second tests, then low test–retest reliability is most likely due to a faulty measuring instrument.

Alternate Form Reliability

Some of the difficulties with the test–retest procedure can be avoided by using **alternate form reliability**. With this method, an individual is tested on one form of the test (Form A) and then again on a comparable second form (Form B). Again, a correlation coefficient is computed between the scores on Forms A and B. The period of time between the two tests is usually a week or two, although this time varies considerably. Usually, the two forms contain the same number of items and cover similar content. Alternate form reliability is used more frequently than test–retest reliability because it has fewer associated problems, including a substantial reduction in practice effects.

Split-Half Reliability

Only a single test session is involved when using the method of **split-half reliability**. Two scores for each individual are obtained by splitting the test into comparable halves. This is usually achieved by assigning odd-numbered items to one form and even-numbered items to the other (odd–even split). Generally, this method is not preferred over others because the scores on the two halves are not independent. To illustrate, if a test taker is “spaced out” while taking the test, both scores will be depressed. With a sufficient number of instances of this sort, a spuriously high correlation will result, and the reliability will appear to be higher than it really is.

Factors That Affect Reliability

A number of factors can lower the reliability of a test. Within a testing period, individuals may become upset or ill, or they may misread the question, make a clerical error when recording an answer, or guess at the answer. Between testing sessions, an individual may change, there may be differences between the two tests (alternate form), or scoring criteria may change. It is also important to remember that reliability is measured using correlation coefficients. One factor that can reduce correlations is a restricted range of scores on one or both variables. A restricted range of scores can result from a testing instrument that does not allow a variety of possible scores or from testing samples of individuals who all score very similarly on the instrument (either very high or very low). For example, you would not want to assess the reliability of a depression scale by using a sample of suicidal individuals!

Experimental procedures are available to address these factors that affect reliability. In addition to selecting an appropriate sample of research participants, researchers can reduce measurement error and improve reliability by writing the items carefully. They should follow a standardized procedure for giving the test so that testing conditions are very similar. They should state instructions clearly so that they are easily understood. Finally, they should use objective scoring

procedures to avoid error due to subjectivity and changing criteria. We should also restate that longer tests (given the same interitem correlation) are more reliable than shorter tests. Obviously, when new tests are being constructed, considerable “pilot testing” (preliminary selection of items, testing, revising) is necessary before a useful instrument can be developed.

Acceptable reliability levels depend on what the test is used for. For basic research purposes, reliabilities of approximately .70 are not uncommon, but correlations of .80 or higher are considered desirable. To obtain higher estimates of reliability requires considerable effort in selecting items, standardizing the procedure, and administering and scoring the test. For applied problems dealing with social issues such as placement in a retraining program or a special class, or for awards, measurement error must be minimized. Under these and similar circumstances, reliability coefficients less than .90 are usually not acceptable.

We should note that this discussion has focused on the reliability of testing instruments. Often in behavioral research, human observers record behavior, and their observations provide the data for analysis. The reliability of researcher observations is also an important issue and involves the degree to which multiple observers record the same observations. This notion of interobserver agreement (interrater reliability) will be discussed in the next chapter, along with other issues of observation.

Case Analysis

Universities continually examine ways to increase the academic success of their students. One factor may be the level of emphasis on academics in the student’s life. Thus, you believe that students who experience an out-of-class living environment that emphasizes academics will be more successful in college. Your university has several housing options for incoming freshmen. You work with the housing office to assign some freshmen to live in off-campus apartments, some freshmen to live in on-campus residence halls, and some freshmen to live in on-campus residence halls called residential colleges. These residential colleges emphasize academics by having faculty live in the residence hall, by holding classes in the residence hall, by creating a peer tutoring system in the residence hall, and by creating a program of activities that revolve around academics. At the end of their freshman year, you assess the students’ study skills and grade point average. You assess study skills by asking the students 63 true/false questions related to actions and attitudes toward studying. The study habits score is the number of questions answered positively.

Critical Thinking Questions

1. Identify your independent variable(s).

2. Identify your dependent variable(s).
3. What is your hypothesis?
4. What is your operational definition of academic emphasis?
5. What is your operational definition of success in college?
6. Your independent variable is measured on what scale of measurement?
7. Study skills are measured on what scale of measurement?
8. What criterion could be measured to support the predictive validity of the study habits instrument?
9. How would you test the reliability of the study habits instrument?

General Summary

Many of the concepts that we study in psychology are subject to different interpretations. Therefore, the researcher must provide operational definitions that define the variables in specific and measurable terms. For the purpose of analysis, the observations that are made must be assigned numerical values. These numbers that result from the measurement process possess particular characteristics defined by the scale of measurement. Numbers on a nominal scale are simply labels with no greater-than or less-than qualities. Numbers on an ordinal scale indicate greater than and less than, but do not tell us how much greater or less. Numbers on an interval scale have equal intervals between adjacent numbers, but the scale has no absolute zero. A ratio scale has order, equal intervals, and an absolute zero.

Quality research involves measurements that are valid and reliable. Validity refers to confidence that you are measuring what you intended to measure, and reliability refers to consistency in your measurements. Validity is composed of content validity, face validity, concurrent validity, predictive validity, and construct validity. Reliability can be assessed with test–retest, alternative forms, and split-half methods. Although all of these concepts may seem a bit tedious, the good researcher knows that attention to them is directly related to confidence that the results will indeed answer the research question.

Now that we understand some of the issues regarding the measurement of variables, the next chapter will focus on the methods we use to collect such data.

Detailed Summary

1. Measurement is a way of refining our observations so that we can assign numerical values to them. Measurement requires precise definitions of psychological variables such as intelligence, anxiety, altruism, hostility, love, alienation, aggression, guilt, reinforcement, frustration, and memory.
2. Operational definitions provide precision by defining psychological variables in terms of specific operations. For example, hunger could be defined as the number of calories consumed. Operational

definitions are useful when they allow us to discover lawful relationships among variables.

Operational definitions, by themselves, are not explanations. Using definitions as explanations results in inappropriate tautological reasoning.

3. The conversion of observations to numbers permits the use of mathematical operations to better describe and analyze observations. The numbers that represent observations exist on one of four scales of measurement: nominal, ordinal, interval, or ratio. The scale of measurement is important because it determines the types of mathematical operations and statistical analyses that are appropriate.
4. On a nominal scale of measurement, numbers simply represent labels and have no quantitative meaning (for example, religious affiliation). On an ordinal scale of measurement, numbers represent rank order but without equal intervals between adjacent numbers (for example, letter grade on a test). An interval scale has equal intervals between adjacent numbers but no absolute zero (for example, score on a Likert-type scale). A ratio scale of measurement has equal intervals and an absolute zero (for example, number of correct test questions).
5. Validity refers to whether the instrument measures what it is intended to measure. Validity can be assessed in different ways.
6. Content validity is the degree to which measurement items are drawn from the content of the course or program to be evaluated. Face validity is the degree to which measurement items appear (on the face of it) to accurately represent the content to be evaluated. Concurrent validity is the degree to which measurements correlate with measurements from an established (validated) instrument. Predictive validity is the degree to which measurements predict some target behavior. Construct validity is the degree to which measurements support a psychological concept that helps to explain and predict behavior.
7. In order for measurements to be valid, they must be reliable; that is, they should provide similar results under similar circumstances.
8. Test-retest reliability is the degree to which measurements are similar when the same instrument is administered a second time. Alternate form reliability is the degree to which measurements are similar when a comparable version of the instrument is administered. Split-half reliability is the degree to which measurements are similar when different halves of the same instrument are compared.
9. Factors that can lower the reliability of a test include a restricted range of scores, feeling ill, misreading questions, making clerical errors, guessing at answers, changes in the test takers, differences between two tests (alternate form), or changes in scoring criteria.

Key Terms

alternate form reliability

concurrent validity

construct validity

content validity

face validity

interval scale

measurement

nominal scale

operational definition

ordinal scale

predictive validity

ratio scale

reliability

split-half reliability

tautological (circular) reasoning

test–retest reliability

validity

Review Questions / Exercises

1. Provide an operational definition for each of the following variables: hunger, reaction time, academic success, college year (such as freshman or sophomore), fear, and type of automobile.
2. For each of the variables in question 1, describe the types of numbers that would be used to measure these variables, and identify the appropriate scale of measurement.
3. Search a library database for a study that addresses the validity of an instrument or construct. Describe which methods were used to assess validity.
4. Search a library database for a study that addresses the reliability of an instrument. Describe which methods were used to assess reliability.
5. In your own words, describe why validity and reliability of measurement are important in behavioral research.