CHAPTER 8

Tools of Research

Operational definitions transform constructs into observable measures.

INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1. Explain the role of measurement in research.
2. Access sources such as *Mental Measurements Yearbook* and *Tests in Print* to obtain information necessary for evaluating standardized tests and other measuring instruments.
3. State the difference between a test and a scale.
4. Distinguish between norm-referenced and criterion-referenced tests.
5. Distinguish between measures of aptitude and achievement.
6. Distinguish between ceiling effect and floor effect and discuss why these may be of concern.
7. Describe the steps to follow in preparing a Likert scale for measuring attitudes.
8. Define performance assessment and discuss its advantages and disadvantages.
9. Describe the characteristics of a bipolar adjective scale.
10. State the kinds of errors that are common to rating scales.
11. State advantages and disadvantages of self-report personality measures.
12. List at least five guidelines that a researcher should follow when using direct observation as a data-gathering technique.
13. Define a situational test, and tell when it might be used in research.
14. State the essential characteristic of a projective technique and name at least two well-known projective techniques.

One aim of quantitative research is to obtain greater understanding of relationships among variables in populations. For example, you might ask, What is the relationship between intelligence and creativity among 6-year-olds? You cannot directly observe either intelligence or creativity. Nor can you directly observe all 6-year-olds. But this does not mean that you must remain in ignorance about this and similar questions. There are observable behaviors that are accepted as being valid indicators of constructs such as intelligence and creativity. Using indicators to approximate constructs is the measurement aspect of research.

Some measurement is very straightforward, using a single indicator to represent a variable. For example, you could measure a person's educational background by asking about the highest grade he or she had completed. Similarly, such variables as grade level, nationality, marital status, or number of children could be measured by a single indicator simply because these variables refer to phenomena that are very clear and for which a single indicator provides an acceptable measure. Other variables, however, are more complex and much more difficult to measure. In these cases, using a single indicator is not appropriate.

Selecting appropriate and useful measuring instruments is critical to the success of any research study. One must select or develop scales and instruments that can measure complex constructs such as intelligence, achievement, personality, motivation, attitudes, aptitudes, interests, and self-concept. There are two basic ways to obtain these measures for your study: Use one that has already been developed or construct your own.

To select a measuring instrument, the researcher should look at the research that has been published on his or her question to determine what other researchers have used to measure the construct of interest. These reports will generally indicate whether the instrument worked well or whether other procedures might be better. Other useful sources for identifying published instruments for one's research purposes are the *Seventeenth Mental Measurements Yearbook* (Geisinger, Spies, Carlson, & Plake, 2007) and a companion volume, *Tests in Print VII* (Murphy, Plake, & Spies, 2006). Each edition of *Tests in Print* provides an index of all known commercially available tests in print at the time, with information on publisher and date of publication. A subject index helps one to locate tests in a specific category. The Buros Center for Testing website (www.unl.edu/buros) allows you to examine a large amount of information on tests and testing. Once you locate an available test, you then consult the *Mental Measurements Yearbook* for more information and a critical review of the test. The "Test Reviews Online," a service of the Buros Center for Testing, provides reviews exactly as they appear in the *Mental Measurements Yearbook* series. Another good source of information about both published and unpublished tests is the Educational Testing Service (ETS) Test Collection. The ETS Test Collection is a library of more than 20,000 commercial and research tests and other measuring devices designed to provide up-to-date test information to educational researchers. It is available on the web (www.ets.org/testcoll). ETS also has the collection *Tests in Microfiche*, which provides not only an index of unpublished tests but also copies of the tests on microfiche.

If researchers cannot find a previously developed instrument, then they must develop their own. The procedure involves identifying and using behavior that can be considered an indicator of the construct. To locate these indicators, researchers should turn first to the theory behind the construct. A good theory generally suggests how the construct will manifest itself and the changes that can be observed; that is, it suggests ways to measure the construct(s). For example, the general (*g* factor) theory of intelligence influenced the choice of tasks in the construction of early intelligence tests. Shavelson, Huber, and Stanton's (1976) multidimensional theory of self-concept served as the blueprint for a number of self-concept measures that have had a major influence on both theory and classroom practice. For instance, the Shavelson model was the basis for Marsh's (1988) widely used SDQ (Self-Description Questionnaire), which measures self-concept in preadolescents, adolescents, and late adolescents/young adults. Following construction of an instrument, additional research is used to support or revise both the instrument and the theory upon which it is based. Researchers can also use their own experiences and expertise to decide on the appropriate indicators of the construct. In this chapter, we briefly discuss some of the main types of measuring instruments that are used in educational research: achievement and aptitude tests, personality tests, attitude scales, and observational techniques.

TESTS

Tests are valuable measuring instruments for educational research. A **test** is a set of stimuli presented to an individual in order to elicit responses on the basis of which a numerical score can be assigned. This score, based on a representative sample of the individual's behavior, is an indicator of the extent to which the subject has the characteristic being measured.

The utility of these scores as indicators of the construct of interest is in large part a function of the objectivity, validity, and reliability of the tests. Objectivity is the extent of agreement among scorers. Some tests, such as multiple-choice and true–false tests, are described as objective because the scoring is done by comparing students' answers with the scoring key, and scorers need make no decisions. Essay tests are less objective because scores are influenced by the judgment and opinions of the scorers. In general, validity is the extent to which a test measures what it claims to measure. Reliability is the extent to which the test measures accurately and consistently. We discuss validity and reliability in Chapter 9.

ACHIEVEMENT TESTS

Achievement tests are widely used in educational research, as well as in school systems. They are used to measure what individuals have learned. Achievement tests measure mastery and proficiency in different areas of knowledge by presenting subjects with a standard set of questions involving completion of cognitive tasks. Achievement tests are generally classified as either standardized or teacher/researcher made.

Standardized Tests

Standardized tests are published tests that have resulted from careful and skillful preparation by experts and cover broad academic objectives common to the majority of school systems. These are tests for which comparative norms have been derived, their validity and reliability established, and directions for administering and scoring prescribed. The directions are contained in the manuals provided by the test publishers. To establish the norms for these tests, their originators administer them to a relevant and representative sample. The norm group may be chosen to represent the nation as a whole or the state, city, district, or local school. The *mean* for a particular grade level in the sample becomes the norm for that grade level. It is important to distinguish between a norm and a standard. A *norm* is not necessarily a goal or a criterion of what should be. It is a measure of what *is*. Test norms are based on the actual performance of a specified group, not on standards of performance. The skills measured are not necessarily what "ought" to be taught at any grade level, but the use of norms does give educators a basis for comparing their groups with an estimate of the mean for all children at that grade level. Currently, as part of the accountability movement, standardized tests are being widely used to measure students' achievement. The No Child Left Behind Act of 2001 mandated that states have instruments that ensure accurate measurement of a body of skills and knowledge judged to be important and that the instruments be administered and scored under standardized conditions. The measurement aims to determine the number of students at a particular grade level who know a particular set of facts or are proficient in a particular set of skills. For example, Indiana has the ISTEP (Indiana Student Test of Educational Progress), Illinois has the ISAT (Illinois Standard Achievement Test), and California has the CST (California Standards Test).

Standardized achievement tests are available for single school subjects, such as mathematics and chemistry, and also in the form of comprehensive batteries measuring several areas of achievement. An example of the latter is the California Achievement Test (CAT/5), which contains tests in the areas of reading, language, and mathematics and is appropriate for grades kindergarten to 12. Other widely used batteries include the Iowa Tests of Basic Skills (ITBS), the Metropolitan Achievement Tests (MAT-8), the SRA Achievement Series, and the Stanford Achievement Test Series (SAT-9). Some well-known single-subject achievement tests are the Gates–MacGinitie Reading Test, the Nelson–Denny Reading Test, and the Modern Math Understanding Test (MMUT). If one is interested in measuring achievement in more than one subject area, it is less expensive and time-consuming to use a battery. The main advantage of the test battery is that each subtest is normed on the same sample, which makes comparisons across subtests, both within and between individuals, easier and more accurate.

In selecting an achievement test, researchers must be careful to choose one that is reliable and is appropriate (valid) for measuring the aspect of achievement in which they are interested. There should be a direct link between the test content and the curriculum to which students have been exposed. The test must also be valid and reliable for the type of subjects included in the study. Sometimes a researcher is not able to select the test but must use what the school system has already selected. The *Mental Measurements Yearbooks* present a comprehensive listing, along with reviews of the different achievement tests available.

If an available test measures the desired behavior and if the reliability, validity, and the norms are adequate for the purpose, then there are advantages in using a standardized instrument. In addition to the time and effort saved, investigators realize an advantage from the continuity of testing procedures—the results of their studies can be compared and interpreted with respect to those of other studies using the same instrument.

Researcher-Made Tests

When using standardized tests of achievement is not deemed suitable for the specific objectives of a research study, research workers may construct their own tests. It is much better to construct your own test than to use an inappropriate standardized one just because it is available. The advantage of a **researcher-made test** is that it can be tailored to be content specific; that is, it will match more closely the content that was covered in the classroom or in the research study. For example, suppose a teacher wants to compare the effects of two teaching methods on students' achievement in mathematics. Although there are excellent standardized tests in mathematics, they are generally designed to measure broad objectives and may not focus sufficiently on the particular skills the researcher wishes to measure. It would be wise in this case to construct the measuring instrument, paying particular attention to evidence of its validity and reliability. The researcher should administer a draft of the test to a small group who will not participate in the study but who are similar to those who will participate. An analysis of the results enables the researcher to check the test's validity and reliability and to detect any ambiguities or other problems before employing the test. For suggestions on achievement test construction, refer to specialized texts in measurement, such as those by Popham (2005), Thorndike (2005), Kubiszyn and Borich (2006), and Haladyna (2004).

Norm-Referenced and Criterion-Referenced Tests

On the basis of the type of interpretation made, standardized and **teacher-made tests** may be further classified as **norm-referenced** or **criterion-referenced.** Norm-referenced tests permit researchers to compare individuals' performance on the test to the performance of other individuals. An individual's performance is interpreted in terms of his or her relative position in a specified reference group known as the *normative group*. Typically, standardized tests are norm referenced, reporting performance in terms of percentiles, standard scores, and similar measures.

In contrast, criterion-referenced tests enable researchers to describe what a specific individual can do, without reference to the performance of others. Performance is reported in terms of the level of mastery of some well-defined content or skill domain. Typically, the level of mastery is indicated by the percentage of items answered correctly. For example, a criterion-referenced test might be used to ascertain what percentage of Indiana fourth-graders know the capitals of the 50 states. Predetermined cutoff scores may be used to interpret the individual's performance as pass–fail. The state tests used in the mandated accountability testing programs are criterion referenced. A well-known standardized instrument, the National Assessment of Educational Progress (NAEP), is criterion referenced. It is administered to a national sample of all U.S. schools to measure student knowledge in a wide variety of subject areas.

Before designing a measuring instrument, you must know which type of interpretation is to be made. In norm-referenced tests, items are selected that will yield a wide range of scores. A researcher must be concerned with the range of difficulty of the items and the power of the items to discriminate among individuals. In criterion-referenced tests, items are selected solely on the basis of how well they measure a specific set of instructional objectives. They may be easy or difficult, depending on what is being measured. The major concern is to have a representative sample of items measuring the stated objectives so that individual performance can be described directly in terms of the specific knowledge and skills that these people are able to achieve.

Test Performance Range

The range of performance that an achievement test permits is important. Researchers want a test designed so that the subjects can perform fully to their ability level without being restricted by the test. Two types of testing effects may occur. A **ceiling effect** occurs when many of the scores on a measure are at or near the maximum possible score. Tests with a ceiling effect are too easy for many of the examinees, and we do not know what their scores might have been if there had been a higher ceiling. For example, if we gave a 60-item test and most of the scores fell between 55 and 60, we would have a ceiling effect. A graph of the frequency distribution of scores would be negatively skewed (see Chapter 6).

Likewise, test performance may be restricted at the lower end of the range, resulting in a **floor effect.** A floor effect occurs when a test is too difficult and many scores are near the minimum possible score. For example, a statistics test administered as a pretest before students had a statistics class would likely show a floor effect. A graph of the frequency distribution of scores would be positively skewed. A test with a floor effect would not detect true differences in examinees' achievement either. Standardized tests typically cover a wide range of student performance, so it is not likely that many students would get all or almost all questions correct (ceiling effect) or almost all questions wrong (floor effect). A researcher should, however, consult the test manual for information about ceiling and floor effects so that he or she can select an instrument that permits a wide range of performance. Test developers gather extensive data on subjects' performance during the test standardization process. Researchers who construct their own tests can try them out with various groups and examine the results for evidence of ceiling and floor effects. If it appears that performance range is restricted, then the researcher needs to revise the test.

Performance Assessments

Another way to classify achievement tests is whether they are verbal or **performance tests.** The most common achievement tests are paper-and-pencil tests measuring cognitive objectives. This familiar format, usually administered to groups, requires individuals to compose answers or choose responses on a printed sheet. In some cases, however, a researcher may want to measure performance—what an individual can *do* rather than what he or she *knows*. Performance assessment, usually administered individually, is a popular alternative to traditional paper-and-pencil tests among educators. A performance test is a technique in which a researcher directly observes and assesses an individual's performance of a certain task and/or judges the finished product of that performance. The test taker is asked to carry out a *process* such as playing a musical instrument or tuning a car engine or to produce a *product* such as a written essay. The performance or product is judged against established criteria. An everyday example of a performance test is the behind-the-wheel examination taken when applying for a driver's license. A paper-and-pencil test covering knowledge of signs and rules for driving is not sufficient to measure driving skill. In investigating a new method of teaching science, for example, you would want to know the effect of the method not only on students' cognitive behavior but also on their learning of various laboratory procedures and techniques or their ability to complete experiments. In this case, the researcher's test would require the students to perform a real task or use their knowledge and skills to solve a science problem. Performance assessment is important in areas such as art, music, home economics, public speaking, industrial training, and the sciences, which typically involve individuals' ability to do something or produce something. Portfolios that contain a collection of student work such as poetry, essays, sketches, musical compositions, audiotapes of speeches, and even mathematics worksheets are popular in performance assessments. They provide an opportunity for teachers and researchers to gain a more holistic view of changes in students' performance over time.

Constructing a Performance Test To create a performance test, follow these three basic steps:

1. Begin with a clear statement of the objectives and what individuals will be asked to do and the conditions under which the task will be performed. A set of test specifications listing the critical dimensions to be assessed will lead to a more comprehensive coverage of the domain. State whether there will be time limits, whether reference books will be available, and so on.
2. Provide a problem or an exercise that gives students an opportunity to perform—either a simulation or an actual task. All individuals must be asked to perform the same task.
3. Develop an instrument (checklist, rating scale, or something similar) that lists the relevant criteria to use in evaluating the performance and/or the product. Make sure that the same criteria are used for each individual's performance or product.

Performance tests are useful for measuring abilities and skills that cannot be measured by paper-and-pencil tests. However, they are time intensive and thus more expensive to administer and score.

APTITUDE TESTS

Aptitude tests differ from achievement tests in that aptitude tests attempt to measure general ability or potential for learning a body of knowledge and skills, whereas achievement tests attempt to measure the actual extent of acquired knowledge and skills in specific areas. Aptitude tests measure a subject's ability to perceive relationships, solve problems, and apply knowledge in a variety of contexts. Some critics question the distinction made between aptitude and achievement tests. They point out that an aptitude test measures achievement to some extent, and an achievement test has an aptitude element. Aptitude tests were formerly referred to as **intelligence tests**, but the latter term has declined in use because of controversy over the definition of intelligence and because people tend to associate intelligence with inherited ability. Aptitude tests should *not* be considered as measures of innate (or "pure") intelligence. As noted previously, performance on such tests partly depends on the background and schooling of the subject.

Educators have found aptitude tests useful and generally valid for the purpose of predicting school success. Many of the tests are referred to as **scholastic aptitude tests,** a term pointing out specifically that the main function of these tests is to predict school performance. Well-known aptitude tests are the ACT (American College Testing Assessment) and the SAT (Scholastic Assessment Test) for high school students and the GRE (Graduate Record Exam) and MAT (Miller Analogies Test) for college seniors.

Researchers often use aptitude tests. Aptitude or intelligence is frequently a variable that needs to be controlled in educational experiments. To control this variable, the researcher may use the scores from a scholastic aptitude test. Of the many tests available, some have been designed for use with individuals and others for use with groups.

Individual Aptitude Tests

The most widely used individually administered instruments for measuring aptitude are the Stanford–Binet Intelligence Scale (4th ed.) and the three Wechsler tests. The Stanford–Binet currently in use is the outcome of several revisions of the device first developed in France in 1905 by Alfred Binet and Theodore Simon for identifying children who were not likely to benefit from normal classroom instruction. It was made available for use in the United States in 1916. This test originally reported an individual's mental age. Later, the concept of *intelligence quotient* (IQ) was introduced. This quotient was derived by dividing mental age (MA) by chronological age (CA) and multiplying the result by 100. The present revision of the Stanford–Binet no longer employs the MA/CA ratio for determining IQ. The IQ is found by comparing an individual's performance (score) with norms obtained from his or her age group through the use of standard scores (see Chapter 6). The latest revision of the test has 15 subtests organized into four areas: Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning, and Short-Term Memory. The scores on the 15 subtests are standard scores with a mean of 50 and a standard deviation of 8. The four area scores and the total IQ score all have a mean of 100 and standard deviation of 16. The Stanford–Binet is appropriate for ages 2 years through adult.

The tests David Wechsler developed to measure aptitude now come in several forms: the Wechsler Intelligence Scale for Children—Third Edition (WISC–III, 1991), the Wechsler Adult Intelligence Scale-III (WAIS–III, 1997), and the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI–R, 1989), which was introduced for the 4 to 6½-year age group. The Wechsler tests yield verbal IQ scores, performance IQ scores, and full-scale IQ scores derived by averaging the verbal subtest scores, the performance subtest scores, and all subtest scores, respectively. The Wechsler scales are more popular than the Stanford–Binet primarily because they require less time to administer.

Group Tests of Aptitude

A Stanford–Binet or Wechsler test must be given by a trained psychometrician to an individual subject, a procedure expensive in both time and money. Thus, they are impractical as aptitude measures for large groups of individuals. In this situation, group tests are used. The first group test of mental ability was developed during World War I for measuring the ability of men in military service. One form of this test, the Army Alpha, was released for civilian use after the war and became the model for a number of group tests. Today, many group tests of mental aptitude are available. Among the most widely used are the Cognitive Abilities Tests (CogAT), Test of Cognitive Skills (TCS/2), and the Otis–Lennon School Ability Tests (OLSAT-7). The CogAT and the OLSAT-7 are appropriate for grades kindergarten to 12, whereas the TCS/2 is used for grades 2 to 12.

TESTING AND TECHNOLOGY

New technologies are presenting opportunities for alternatives to paper-and-pencil tests. For example, the PRAXIS I test designed to assess basic skills prior to entry into teacher education is given electronically with immediate scoring and feedback on performance provided to the examinee. A computer is also used to administer the GRE and a number of other well-known tests. Many of you may have encountered computer-based testing when you took the knowledge portion of your test to obtain a driver's license.

MEASURES OF PERSONALITY

Educational researchers often use measures of personality. There are several different types of personality measures, each reflecting a different theoretical point of view. Some reflect trait and type theories, whereas others have their origins in psychoanalytic and motivational theories. Researchers must know precisely what they wish to measure and then select the instrument, paying particular attention to the evidence of its validity. Two approaches are used to measure personality: objective personality assessment and projective personality assessment.

OBJECTIVE PERSONALITY ASSESSMENT

Self-report inventories present subjects with an extensive collection of statements describing behavior patterns and ask them to indicate whether or not each statement is characteristic of their behavior by checking *yes*, *no*, or *uncertain*. Other formats use multiple choice and true–false items. The score is computed by counting the number of responses that agree with a trait the examiner is attempting to measure. For example, someone with paranoid tendencies would be expected to answer *yes* to the statement “People are always talking behind my back” and *no* to the statement “I expect the police to be fair and reasonable.” Of course, similar responses to only two items would not indicate paranoid tendencies. However, such responses to a large proportion of items could be considered an indicator of paranoia.

Some of the self-report inventories measure only one trait, such as the California F-Scale, which measures authoritarianism. Others, such as Cattell's Sixteen Personality Factor Questionnaire, measure a number of traits. Other multiple-trait inventories used in research are the Minnesota Multiphasic Personality Inventory (MMPI-2), the Guilford–Zimmerman Temperament Survey, the Mooney Problem Check List, the Edwards Personal Preference Schedule (EPPS), the Myers–Briggs Type Indicator, and the Strong Interest Inventory. A popular inventory, the Adjective Checklist, asks individuals to check from a list of adjectives those that are applicable to themselves. It is appropriate for individuals in grade 9 through adults and only takes 15 minutes to complete. It yields scores on self-confidence, self-control, needs, and other aspects of personality adjustment.

Inventories have been used in educational research to obtain trait descriptions of certain defined groups, such as underachievers and dropouts. They are useful for finding out about students' self-concepts, their concerns or problems, and their study skills and habits. Inventories have also been used in research concerned with interrelationships between personality traits and such variables as aptitude, achievement, and attitudes.

Inventories have the advantages of economy, simplicity, and objectivity. They can be administered to groups and do not require trained psychometricians. Most of the disadvantages are related to the problem of validity. The validity of self-report inventories depends in part on the respondents' being able to read and understand the items, their understanding of themselves, and especially their willingness to give frank and honest answers. As a result, the information obtained from inventories may be superficial or biased. This possibility must be taken into account when using results obtained from such instruments. Some inventories have built in validity scales to detect faking, attempts to give socially desirable responses, or reading comprehension problems.

PROJECTIVE PERSONALITY ASSESSMENT

Projective techniques are measures in which an individual is asked to respond to an ambiguous or unstructured stimulus. They are called *projective* because a person is expected to project into the stimulus his or her own needs, wants, fears, beliefs, anxieties, and experiences. On the basis of the subject's interpretation of the stimuli and his or her responses, the examiner attempts to construct a comprehensive picture of the individual's personality structure. Projective methods are used mainly by clinical psychologists for studying and diagnosing people with emotional problems. They are not frequently used in educational research because of the necessity of specialized training for administration and scoring and the expense involved in individual administration. Furthermore, many researchers question their validity primarily because of the complex scoring. The two best known projective techniques are the Rorschach Inkblot Technique and the Thematic Apperception Test (TAT). The Rorschach consists of 10 cards or plates each with either a black/white or a colored inkblot. Individuals are asked what they "see." Their responses are scored according to whether they used the whole or only a part of the inkblot or if form or color was used in structuring the response, whether movement is suggested, and other aspects. In the TAT, the respondent is shown a series of pictures varying in the extent of structure and ambiguity and asked to make up a story about each one. The stories are scored for recurrent themes, expression of needs, perceived problems, and so on. The TAT is designed for individuals age 10 years through adult. There is also a form available for younger children (Children's Apperception Test) and one for senior citizens (Senior Apperception Test).

SCALES

Scales are used to measure attitudes, values, opinions, and other characteristics that are not easily measured by tests or other measuring instruments. A **scale** is a set of categories or numeric values assigned to individuals, objects, or behaviors for the purpose of measuring variables. The process of assigning scores to those objects in order to obtain a measure of a construct is called *scaling*. Scales differ from tests in that the results of these instruments, unlike those of tests, do not indicate success or failure, strength or weakness. They measure the degree to which an individual exhibits the characteristic of interest. For example, a researcher may use a scale to measure the attitude of college students toward religion or any other topic. A number of scaling techniques have been developed throughout the years.

ATTITUDE SCALES

Attitude scales use multiple responses—usually responses to statements—and combine the responses into a single scale score. Rating scales, which we discuss later in this chapter, use judgments—made by the individual under study or by an observer—to assign scores to individuals or other objects to measure the underlying constructs.

Attitudes of individuals or groups are of interest to educational researchers. An attitude may be defined as a positive or negative affect toward a particular group, institution, concept, or social object. The measurement of attitudes presumes the ability to place individuals along a continuum of favorableness–unfavorableness toward the object.

If researchers cannot locate an existing attitude scale on their topic of interest, they must develop their own scales for measuring attitudes. We discuss two types of attitude scales: summated or Likert (pronounced *Lik'ert*) scales and bipolar adjective scales.

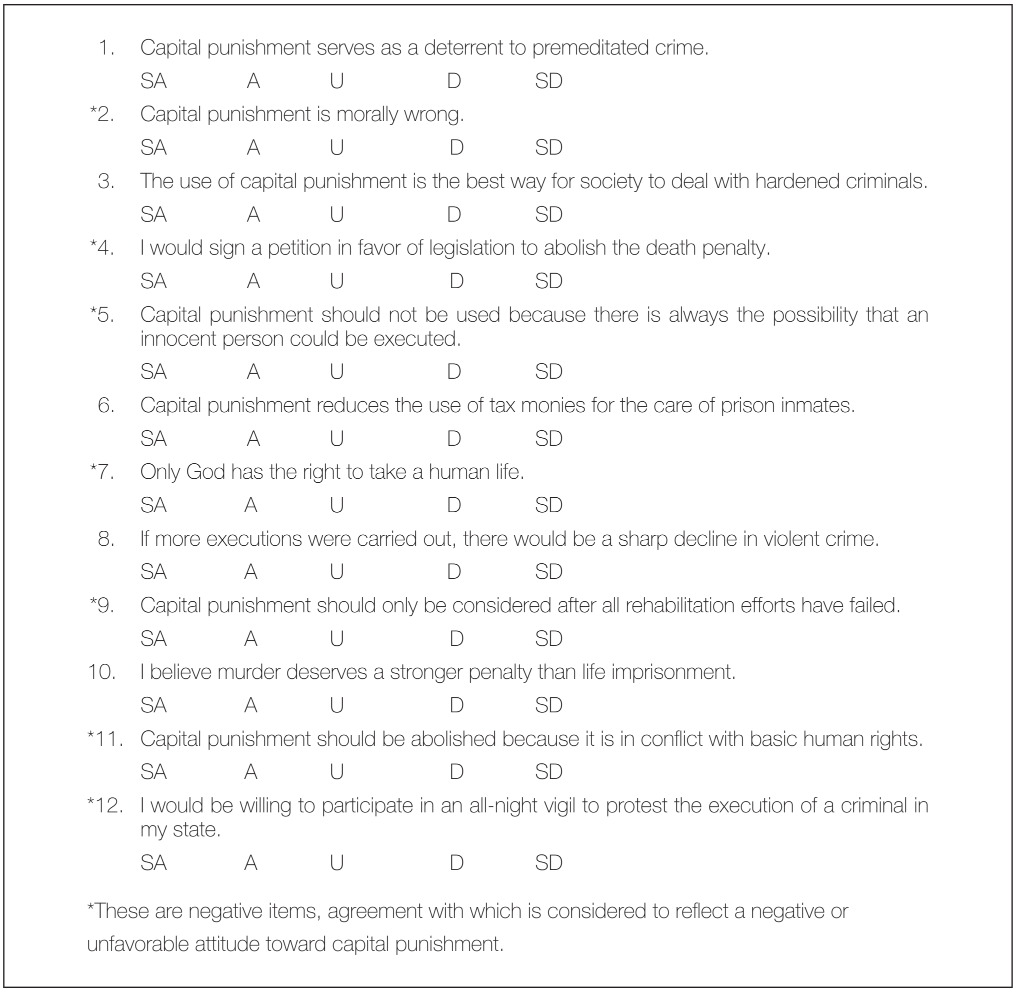
Likert Scales: Method of Summated Ratings

The Likert scale (1932), named for Rensis Likert who developed it, is one of the most widely used techniques to measure attitudes. A **Likert scale** (a **summated rating scale**) assesses attitudes toward a topic by presenting a set of statements about the topic and asking respondents to indicate for each whether they strongly agree, agree, are undecided, disagree, or strongly disagree. The various agree–disagree responses are assigned a numeric value, and the total scale score is found by summing the numeric responses given to each item. This total score assesses the individual's attitude toward the topic.

A Likert scale is constructed by assembling a large number of statements about an object, approximately half of which express a clearly favorable attitude and half of which are clearly unfavorable. Neutral items are not used in a Likert scale. It is important that these statements constitute a representative sample of all the possible opinions or attitudes about the object. It may be helpful to think of all the subtopics relating to the attitude object and then write items on each subtopic. To generate this diverse collection of items, the researcher may find it helpful to ask people who are commonly accepted as having knowledge about and definite attitudes toward the particular object to write a number of positive and negative statements. Editorial writings about the object are also good sources of potential statements for an attitude scale. Figure 8.1 shows items from a Likert scale designed to measure attitudes toward capital punishment.

For pilot testing, the statements, along with five response categories arranged on an agreement–disagreement continuum, are presented to a group of subjects. This group should be drawn from a population that is similar to the one in which the scale will be used. The statements should be arranged in random order so as to avoid any response set on the part of the subjects.

The subjects are directed to select the response category that best represents their reaction to each statement: *strongly agree* (SA), *agree* (A), *undecided* (U), *disagree* (D), or *strongly disagree* (SD). There has been some question regarding whether the undecided option should be included in a Likert scale. Most experts in the field recommend that the researcher include a neutral or undecided choice



1. Capital punishment serves as a deterrent to premeditated crime.
SA A U D SD

*2. Capital punishment is morally wrong.
SA A U D SD

3. The use of capital punishment is the best way for society to deal with hardened criminals.
SA A U D SD

*4. I would sign a petition in favor of legislation to abolish the death penalty.
SA A U D SD

*5. Capital punishment should not be used because there is always the possibility that an innocent person could be executed.
SA A U D SD

6. Capital punishment reduces the use of tax monies for the care of prison inmates.
SA A U D SD

*7. Only God has the right to take a human life.
SA A U D SD

8. If more executions were carried out, there would be a sharp decline in violent crime.
SA A U D SD

*9. Capital punishment should only be considered after all rehabilitation efforts have failed.
SA A U D SD

10. I believe murder deserves a stronger penalty than life imprisonment.
SA A U D SD

*11. Capital punishment should be abolished because it is in conflict with basic human rights.
SA A U D SD

*12. I would be willing to participate in an all-night vigil to protest the execution of a criminal in my state.
SA A U D SD

*These are negative items, agreement with which is considered to reflect a negative or unfavorable attitude toward capital punishment.

Figure 8.1 Example of a Likert Scale

Source: These items were taken from an attitude scale constructed by a graduate student in an educational research class.

because some respondents actually feel that way and do not want to be forced into agreeing or disagreeing.

Scoring Likert Scales To score the scale, the response categories must be weighted. For favorable or positively stated items, *strongly agree* is scored 5, *agree* is scored 4, *undecided* is scored 3, *disagree* is scored 2, and *strongly disagree* is scored 1. For unfavorable or negatively stated items, the weighting is reversed because disagreement with an unfavorable statement is psychologically equivalent to agreement with a favorable statement. Thus, for unfavorable statements, *strongly agree* would receive a weight or score of 1 and *strongly disagree* a weight of 5. (The weight values do not appear on the attitude scale presented to respondents, nor do the asterisks seen in Figure 8.1.)

The sum of the weights of all the items checked by the subject is the individual's total score. The highest possible scale score is $5 \times N$ (the number of items); the lowest possible score is $1 \times N$.

Let us consider an example of scoring a Likert scale by looking at just the first six statements of the scale shown in Figure 8.1. An individual would complete this scale by circling the appropriate letter(s) for each statement.

The following are the responses circled by a hypothetical respondent and the score for each item:

Response	Score
1. D	2
2. SA	1
3. D	2
4. A	2
5. A	2
6. U	3

The individual's total score on the six items is 12 (out of a possible 30). Divide the total score by the number of items to arrive at a mean attitude score: $2 + 1 + 2 + 2 + 2 + 3)/6 = 2.0$. Because the mean score is less than 3, we conclude that this individual has a moderately negative attitude toward capital punishment.

Item Analysis After administering the attitude scale to a preliminary group of respondents, the researcher does an **item analysis** to identify the best functioning items. The item analysis typically yields three statistics for each item: (1) an item discrimination index, (2) the percentage of respondents marking each choice to each item, and (3) the item mean and standard deviation.

The item discrimination index shows the extent to which each item discriminates among the respondents in the same way as the total score discriminates. The item discrimination index is calculated by correlating item scores with total scale scores, a procedure usually done by computer. If high scorers on an individual item have high total scores and if low scorers on this item have low total scores, then the item is discriminating in the same way as the total score. To be useful, an item should correlate at least .25 with the total score. Items that have very low correlation or negative correlation with the total score should be eliminated because they are not measuring the same thing as the total scale and hence are not contributing to the measurement of the attitude. The researcher will want to examine those items that are found to be nondiscriminating. The items may be ambiguous or double barreled (containing two beliefs or opinions in one statement), or they may be factual statements not really expressing feelings about the object. Revising these items may make them usable. The item analysis also shows the percentage of respondents choosing each of the five options and the mean and standard deviation for each item. Items on which respondents are spread out among the options are preferred. Thus, if most respondents choose only one or two of the options, the item should be rewritten or eliminated. After selecting the most useful items as indicated by the item analysis, the researcher should then try out the revised scale with a different group of subjects and again check the items for discrimination and variability.

Validity Validity concerns the extent to which the scale really measures the attitude construct of interest. It is often difficult to locate criteria to be used in obtaining evidence for the validity of attitude scales. Some researchers have used observations of actual behavior as the criterion for the attitude being measured. This procedure is not often used because it is often difficult to determine what behavior would be the best criterion for the attitude and also because it is expensive.

One of the easiest ways to gather validity evidence is to determine the extent to which the scale is capable of discriminating between two groups whose members are known to have different attitudes (see Chapter 9). To validate a scale that measures attitudes toward organized religion, a researcher would determine if the scale discriminated between active church members and people who do not attend church or have no church affiliation. A scale measuring attitudes toward abortion should discriminate between members of pro-life groups and members of pro-choice groups. By "discriminate," we mean that the two groups would be expected to have significantly different mean scores on the scale. Another method of assessing validity is to correlate scores on the attitude scale with those obtained on another attitude scale measuring the same construct and whose validity is well established.

Reliability The reliability of the new scale must also be determined. Reliability is concerned with the extent to which the measure would yield consistent results each time it is used. The first step in ensuring reliability is to make sure that the scale is long enough—that it includes enough items to provide a representative sampling of the whole domain of opinions about the attitudinal object. Other things being equal, the size of the reliability coefficient is directly related to the length of the scale. Research shows, however, that if the items are well constructed, scales having as few as 20 to 22 items will have satisfactory reliability (often above .80). The number of items needed depends partly on how specific the attitudinal object is; the more abstract the object, the more items are needed.

You would also want to calculate an index of reliability. The best index to use for an attitude scale is coefficient alpha (see Chapter 9), which provides a measure of the extent to which all the items are positively intercorrelated and working together to measure one trait or characteristic (the attitude). Many statistical computer programs routinely calculate coefficient alpha as a measure of reliability. For further discussion on the construction of Likert and other attitude scales, the reader is referred to Mueller (1986).

Bipolar Adjective Scales

The **bipolar adjective scale** presents a respondent with a list of adjectives that have bipolar or opposite meanings. Respondents are asked to place a check mark at one of the seven points in the scale between the two opposite adjectives to indicate the degree to which the adjective represents their attitude toward an object, group, or concept. Figure 8.2 shows a bipolar adjective scale designed to measure attitude toward school. Notice that the respondent checked the extreme right position for item a and the extreme left position for item d. The adjective pairs making up a scale are listed in both directions; on some pairs the rightmost position is the most positive response, and on other pairs the leftmost position is the most positive. This is done to minimize a response set or a tendency to favor certain positions in a list of options. An individual might have a tendency to choose

School

a. bad							✓	good
b. fast		✓						slow
c. dull						✓		sharp
d. pleasant	✓							unpleasant
e. light			✓					heavy
f. passive							✓	active
g. worthless						✓		valuable
h. strong				✓				weak
i. still					✓			moving

Figure 8.2 Bipolar Adjective Scale Showing Responses of One Subject Toward the Concept “School”

the extreme right end and would check that position for each item. However, if the direction of the scale is changed in a random way so that the right end is not always the more favorable response, the individual must read each item and respond in terms of its content rather than in terms of a positional preference. The responses are scored by converting the positions checked into ratings (1 to 7). Seven represents the most positive and 1 the least positive response on each scale. The weights on each item would then be summed and averaged. In Figure 8.2, item weights are $7 + 6 + 6 + 7 + 3 + 7 + 6 + 4 + 5 = 51/9 = 5.67$. The score of 5.67 indicates a very positive attitude toward school.

The bipolar adjective scale is a very flexible approach to measuring attitudes. A researcher can use it to investigate attitudes toward any concept, person, or activity in any setting. It is much easier and less time-consuming to construct than a Likert scale. Instead of having to come up with approximately 20 statements, you need only select four to eight adjective pairs. It requires very little reading time by participants. The main difficulty is the selection of the adjectives to use. If one has a problem with this task, there are references such as Osgood, Suci, and Tannenbaum (1967) that provide lists of bipolar adjectives. It is probably better, however, to think of adjective pairs that are especially relevant to one’s own project.

RATING SCALES

Rating scales present a number of statements about a behavior, an activity, or a phenomenon with an accompanying scale of categories. Observers or respondents are asked to indicate their assessment or judgment about the behavior or activity on the rating scale. For example, a teacher might be asked to rate the leadership ability of a student. The teacher would indicate his or her assessment of the student’s characteristic leadership behavior by checking a point on a continuum or choosing a response category. It is assumed that raters are familiar with the behavior they are asked to assess. A numeric value may be attached to the points or categories so that an overall score could be obtained.

One of the most widely used rating scales is the **graphic scale,** in which the respondent indicates the rating by placing a check at the appropriate point on a

	Low	Medium	High
Personal appearance	______	______	______
Social acceptability	______	______	______
Speaking skills	______	______	______

Figure 8.3 Example of a Graphic Scale

horizontal line that runs from one extreme of the behavior in question to the other. Figure 8.3 is an example of a graphic scale. The rater can check any point on the continuous line. Graphic scales usually assign numeric values to the descriptive points. Such scales are referred to as *numeric rating scales*. The speaking skills item in Figure 8.3 could look like this in a numeric scale:

1	2	3	4	5	6	7
one of the poorest speakers			an average speaker			one of the very best speakers

Category Scales

The **category scale** consists of a number of categories that are arranged in an ordered series. Five to seven categories are most frequently used. The rater picks the one that best characterizes the behavior of the person being rated. Suppose a student's abilities are being rated and one of the characteristics being rated is creativity. The following might be one category item:

How creative is this person? (check one)

exceptionally creative ______
very creative ______
not creative ______
not at all creative ______

To provide greater meaning, brief descriptive phrases are sometimes used to comprise the categories in this type of scale. Clearly defined categories contribute to the accuracy of the ratings. For example,

How creative is this person? (check one)

always has creative ideas ______
has many creative ideas ______
sometimes has creative ideas ______
rarely has creative ideas ______

Comparative Rating Scales

In using the graphic and category scales, raters make their judgments without directly comparing the person being rated to other individuals or groups. In **comparative rating scales,** in contrast, raters are instructed to make their judgment with direct reference to the positions of others with whom the individual might be compared. The positions on the rating scale are defined in terms of a given population with known characteristics. A comparative rating scale is shown in Figure 8.4. Such a scale might be used in selecting applicants for admission to graduate school. Raters are asked to judge the applicant's ability to do graduate work compared with that of all the students the rater has known. If the rating is to be valid, the judge must understand the range and distribution of abilities in the total group of graduate students.

Errors in Rating Because ratings depend on the perceptions of human observers, who are susceptible to various influences, rating scales are subject to

Area of Competency (to be rated)	Unusually low	Poorer than most students	About average among students	Better than most	Really superior	Not able to judge
1. Does this person show evidence of clear-cut and worthy professional goals?						
2. Does this person attack problems in a constructive manner?						
3. Does he or she take well-meant criticism and use it constructively?						

Figure 8.4 Example of a Comparative Rating Scale

considerable error. Among the most frequent systematic errors in rating people is the **halo effect,** which occurs when raters allow a generalized impression of the subject to influence the rating given on very specific aspects of behavior. This general impression carries over from one item in the scale to the next. For example, a teacher might rate a student who does good academic work as also being superior in intelligence, popularity, honesty, perseverance, and all other aspects of personality. Or, if you have a generally unfavorable impression of a person, you are likely to rate the person low on all aspects.

Another type of error is the **generosity error,** which refers to the tendency for raters to give subjects the benefit of any doubt. When raters are not sure, they tend to rate people favorably. In contrast, the **error of severity** is a tendency to rate all individuals too low on all characteristics. Another source of error is the **error of central tendency,** which refers to the tendency to avoid either extreme and to rate all individuals in the middle of the scale. For example, the ratings that teachers of English give their students have been found to cluster around the mean, whereas mathematics teachers' ratings of students show greater variation.

One way of reducing such errors is to train the raters thoroughly before they are asked to make ratings. They should be informed about the possibility of making these "personal bias" types of errors and how to avoid them. It is absolutely essential that raters have adequate time to observe the individual and his or her behavior before making a rating. Another way to minimize error is to make certain that the behavior to be rated and the points on the rating scale are clearly defined. The points on the scale should be described in terms of overt behaviors that can be observed, rather than in terms of behaviors that require inference on the part of the rater.

The accuracy or reliability of ratings is usually increased by having two (or more) trained raters make independent ratings of an individual. These independent ratings are pooled, or averaged, to obtain a final score. A researcher may also correlate the ratings of the two separate raters in order to obtain a coefficient of interrater reliability (see Chapter 9). The size of the coefficient indicates the extent to which the raters agree. An **interrater reliability** coefficient of .70 or higher is considered acceptable for rating scales.

DIRECT OBSERVATION

In many cases, systematic or **direct observation** of behavior is the most desirable measurement method. Observation is used in both quantitative and qualitative research. When observations are made in an attempt to obtain a comprehensive picture of a situation, and the product of those observations is notes or narratives, the research is qualitative. In Chapter 15, we discuss the use of observation in qualitative research. The current chapter focuses on observation in quantitative research where the product of using the various observational instruments is numbers. The purpose of direct observation is to determine the extent to which a particular behavior(s) is present. The observer functions like a camera or recording device to provide a record of the occurrence of the behavior in question. The researcher identifies the behavior of interest and devises a systematic procedure for identifying, categorizing, and recording the behavior in either a natural or a contrived situation. The behaviors observed in quantitative studies may be categorized as high inference and low inference. High-inference behaviors such as teacher warmth or creativity require more judgment on the part of the observer. Low-inference behaviors require less judgment by the observer. Examples of low-inference behaviors include classroom behaviors such as teachers' asking questions, praising students, or accepting students' ideas. In educational research, one of the most common uses of direct observation is in studying classroom behavior. For example, if you were interested in investigating the extent to which elementary teachers use positive reinforcement in the classroom, you could probably obtain more accurate data by actually observing classrooms rather than asking teachers about their use of reinforcement. Or if you wanted to study students' disruptive behavior in the classroom and how teachers deal with it, direct observation would provide more accurate data than reports from students or teachers.

There are five important preliminary steps to take in preparing for quantitative direct observation:

1. *Select the aspect of behavior to be observed*. Because it is not possible to collect data on everything that happens, the investigator must decide beforehand which behaviors to record and which not to record.
2. *Clearly define the behaviors falling within a chosen category*. Know what behaviors would be indicators of the attribute. In studying aggressive behavior in the classroom, would challenging the teacher or speaking out of turn be classified as aggressive, or would it be restricted to behaviors such as pushing, hitting, throwing objects, and name-calling? If observing multiple categories of behavior, make sure the categories are mutually exclusive.
3. *Develop a system for quantifying observations*. The investigator must decide on a standard method for counting the observed behaviors. For instance, establish beforehand whether an action and the reaction to it are to count as a single incident of the behavior observed or as two incidents. A suggested approach is to divide the observation period into brief time segments and to record for each period—for example, 10 seconds—whether the subject showed the behavior or not.

4. *Develop specific procedures for recording the behavior*. Record the observations immediately after they are made because observers' memory is not sufficiently reliable for accurate research. The best solution is a coding system that allows the immediate recording of what is observed, using a single letter or digit. A coding system is advantageous in terms of the observers' time and attention.
5. *Train the people who will carry out the observations*. Training and opportunity for practice are necessary so that the investigator can rely on the observers to follow an established procedure in observing and in interpreting and reporting observations. Having the observers view a videotape and discuss the results is a good training technique.

DEVICES FOR RECORDING OBSERVATIONS

Researchers use checklists, rating scales, and coding sheets to record the data collected in direct observation.

Checklists

The simplest device used is a **checklist,** which presents a list of the behaviors that are to be observed. The observer then checks whether each behavior is present or absent. A checklist differs from a scale in that the responses do not represent points on a continuum but, rather, nominal categories. For example, a researcher studying disruptive behavior would prepare a list of disruptive behaviors that might occur in a classroom. An observer would then check items such as "Passes notes to other students" or "Makes disturbing noises" each time the behavior occurs. The behaviors in a checklist should be operationally defined and readily observable.

Rating Scales

Rating scales, discussed previously, are often used by observers to indicate their evaluation of an observed behavior or activity. Typically, rating scales consist of three to five points or categories. For example, an observer studying teachers' preparation for presentation of new material in a classroom might use a scale with the following points: 5 (*extremely well prepared*), 4 (*well prepared*), 3 (*prepared*), 2 (*not well prepared*), or 1 (*totally unprepared*). A 3-point scale might include 3 (*very well prepared*), 2 (*prepared*), or 1 (*not well prepared*). Scales with more than five rating categories are not recommended because it is too difficult to accurately discriminate among the categories.

Coding Systems

Coding systems are used in observational studies to facilitate the categorizing and counting of specific, predetermined behaviors as they occur. The researcher does not just indicate whether a behavior occurred as with a checklist but, rather, uses agreed-on codes to record what actually occurred. Whereas rating scales can be completed after an observation period, coding is completed at the time the observer views the behavior.

Two kinds of coding systems are typically used by researchers: sign coding and time coding. *Sign coding* uses a set of behavior categories; each time one of the behaviors occurs, the observer codes the happening in the appropriate category. If a coding sheet used in classroom observational research listed "summarizing" as a teacher behavior, the observer would code a happening every time a teacher summarized material.

In a study using sign coding, Skinner, Buysse, and Bailey (2004) investigated how total duration and type of social play of preschool children with disabilities varied as a function of the chronological and developmental age of their social partners. They hypothesized that developmental age of each partner would better predict the duration of social play than chronological age. The 55 focal children were preschool children with mild to moderate developmental delays who were enrolled in some type of inclusive developmental day program. Each focal child was paired with 4 different same-sex partners in a standardized dyadic play situation. The observations took place outside the classroom in a specially designed and well-equipped play area. The observation consisted of two 15-minute sessions with each of the 4 play partners, or a total of 120 minutes per focal child over a period of 2 days. A video camera recorded the play behavior and trained coders used Parten's (1932) seven categories of play to code the extent to which children were engaged socially. The Battelle Developmental Inventory (Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1988) was used to assess the overall developmental status of both focal children and their social partners. A mixed-model regression analysis was employed, with the independent variables being the chronological and developmental ages of both the focal children and the partners; the dependent variable was the total duration of the category called associative play. No impact was observed for the focal children's chronological age once they accounted for developmental age. Also, they found that the influence of partner's developmental age on social play was different depending on the developmental age of the focal child. The researchers concluded that advantages accrued to preschoolers with disabilities from mixed-aged play groupings depend on the child's developmental age and those of available social partners.

In the second type of coding, called *time coding*, the observer identifies and records all predetermined behavior categories that occur during a given time period. The time period might be 10 seconds, 5 minutes, or some other period of time. Miller, Gouley, and Seifer (2004) used time coding in a study designed to document observed emotional and behavioral dysregulation in the classroom and to investigate the relationships between observed dysregulation and teachers' ratings of children's classroom adjustment and their social engagement with peers. Dysregulation was defined as emotional and behavioral displays disruptive to the preschool classroom setting. The participants were 60 low-income children attending Head Start classes. Each child was observed in a naturalistic context for two sessions of 10 minutes each, or a total of 20 minutes. The researchers used handheld computers with *The Observer* (Noldus Information Technology, 1995) software, which permitted coding of behavior along several dimensions. Analysis showed that although the majority of children did not display much dysregulated emotion or behavior in the classroom, almost one-fourth of children did display high levels of dysregulation in the observation period.

High levels of classroom dysregulation were related to teacher ratings of poor classroom adjustment and observed peer conflict behaviors, as well as negative emotional displays.

Coding has the advantage of recording observations at the time the behavior occurs, and it may yield more objective data than do rating scales. The disadvantage is that a long training period may be required for observers to learn to code behavior reliably. A number of standardized coding systems and observation forms are available. Beginning researchers should check references such as the ETS Test Collection Database for a suitable one before attempting to construct their own.

ADVANTAGES AND DISADVANTAGES OF DIRECT OBSERVATION

The most obvious advantage of systematic observation is that it provides a record of the actual behavior that occurs. We do not have to ask subjects what they would do or what they think; we have a record of their actions. Probably the most important advantage of systematic observation is its appropriateness for use with young children. It is used extensively in research on infants and on preschool children who have difficulty communicating through language and may be uncomfortable with strangers. Another advantage is that systematic observation can be used in natural settings. It is often used in educational research to study classroom or playground behavior.

The main disadvantage of systematic observation is the expense. Observations are more costly because of the time required of trained observers. Subjects may be observed for a number of sessions, requiring extended hours.

VALIDITY AND RELIABILITY OF DIRECT OBSERVATION

As with other types of measures, the validity and reliability of direct observation must be assessed. The best way to enhance validity is to carefully define the behavior to be observed and to train the people who will be making the observations. Observers must be aware of two sources of bias that affect validity: observer bias and observer effect. **Observer bias** occurs when the observer's own perceptions, beliefs, and biases influence the way he or she observes and interprets the situation. Having more than one person make independent observations helps to detect the presence of bias. **Observer effect** occurs when people being observed behave differently just because they are being observed. One-way vision screens may be used in some situations to deal with this problem. In many cases, however, after an initial reaction the subjects being observed come to pay little attention to the observer, especially one who operates unobtrusively. Some studies have used interactive television to observe classrooms unobtrusively. Videotaping for later review and coding may also be useful. Researchers who have used videotapes, for example, have found that although the children initially behaved differently with the equipment in the room, after a few days they paid no attention and its presence became routine. Handheld technologies, such as a PalmPilot, can be used to record data during observations rather than the traditional pencil-and-paper recording techniques. Professional software such as *The Observer XT* 8.0 (Noldus Information Technology, 2008) is available for use in the collection, analysis, and presentation of observational data. Information on *The Observer XT 8.0* is available at www.noldus.com/site/doc200806003.

The accuracy or reliability of direct observation is usually investigated by having at least two observers independently observe the behavior and then determining the extent to which the observers' records agree. Reliability is enhanced by providing extensive training for the observers so that they are competent in knowing what to observe and how to record the observations. Further discussion of methods for assessing the reliability of direct observation is presented in Chapter 9.

CONTRIVED OBSERVATIONS

In **contrived observations,** the researcher arranges for the observation of subjects in simulations of real-life situations. The circumstances have been arranged so that the desired behaviors are elicited.

One form of contrived observation is the **situational test.** A classic example of a situational test—although not labeled as such at the time—was used in a series of studies by Hartshorne and May (1928) for the Character Education Inquiry (CEI). These tests were designed for use in studying the development of such behavior characteristics as honesty, self-control, truthfulness, and cooperativeness. Hartshorne and May observed children in routine school activities but also staged some situations to focus on specific behavior. For example, they gave vocabulary and reading tests to the children, collected the tests, and without the children's knowledge made duplicate copies of their answers. Later, the children were given answer keys and were asked to score their original papers. The difference between the scores the children reported and the actual scores obtained from scoring the duplicate papers provided a measure of cheating. Another test asked the children to make a mark in each of 10 small, irregularly placed circles while keeping their eyes shut. Previous control tests under conditions that prevented peeking indicated that a score of more than 13 correctly placed marks in a total of three trials was highly improbable. Thus, a score of more than 13 was recorded as evidence that the child had peeked.

Hartshorne and May (1928) found practically no relationship between cheating in different situations, such as on a test and in athletics. They concluded that children's responses were situationally specific—that is, whether students cheated depended on the specific activity, the teacher involved, and other situations rather than on some general character trait.

DATA COLLECTION IN QUALITATIVE RESEARCH

Qualitative researchers also have a number of data-gathering tools available for their investigations. The most widely used tools in qualitative research are interviews, document analysis, and observation. We discuss these methods in Chapter 15.

SUMMARY

One of the most important tasks of researchers in the behavioral sciences is the selection and/or development of dependable measuring instruments. A research study can be no better than the instruments used to collect the data. A variety of tests, scales, and inventories are available for gathering data in educational research, especially for quantitative studies. Researchers need to be aware of the strengths and limitations of these data-gathering instruments so that they can choose the one(s) most appropriate for their particular investigation. If an appropriate standardized instrument is available, the researcher would be wise to choose it because of the advantage in terms of validity, reliability, and time saved.

A test is a set of stimuli presented to an individual to elicit responses on the basis of which a numerical score can be assigned. Achievement tests measure knowledge and proficiency in a given area and are widely used in educational research. Standardized achievement tests permit the researcher to compare performance on the test to the performance of a normative reference group. Tests may be classified as paper-and-pencil or as performance tests, which measure what someone can *do* rather than what he or she *knows*. Aptitude tests are used to assess an individual's verbal and nonverbal capacities. Personality inventories are designed to measure the subject's personal characteristics and typical performance.

Attitude scales are tools for measuring individuals' beliefs, feelings, and reactions to certain objects. The major types of attitude scales are Likert-type scales and the bipolar adjective scale.

Rating scales permit observers to assign scores to the assessments made of observed behavior or activity. Among the types of rating scales are the graphic scale, the category scale, and comparative rating scales.

Rating scales, checklists, and coding systems are most commonly used to record the data in quantitative direct observation research. In coding systems, behavior can be categorized according to individual occurrences (sign coding) or number of occurrences during a specified time period (time coding).

KEY CONCEPTS

achievement test
aptitude test
attitude scale
bipolar adjective scale
category scale
ceiling effect
checklist
coding system
comparative rating scales
contrived observation
criterion-referenced test
direct observation
error of central tendency
error of severity
floor effect
generosity error
graphic scale
halo effect
intelligence test
interrater reliability
inventories
item analysis
Likert scale
norm-referenced test
observer bias
observer effect
performance test
projective technique
rating scale
researcher-made test
scale scholastic aptitude test
self-report inventories
situational test
standardized test
summated rating scale
teacher-made test
test

EXERCISES

1. What is the meaning of the term *standardized* when applied to measuring instruments?
2. What is the difference between comparative rating scales and graphic and category scales?
3. List some of the common sources of bias in rating scales.
4. What type of instrument would a researcher choose in order to obtain data about each of the following?

a. How college professors feel about the use of technology in their teaching
b. The potential of the seniors at a small college to succeed in graduate school
c. To determine if high school chemistry students can analyze an unknown chemical compound
d. How well the students at Brown Elementary School compare to the national average in reading skills
e. The advising-style preferences of a group of college freshmen
f. How well students perform in a public speaking contest
g. To determine the winner in a history essay contest
h. The general verbal and nonverbal abilities of a student with attention deficit disorder
i. The extent to which elementary teachers use negative reinforcement in the classroom, and the effect of that reinforcement on students' behavior
j. The problems faced by minority students during the first year at a large research university
k. How parents in a school system feel about moving the sixth grade from the elementary school to the middle school

5. How would you measure parents' attitudes toward a new dress code proposed for a middle school?
6. What are some procedures for increasing the accuracy of direct observation techniques?
7. Construct a five-item Likert scale for measuring peoples' attitudes toward stem cell research.
8. Intelligence tests can most accurately be described as
   a. Measures of innate mental capacity
   b. Academic achievement measures
   c. Reading tests
   d. Scholastic aptitude tests
9. List and briefly describe the instruments available for recording data in observational research.
10. What type of instrument would be most appropriate to measure each of the following?
    a. To determine if high school chemistry students can use laboratory scales to weigh specified amounts of a given chemical compound
    b. How students in the various elementary schools in Brown County compare in math skills
    c. How parents feel about an extended school day for elementary schools in the district
    d. The general verbal and nonverbal abilities of a child with dyslexia
    e. To study bullying in an elementary classroom
    f. To get a major professor's evaluation of the potential of a student for advanced work in chemistry
    g. To get a quick measure of students' attitudes toward the extracurricular programs available at the school

ANSWERS

1. *Standardized* refers to instruments for which comparative norms have been derived, their reliability and validity have been established, and directions for administration and scoring have been prescribed.
2. In judging an individual on a comparative rating scale, the rater must have knowledge of the group with which the individual is being compared. In judging an individual on graphic and category scales, raters do not make a direct comparison of the subject with other people.
3. Raters may be less than objective in judging individuals when influenced by such tendencies as the halo effect, the generosity error, the error of severity, or the error of central tendency.
4. a. Attitude scale
   b. Aptitude test (group)
   c. Performance test
   d. Standardized reading achievement test
   e. Inventory
   f. Rating scale (performance test)
   g. Performance test
   h. Aptitude or intelligence test (individual)
   i. Direct observation
   j. Inventory
   k. Attitude scale

5. Construct a Likert scale containing approximately 20 statements expressing positive and negative feelings about the proposed dress code or construct a bipolar adjective scale.
6. The behaviors to be observed must be specified; behaviors falling within a category must be defined; a system for quantification must be developed; and the observers must be trained to carry out the observations according to this established procedure.
7. Answers will vary.
8. d
9. Checklists indicate the presence or absence of certain behaviors. Rating scales and coding schemes both yield quantitative measures. In ratings, the person indicates his or her judgment of the behavior on a continuum. Ratings are sometimes completed in retrospect. Coding schemes are used to categorize observed behavior as it occurs.
10. **a.** Performance test
    b. Standardized achievement test
    c. Attitude scale
    d. Individual intelligence test, such as the Wechsler
    e. Observation
    f. Comparative rating scale
    g. Bipolar adjective scale

REFERENCES

Geisinger, K., Spies, R., Carlson, J., & Plake, B. (Eds.). (2007). *The Seventeenth Mental Measurements Yearbook*. Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Erlbaum.

Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character: Studies in deceit* (Vol. 1). New York: Macmillan. [Reprinted in 1975 by Ayer, New York]

Kubiszyn, T., & Borich, G. (2006). *Educational testing and measurement*. New York: Wiley.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, no. 140.

Marsh, H. W. (1988). *Self-description questionnaire: A theoretical and empirical basis for the measurement of multiple dimensions of preadolescent self-concept: A test manual and a research monograph*. San Antonio, TX: The Psychological Corporation.

Miller, A., Gouley, K., & Seifer, R. (2004). Emotions and behaviors in the Head Start classroom: Associations among observed dysregulation, social competence, and preschool adjustment. *Early Education and Development*, *15*(2), 147–165.

Mueller, D. J. (1986). *Measuring social attitudes: A handbook for researchers and practitioners*. New York: Teachers College Press.

Murphy, L., Plake, B., & Spies, R. (Eds.). (2006). *Tests in print VII: An index to tests, test reviews, and the literature on specific tests*. Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Newborg, J., Stock, J., Wnek, L., Guidubaldi, J., & Svinicki, J. (1988). *The Battelle Developmental Inventory (BDI)*. Chicago: Riverside.

Noldus Information Technology. (1995). *The observer: System for collection and analysis of observational data* (Version 3.0). Sterling, VA: Author.

Noldus Information Technology. (2008). *The observer XT 8.0*. Sterling, VA: Author.

Osgood, C. E., Suci, G. J., & Tannenbaum, P.H. (1967). *The measurement of meaning*. Urbana: University of Illinois Press.

Parten, M. B. (1932). Social participation among preschool children. *Journal of Abnormal Social Psychology*, *27*, 243–269.

Popham, W. J. (2005). *Classroom assessment: What teachers need to know*. Boston: Allyn & Bacon.

Shavelson, R. J., Huber, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, *46*, 407–441.

Skinner, M., Buysse, V., & Bailey, D. (2004). Effects of age and developmental status of partners on play of preschoolers with disabilities. *Journal of Early Intervention*, *26*(3), 194–203.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. Upper Saddle River, NJ: Pearson Education.

Information derived from measuring instruments ranges from excellent to useless to downright misleading. There are systematic ways to assess the usefulness of the scores derived from measuring instruments.

CHAPTER 9

Validity and Reliability

INSTRUCTIONAL OBJECTIVES

After studying this chapter, the student will be able to:

1 Distinguish between validity and reliability.

2 List the major types of evidence used to support the valid interpretation of test scores.

3 Define construct underrepresentation and construct-irrelevant variance and explain their relevance to the validity of test scores.

4 Distinguish between convergent and discriminant evidence of validity.

5 Distinguish between random and systematic errors of measurement and their relationship to validity and reliability of test scores.

6 State the different sources of random error in educational and psychological measures.

7 Describe the different procedures (test–retest, equivalent forms, split-half, Kuder–Richardson, and others) for estimating the reliability of a measure.

8 Compute reliability coefficients for given data.

9 Define interobserver reliability and explain how it is calculated.

10 Apply the Spearman–Brown formula to determine the effect of lengthening a test on test reliability.

11 Explain the factors affecting the size of a reliability coefficient.

12 Compute the standard error of measurement and interpret score bands as indications of reliability.

13 Compute indexes to show the reliability of a criterion-referenced test.

Quantitative research always depends on measurement. Chapter 8 introduced you to some of the measuring instruments used in research. Two very important concepts that researchers must understand when they use measuring instruments are *validity* and *reliability*. Validity is defined as the extent to which scores on a test enable one to make meaningful and appropriate interpretations. Reliability indicates how consistently a test measures whatever it does measure. Researchers must be concerned about the validity and reliability of the scores derived from instruments used in a study and must include this information in the research report. If a researcher's data are not obtained with instruments that allow valid and reliable interpretations, one can have little faith in the results obtained or in the conclusions based on the results.

VALIDITY

Validity is the most important consideration in developing and evaluating measuring instruments. Historically, **validity** was defined as the extent to which an instrument measured what it claimed to measure. The focus of recent views of validity is not on the instrument itself but on the interpretation and meaning of the scores derived from the instrument. The most recent *Standards for Educational and Psychological Testing* (1999),* prepared by the American Educational Research Association, the National Council on Measurement in Education, and the American Psychological Association, defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Measuring instruments yield scores; however, the important issue is the interpretation we make of the scores, which may or may not be valid. For example, a fourth-grade math test that might allow a teacher to make valid interpretations about the math achievement of her fourth-grade students would not yield valid interpretations about the fourth-graders' abilities to solve algebra problems. If one tried to use the math achievement test for this purpose, it would be the interpretations about the students' ability to solve algebra problems that would be invalid, not the test. Thus, we no longer speak of the validity of the instrument but, rather, the validity of the interpretations or inferences that are drawn from the instrument's scores. Validity does not travel with the instrument. A test may be valid for use with one population or setting but not with another.

Assessing the validity of score-based interpretations is important to the researcher because most instruments used in educational and psychological investigations are designed for measuring hypothetical constructs. Recall that constructs such as intelligence, creativity, anxiety, critical thinking, motivation, self-esteem, and attitudes represent abstract variables derived from theory or observation. Researchers have no direct means of measuring these constructs such as exist in the physical sciences for the measurement of characteristics such as length, volume, and weight. To measure these hypothetical constructs, you must move from the theoretical domain surrounding the construct to an empirical level that operationalizes the construct. That is, we use an operational definition to measure the construct. We do this by selecting specific sets of observable tasks believed to serve as indicators of the particular theoretical construct. Then we assume that performance (scores) on the tasks reflects the particular construct of interest as distinguished from other constructs. Essentially, validity deals with how well the operational definition fits with the conceptual definition.

Tests may be imprecise measures of the constructs they are designed to assess because they leave out something that theory states should be included, include something that should be left out, or both. Messick (1995) identified two problems that threaten the interpretation (validity) of test scores: construct underrepresentation and construct-irrelevant variance. The term **construct underrepresentation** refers to assessment that is too narrow and fails to include important dimensions of the construct. The test may not adequately sample some kinds of content or some types of responses or psychological processes and thus fails to adequately represent the theoretical domain of the construct. Individuals'

*The 1999 edition of the *Standards* is currently being revised (see www.apa.org/science/standards.html).

scores on a math test may be misleading because the test did not measure some of the relevant skills that, if represented, would have allowed the individuals to display their competence. Or a scale designed to measure general self-concept might measure only social self-concept and not academic and physical components of self-concept.

The term **construct-irrelevant variance** refers to the extent to which test scores are affected by variables that are extraneous to the construct. Low scores should not occur because the test contains something irrelevant that interferes with people's demonstration of their competence. Construct-irrelevant variance could lower scores on a science achievement test for individuals with limited reading skills or limited English skills. Reading comprehension is thus a source of construct-irrelevant variance in a science achievement test and would affect the validity of any interpretations made about the individuals' science achievement.

VALIDATION

The process of gathering evidence to support (or fail to support) a particular interpretation of test scores is referred to as validation. We need evidence to establish that the inferences, which are made on the basis of the test results, are appropriate. Numerous studies may be required to build a body of evidence about the validity of these score-based interpretations. The *Standards for Educational and Psychological Testing* lists three categories of evidence used to establish the validity of score-based interpretations: evidence based on content, evidence based on relations to a criterion, and construct-related evidence of validity. Using these categories does not imply that there are distinct types of validity but, rather, that different types of evidence may be gathered to support the intended use of a test. The categories overlap and all are essential to a unitary concept of validity.

1. Evidence Based on Test Content

Evidence based on test content involves the test's content and its relationship to the construct it is intended to measure. The *Standards* defines content-related evidence as "The degree to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content." That is, the researcher must seek evidence that the test to be used represents a balanced and adequate sampling of all the relevant knowledge, skills, and dimensions making up the content domain. Evidence based on test content is especially important in evaluating achievement tests. In this age of educational accountability, content validity is receiving renewed attention. Crocker (2003) wrote, "When scores are used for educational accountability, the 'load-bearing wall' of that validity argument is surely content representativeness" (p. 7). Validation of an achievement test, for instance, would consider the appropriateness of the test's content to the total content area to be measured as well as how adequately the test samples the total domain. One would not attempt to measure chemistry students' knowledge of oxidation, for example, with only two questions.

A researcher who wished to develop a test in fourth-grade mathematics for a particular school system would need to determine what kinds of content (skills and objectives) are covered in fourth-grade classes throughout the system. After examining textbooks, syllabi, objectives, and talking with teachers, the researcher would prepare an outline of the topics, computational and conceptual skills, and performances that make up fourth-grade mathematics (content domain) in that system, along with an indication of the emphasis given to each. Using the outline as a guide, the researcher would write a collection of test items that cover each topic and each objective in proportion to the emphasis given to each in the total content domain. The result should be a representative sample of the total domain of knowledge and skills included in that school system's fourth-grade math classes.

If a math test were designed to be used nationally, the researcher would need to examine widely used textbooks, states' curriculum guides, syllabi, and so on throughout the country to determine what content (concepts and skills) is included in fourth-grade math. The test content would be sampled to provide a representative and balanced coverage of this national curriculum. Subject matter experts and curriculum specialists would be asked to judge the adequacy of the test's content for measuring fourth-grade math achievement. Developers of nationally used achievement tests, such as the Stanford Achievement Test, are expected to provide extensive evidence of content validity. If a publisher says a test measures reading comprehension, for example, then the publisher should provide evidence that higher scores on the test are attributable to higher levels of reading comprehension rather than, for example, better memory.

To ensure content validity in a classroom test, a teacher should prepare a "blueprint" showing the content domain covered and the relative emphasis given to each aspect of the domain. If the pie chart in Figure 9.1 represents a teacher's assessment of the relative importance of topics within a unit on the American Civil War, a 50-item exam should include 10 items on topic A; 5 each on B, C, D, and G; 12 on E; and 8 on F.

There is no numerical index to indicate content validity. Evidence based on content is mainly the result of a logical examination or analysis by content experts that shows whether the instrument adequately represents the content

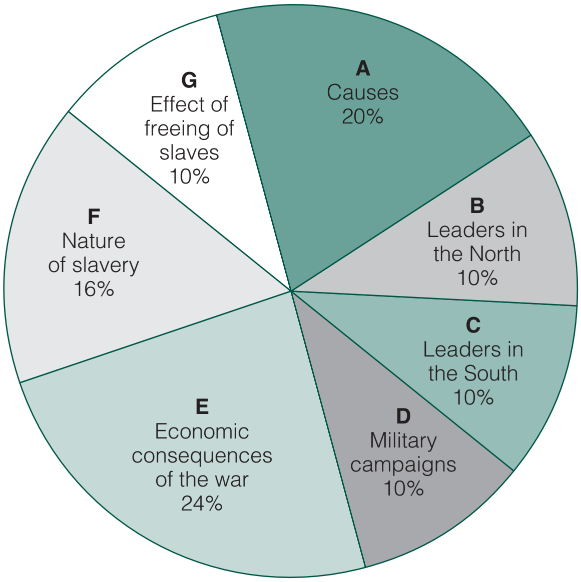


Figure 9.1 Unit on the American Civil War

and objectives making up the domain. An achievement test may have content validity when used for the purposes defined by the test maker but not yield valid interpretations for a user who defines the content domain in a different way. Only the user of a test can ultimately judge its validity for his or her purpose. Brennan (2001) stated, "For test users, the single best thing to do in advancing proper score use and interpretation is to take the test, or at least, study its content" (p. 12).

Although **content-related validity evidence** is especially important for achievement tests, it is also a concern for other types of measuring instruments, such as personality and aptitude measures. An instrument for measuring attitudes toward capital punishment, for example, would be examined to ensure that it contains, in sufficient number, a balanced set of positive and negative statements about capital punishment. An academic aptitude test should measure skills and abilities judged to be important to success in academic tasks. If you were developing a test to select among applicants for a particular job, you would need to specify all the major aspects of the job and then write test items that measure each aspect.

Face validity is a term sometimes used in connection with a test's content. Face validity refers to the extent to which examinees believe the instrument is measuring what it is supposed to measure. The question is, "on the face of it," does the test appear to be valid? Although it is not a technical form of validity, face validity can be important to ensure acceptance of the test and cooperation on the part of the examinees. Students taking a test to qualify for an advanced chemistry class would not expect it to contain items dealing with world history or geography.

2. Evidence Based on Relations to a Criterion

Criterion-related validity evidence refers to the extent to which test scores are systematically related to one or more outcome criteria. The emphasis is on the criterion because one will use the test scores to infer performance on the criterion. Historically, two types of criterion-related validity evidence have been distinguished: concurrent and predictive. The distinction is made on the basis of the time the criterion data are collected.

Concurrent Validity **Concurrent validity evidence** is the relationship between scores on a measure and criterion scores obtained at the same time. Assume a researcher has developed a foreign language aptitude test and needs evidence that the test really measures foreign language aptitude. The researcher could select a well-known and previously validated foreign language aptitude test, administer it and the new test to a group of students, and determine the correlation between the two sets of scores. A substantial correlation between the new test and the widely accepted test is evidence that the new test is also measuring foreign language aptitude. Other criteria available at the time might be current grades in a foreign language class or scores on a teacher-made test. Or assume a researcher at Educational Testing Service has developed a new scholastic aptitude test that might replace the more expensive Scholastic Assessment Test (SAT). In order to obtain evidence about the meaningfulness of the scores from this new test, the researcher would administer both the new test and the SAT (the criterion) to a representative sample of high school students. A substantial correlation between the two sets of scores would indicate that inferences made on the basis of the new test's scores would have validity for measuring scholastic aptitude. A low correlation would indicate that the validity of the inferences based on the new test's scores would be suspect. One would not consider the test a worthwhile replacement for the SAT.

Predictive Validity **Predictive validity evidence** is the relationship between scores on a measure and criterion scores available at a future time. In gathering predictive validity evidence of a foreign language aptitude test, one would look at the relationship between scores on the test and the grades students eventually earned in a future foreign language course (criterion). If a relationship is demonstrated, the scores on the aptitude test could be used later to predict performance in foreign language courses. In the case of a new scholastic aptitude test, predictive validity evidence would involve administering the test to a sample of high school juniors or seniors and then putting the scores away until the students complete their first semester or two of college. When the students' college grade point averages (GPAs) become available, one would correlate the test scores and GPAs. If the correlation were high, one has evidence for the usefulness of the aptitude test for predicting college achievement. Large numbers of high school students take the SAT or the ACT test each year because evidence has revealed a correlation between SAT and ACT scores and freshman college GPA. Likewise, the GRE is used in the selection process for admission to graduate school because there is evidence that scores on the GRE are correlated with achievement in graduate school and thus have validity for predicting future achievement. Figure 9.2 illustrates concurrent- and predictive-related evidence used in the validation of an aptitude test.

Choosing the Criterion The choice of the criterion and its measurement are crucial in criterion-related evidence. What does one look for when choosing a criterion?

1. The worth of the entire procedure depends first and foremost on the *relevance* of the criterion. The criterion must well represent the attribute being measured or else it would be meaningless to use it. For example, GPA is considered a relevant measure of success in college and is generally chosen

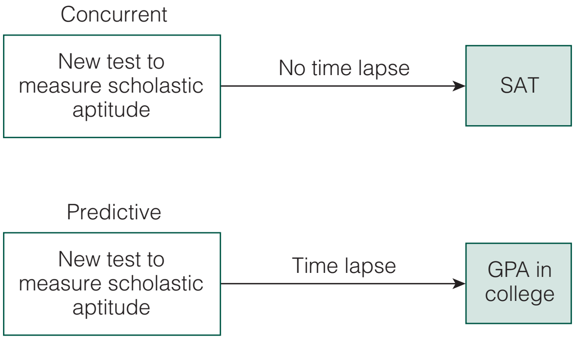

Concurrent
New test to measure scholastic aptitude
No time lapse
SAT
Predictive
New test to measure scholastic aptitude
Time lapse
GPA in college

Figure 9.2 Criterion-Related Evidence of Validity

as the criterion for validation studies of scholastic aptitude tests. A relevant criterion for a test designed to select salespeople might be the dollar value of sales made in a specified time. Supervisor ratings might be used as a criterion in the validation of a test designed to predict success in data-entry positions at a corporation. It is sometimes difficult to find a relevant criterion measure, as in the validation of measures designed to predict teacher effectiveness. With neither an agreed-on description of teacher effectiveness nor an effective method of measuring that variable, it is extremely difficult to validate such instruments.

2. The criterion must also first be *reliable*, which means that it is a consistent measure of the attribute over time or from situation to situation. If the criterion is not consistent, you would not expect it to relate consistently to any tests.
3. The criterion should also be *free from bias*, which means that the scoring of the criterion measure itself should not be influenced by any factors other than actual performance on the criterion. For example, if ratings are used as the criterion, it is essential that the raters be trained and be very careful not to let any factors other than actual performance influence their ratings. The criterion may also be biased through contamination, which occurs when scores on the criterion are influenced by the scorer's knowledge of the individuals' predictor test scores. For example, assume that the criterion used to validate an art aptitude test is grades in art class. If the teachers who grade the students' work are aware of the students' scores on the aptitude test, this awareness may influence the teachers' evaluation of the students and hence the grades. This type of contamination of the criterion can be prevented by not permitting the person who grades or rates the criterion to see the original scores on the test.

Validity Coefficient

The coefficient of correlation between test scores and criterion is called a **validity coefficient** (r_{xy}). Like any correlation coefficient, the size of a validity coefficient is influenced by the strength of the relationship between test and criterion and the range of individual differences in the group. As usual, the nearer the coefficient is to 1.00 (+ or −), the stronger the evidence is that the test is useful for the stated purpose.

Validity coefficients indicate whether the test will be useful as a predictor or as a substitute measure. If it has been shown that a test has a high correlation with a future criterion, then that test can later be used to predict that criterion. Accumulating predictive evidence requires time and patience. In some cases, researchers must wait for several years to determine whether performance on a measure is useful for predicting success on a criterion.

Concurrent criterion-related validity evidence is important in tests used for classification, certification, or diagnosis. For example, one would seek concurrent validity evidence for a new psychiatric screening device by examining its correlation with a well-established instrument already available. If there is a substantial correlation between the new test and the established instrument, one would assume they are measuring the same construct, and the new test could be used as a substitute for the older instrument. Concurrent validity evidence is necessary when new tests are designed to replace older tests that may be more expensive or more difficult and time-consuming to administer.

Students often ask, "How high does a validity coefficient need to be?" As a general rule, the higher the validity coefficient, the better the evidence. But whether high or low, useful or not useful, depends on the purpose of the test and the context in which it is to be used. A correlation coefficient of .40 could be very helpful in cases for which no predictive instrument has previously been available. In other cases, a correlation of .65 might be considered low and unsatisfactory if other predictors are available that have a higher relationship with the criterion. In general, an instrument has "good" validity as a selection device if evidence shows it has a higher correlation with the criterion than do competing instruments.

3. Construct-Related Evidence of Validity

Construct-related evidence of validity focuses on test scores as a measure of a psychological construct. To what extent do the test scores reflect the theory behind the psychological construct being measured? Recall that psychological constructs such as intelligence, motivation, anxiety, or critical thinking are hypothetical qualities or characteristics that have been "constructed" to account for observed behavior. They cannot be seen or touched or much less measured directly. How does one know that a measure of a particular construct really reflects this hypothetical characteristic? The test developer of such a measure would have to provide evidence that the scores really reflect the construct in question. The process begins with a definition of the construct based on the theory and previous research. The test developer then specifies the aspects of the construct that are to be measured in the test and develops items that require test takers to demonstrate the behaviors that define the construct. One collects any logical and empirical evidence that supports the assertion that a test measures the construct as defined and not something else. Construct-related evidence is more comprehensive than content- and criterion-related evidence and subsumes the other types. In general, any information that sheds light on the construct being measured is relevant.

Let us consider some of the strategies used to gather construct-related evidence.

1. *Related measures studies:* The aim is to show that the test in question measures the construct it was designed to measure and not some other theoretically unrelated construct. The *Standards* (1999) distinguishes two types of evidence based on relations to other variables: **convergent** and **discriminant.** "Relationships between test scores and other measures intended to assess *similar* constructs provide convergent evidence, whereas relationships between test scores and measures purportedly of *different* constructs provide discriminant evidence" (*Standards*, p. 14). In the case of convergent evidence, the researcher tries to show that the intended construct is being measured; in the case of **divergent evidence,** he or she shows that a wrong construct is not being measured. A mathematical reasoning test would be expected to correlate with grades in mathematics or with other math reasoning tests (convergent evidence). The math test and these

other measures correlate because they all converge on the same construct. Conversely, the scores on the math reasoning test would be expected to have little or no relationship (discriminant evidence) with measures of other skills, such as reading. If a substantial correlation is found between the math test and the reading test, then the math test is being affected by reading ability, and instead of measuring mathematical reasoning, it is really measuring reading ability. Such evidence would lead one to conclude that the math test is not measuring the intended construct (math reasoning) and thus would not yield valid interpretations about math reasoning. Of course, a mathematical reasoning test will inevitably involve some reading skill, so one would not expect a zero correlation with a reading test. However, if two mathematical reasoning tests are alike in all other aspects, the one with a correlation of .15 with a reading test would be preferred over the one with v correlation of .35.

In a classic article, Campbell and Fiske (1959) discussed a **multitrait–multimethod matrix** (MTMM) of corrrelation coefficients as a straightforward way to simultaneously evaluate convergent and discriminant validity of a construct. Their approach was based on the belief that measures of the same construct should correlate with each other even if they use different methods (convergent validity), and that measures of different constructs should not correlate with each other even if they employ the same method (discriminant validity). To illustrate, let's assume a researcher has a theory about a personality characteristic called teacher warmth and has developed an attitude scale as a measure of this construct. In order to establish its construct validity, he or she would need to show not only that it converges with other measures of teacher warmth but also that it could be distinguished from other teacher traits such as sociability. The researcher could administer the attitude scale (method A) to assess teacher warmth and also assess teacher warmth through face-to-face interviews (method B) with the same group of participants. Sociability would similarly be measured in two ways: scores on an existing attitude scale designed to measure sociability (method C) and through face-to-face interviews (method D) with the same participants.

The next step is to calculate the intercorrelations of the participants' scores on all four measures and present the intercorrrelations in what is called a multitrait–multimethod matrix. Table 9.1 shows hypothetical correlations between teacher warmth measured by A, attitude scale, and B,

Table 9.1 Multitrait–Multimethod Matrix of Correlations between Two Teacher Traits across Two Methods of Measurement

		Warmth		Sociability	
		Scale	**Interview**	**Scale**	**Interview**
		A	B	C	D
Warmth	Scale	A	.75	.30	.10
	Interview	B		.25	.20
Sociability	Scale	C			.70
	Interview	D			

interview, and sociability measured by C, attitude scale, and D, interview. Let us look at the correlations that are relevant to the construct validity of the attitude scale for measuring teacher warmth.

The high correlation of .75 between teacher warmth measured by method A (attitude scale) and by method B (interview) is evidence of convergent validity. The low correlations of .30 between teacher warmth measured by method A and sociability measured by method C and .10 between method A and method D are evidence of divergent validity of the teacher warmth measure. These data provide evidence for the construct validity of the teacher warmth attitude scale. Of course, one would want to conduct further analyses involving more traits and more measures to determine if the pattern of correlations fits the theory behind the constructs. The rule is as follows: If there is a good fit between the theory and the data, then keep both the theory and the measures. If not, you need to revise the theory or the measures or both. The previous example, using only two traits and two methods, is the simplest possible form of a multitrait–multimethod analysis.

2. *Known-groups technique:* Another procedure for gathering construct-related evidence is the **known-groups technique,** in which the researcher compares the performance of two groups already known to differ on the construct being measured. One hypothesizes that the group known to have a high level of the construct will score higher on the measure than the group known to have a low level of the construct. If the expected difference in performance is found, one concludes that the test is measuring that construct. You would expect that scores on a musical aptitude test, for instance, would differ for students currently enrolled in a school of music versus an unselected group of college students. If an inventory measures psychological adjustment, the scores of a group previously identified as adjusted and a group previously identified as maladjusted should be markedly different on the inventory.
3. *Intervention studies:* Another strategy for gathering construct-related evidence is to apply an experimental manipulation and determine if the scores change in the hypothesized way. You would expect the scores on a scale designed to measure anxiety to increase if individuals are put into an anxiety-provoking situation. The scores of a control group not exposed to the experimental manipulation should not be affected. If anxiety were manipulated in a controlled experiment and the resulting scores change in the predicted way, you have evidence that the scale is measuring anxiety.
4. *Internal structure studies:* Analyzing the internal structure of a test is another source of evidence that the test is measuring the construct it is supposed to be measuring. This procedure involves showing that all the items making up the test or scale are measuring the same thing—that is, that the test has internal consistency. We would expect that individuals who answer some questions in a certain way would also answer similar questions in the same way. In a scale measuring attitudes toward stem cell research, for instance, one would determine if individuals who support stem cell research were consistent in their agreeing with positive statements and disagreeing with the negative statements in the scale. A procedure called **factor analysis**

provides a way to study the constructs that underlie performance of a test. Factor analysis calculates the correlations among all the items and then identifies factors by finding groups of items that are correlated highly with one another but have low correlations with other groups. More than one factor may be needed to account for the correlations among the items. You then decide if the observed intercorrelations conform to the theory behind the construct being measured. If the theory suggests a single one-dimension construct, then we look for high intercorrelations among all the items. If the theory suggests more than one dimension, we should have subscales to measure each separate dimension. In that case, the subscales should have high internal consistency, but they should not correlate highly with other subscales. A measure of feminism, for example, would probably have several subscales covering family, work, pay, politics, authority relations, and the like. The extent to which the observed item intercorrelations agree with the theoretical framework provides evidence concerning the construct being measured. Further discussion of factor analysis is presented in Chapter 13.

5. *Studies of response processes:* Another way to obtain evidence about how well a test is measuring the construct of interest is to look at the **evidence based on response processes** of individuals actually taking the test. Questioning test takers about the mental processes and skills that they use when responding to the items of a test can provide information about what construct is being measured. If one were gathering validity evidence about a new verbal reasoning test, for instance, one might ask individuals to "think aloud" as they work through the test. This procedure may reveal that the test is measuring verbal reasoning, or it may reveal that other factors such as vocabulary or reading comprehension are being measured. Examining response processes may indicate certain construct-irrelevant factors that differentially influence the performance of different subgroups. Thus, it provides evidence about whether the test scores have the same meaning or can be interpreted in the same way across different subgroups. Table 9.2 summarizes the three major types of evidence for validity.

VALIDITY GENERALIZATION

A concern in validity studies of educational and employment tests is the extent to which evidence of validity based on test–criterion relationships can be generalized to new settings without further investigations of validity in the new setting. Research shows that test–criterion correlations may vary greatly from time to time and place to place because of the type of criterion measure used, the way the predictor is measured, the type of test takers, and the time period involved.

Validity generalization studies have used meta-analysis, which provides statistical summaries of past validation studies in similar situations. If the meta-analytic database is large and the studies adequately represent the type of situation to which a researcher wishes to generalize, we find support for validity generalization. In other circumstances in which the findings of the meta-analytic studies are less consistent and in which there are more differences between the new and old settings, it is more risky to generalize. Local validation studies providing situation-specific evidence would be more valuable.

Table 9.2 Types of Evidence for Validity of a Test

Type	Question	Method
Content related	Is the test a representative sample of the domain being measured?	Make a logical analysis of the content to determine how well it covers the domain.
Criterion related (concurrent)	Does a new test correlate with a currently available test (criterion) so that the new test could be a substitute?	Correlate scores from new test with scores of a criterion available at the time.
Criterion related (predictive)	Does a new test correlate with a future criterion so that the test can be used to predict later performance on the criterion?	Correlate test scores with a measure (criterion) available at a future time.
Construct related	Does the test really measure the intended construct?	Gather various kinds of evidence: convergent and divergent evidence, known-groups technique, intervention study, internal structure, and response processes.

VALIDITY OF CRITERION-REFERENCED TESTS

Recall that criterion-referenced tests are designed to measure a rather narrow body of knowledge or skills. Thus, the main concern in assessing the validity of criterion-referenced tests is *content validity*. The basic approach to determining content validity is to have teachers or subject matter experts examine the test and judge whether it is an adequate sample of the content and objectives to be measured.

Another approach that has been used is to administer the test and divide the examinees into two groups: masters versus nonmasters. Then, one determines the proportion of examinees in each group who answered each item correctly. Valid items are those for which the success rate in the master group is substantially higher than the success rate in the nonmaster group. To be very strict, the success rate on each item should be 100% for masters, whereas nonmasters have a very low or a chance rate of success. In the ideal test, there should be no misclassifications (Thorndike, 2005, p. 192).

APPLICATION OF THE VALIDITY CONCEPT

Validity is always specific to the particular purpose for which the instrument is being used. “It is incorrect to use the unqualified phrase ‘the validity of the test.’ No test is valid for all purposes or in all situations” (*Standards*, 1999, p. 17). Validity should be viewed as a characteristic of the interpretation and use of test scores and not of the test itself. A test that has validity in one situation and for one purpose may not be valid in a different situation or for a different purpose. A teacher-made achievement test in high school chemistry might be useful for measuring end-of-year achievement in chemistry but not useful for predicting achievement in college chemistry. A German proficiency test might be appropriate for placing undergraduates in German classes at a university but not be a valid exit exam for German majors. Thus, validation is always a responsibility of the test user as well as of the test developer.

We have viewed "test validation" as a process of gathering different types of evidence (content, criterion-related, and construct) in support of score-based interpretations and inferences. The goal of the process is to derive the best possible case for the inferences we want to make.

THINK ABOUT IT 9.1

Identify the type of validity evidence (content, concurrent criterion, predictive criterion, or construct related) being gathered in each of the following examples:

a. A test administered to applicants for law school correlates .65 with first semester grades in law school.
b. A group of math professors examine the math placement test administered to freshmen at the university. They conclude that the test is an excellent sample of the math skills students need to succeed in college-level courses.
c. A high school teacher administers a standardized chemistry test and correlates the scores with the scores that students earned the next day on a teacher-made chemistry test.
d. As predicted, the scores for a group of Young Republicans on a scale measuring political conservatism were markedly higher than those for a group of Young Democrats.
e. Scores on a new scale to detect depression are correlated with scores on a well-established scale measuring optimism. The correlation was negligible.

Answers

a. Predictive (criterion related)
b. Content related
c. Concurrent (criterion related)
d. Construct related
e. Construct related (theory would predict that measures of depression would not be correlated with measures of optimism; this is divergent construct-related validity evidence)

RELIABILITY

As we mentioned at the beginning of this chapter, the **reliability** of a measuring instrument is the degree of consistency with which it measures whatever it is measuring. This quality is essential in any kind of measurement. A post office will soon take action to repair a scale if it is found that the scale sometimes underestimates and sometimes overestimates the weight of packages. A bathroom scale would be reliable if it gives you nearly the same weight on five consecutive days. However, if you got vastly different readings on each of the five days, you would consider the scale unreliable as a measure of your weight and would probably replace it. Psychologists and educators are concerned about the consistency of their measuring devices when they attempt to measure such complex constructs as scholastic aptitude, achievement, motivation, anxiety, and the like. They would not consider a scholastic aptitude test worthwhile if it yielded markedly different results when administered to the same students on two occasions within the same time frame. People who use such measuring instruments must identify and use techniques that will help them determine to what extent their measuring instruments are consistent and reliable.

On a theoretical level, reliability is concerned with the effect of error on the consistency of scores. In this world measurement always involves some error. There are two kinds of errors: **random errors of measurement** and **systematic errors of measurement.** Random error is error that is a result of pure chance. Random errors of measurement may inflate or depress any subject's score in an unpredictable manner. Systematic errors, on the other hand, inflate or depress scores of identifiable groups in a predictable way. Systematic errors are the root of validity problems; random errors are the root of reliability problems.

SOURCES OF RANDOM ERROR

Chance or random error that leads to inconsistency in scores can come from three sources:

1. *The individual being measured may be a source of error*. Fluctuations in individuals' motivation, interest, level of fatigue, physical health, anxiety, and other mental and emotional factors affect test results. As these factors change randomly from one measurement to the next, they result in a change or inconsistency in one's scores. Individuals may make more lucky guesses at one time than another. A student's breaking a pencil point on a timed test would increase the error component in the test results.
2. *The administration of the measuring instrument may introduce error*. An inexperienced person may depart from standardized procedures in administering or scoring a test. Testing conditions such as light, heat, ventilation, time of day, and the presence of distractions may affect performance. Instructions for taking the test may be ambiguous. The scoring procedure may be a source of error. Objectivity and precise scoring procedures enhance consistency, whereas subjectivity and vague scoring instructions depress it.
3. *The instrument may be a source of error*. Brevity of a test is a major source of unreliability. A small sample of behavior results in an unstable score. For example, if a test is very short, those subjects who happen to know the few answers required will get higher scores than they deserve, whereas those who do not know those few answers will get lower scores than they deserve. For example, if a test is given to assess how well students know the capitals of the 50 states but only five questions are asked, it is possible that a student who knows only 10 capitals could get all five questions correct, whereas a student who knows 40 could get none correct. Luck is more of a factor in a short test than in a long test.

If a test is too easy and everyone knows most of the answers, students' relative scores again depend on only a few questions and luck is a major factor. If questions are ambiguous, "lucky" examinees respond in the way the examiner intended, whereas "unlucky" subjects respond in another equally correct manner, but their answers are scored as incorrect.

One element in a physical fitness test for elementary students is the baseball throw. Subjects are instructed to throw a baseball as far as they can, and the distance of the throw is measured. Although the object of the test is to get a score that is typical of a subject's performance, certainly if you had a single subject throw a baseball on several occasions, you would find that the child does not throw it the same distance every time.

Assume you had each student make a throw on two consecutive days. If you then compared the two scores (distance thrown) for each student, you would find that they were almost never exactly the same. Most of the differences would be small, but some would be moderately large and a few would be quite large. Because the results are inconsistent from one day's throw to the next, one throw is not completely reliable as a measure of a student's throwing ability. Three types of chance, or random, influences lead to inconsistency between scores on the two days:

1. The student may change from one time to another. On one day he or she may feel better than on the other. On one day the student may be more motivated or less fatigued. Maybe the student loses balance when starting to throw the ball, or maybe his or her fingers slip while gripping the ball. Perhaps the student's father, hearing about the task, decides to coach the child in throwing a baseball before the next day.
2. The task may change from one measurement to the next. For example, the ball used one day may be firm, whereas on the second day it may be wet and soggy. One day perhaps the examiner permits the students to take a running start up to the throwing line, whereas on the second day a different examiner permits only a couple of steps. There may be gusts of wind at certain times that help some students more than others.
3. The limited sample of behavior results in a less reliable score. The average of a student's baseball throw scores on two days would yield a better estimate of his or her true baseball throwing skill than one day's score. The average of three days' scores would be a still better estimate and so on.

Reliability is concerned with the effect of such random errors of measurement on the consistency of scores. But some errors involved in measurement are predictable or systematic. Using the example of the baseball throw, imagine a situation in which the instructions for the throw are given in English but not all the subjects understand English. The scores of the non-English-speaking subjects could be systematically depressed because the subjects do not comprehend what they are expected to do. Such systematic errors of measurement are a validity problem. The validity of score-based inferences is lowered whenever scores are systematically changed by the influence of anything other than what you are trying to measure (irrelevant variance). In this instance, you are measuring not only baseball-throwing skill but also, in part, English comprehension.

To decide whether you are dealing with reliability or validity, you determine whether you are considering random errors or systematic errors. If a class is being given the baseball throw test and two balls are being employed, one firm and one soggy, and it is purely a matter of chance who gets which ball, the variation caused by the ball used is a reliability problem. The variation caused by the ball represents random error that affects the consistency of the measurements. If the girls are tested using a dry, firm ball and the boys get a wet, soggy ball, scores are a function of gender as well as of skill, resulting in systematic errors that give rise to a validity problem.

RELATIONSHIP BETWEEN RELIABILITY AND VALIDITY

Reliability is concerned with how consistently you are measuring whatever you are measuring. It is not concerned with the meaning and interpretation of the scores, which is the validity question. We express the relationship between these two concepts as follows: A measuring instrument can be reliable without being valid, but it cannot be valid unless it is first reliable. For example, someone could decide to measure intelligence by determining the circumference of the head. The measures might be very consistent from time to time (reliable), but this method would not yield valid inferences about intelligence because circumference of the head does not correlate with any other criteria of intelligence, nor is it predicted by any theory of intelligence. So a test can be very reliable but consistently yield scores that are meaningless.

To be able to make valid inferences from a test's scores, the test must first be consistent in measuring whatever is being measured. Reliability is a necessary but not a sufficient condition for valid interpretations of test scores.

EQUATIONS FOR RELIABILITY

It is generally accepted that all measurements of human qualities contain random error. Although scientists cannot remove all this error, they do have ways to assess the aggregate magnitude of measurement errors. Reliability procedures are concerned with determining the degree of inconsistency in scores caused by random error.

When you administer a test to a student, you get a score, which is called the **observed score.** If you had tested this student on some other occasion with the same instrument, you probably would not have obtained exactly the same observed score because of the influence of random errors of measurement. Assuming that test scores have an error component implies that there is a hypothetical error-free score for an individual that would be obtained if the measurement were perfectly accurate. This error-free value is called the individual's **true score** on the test. The true score is conceptualized as "the hypothetical average score resulting from many repetitions of the test or alternate forms of the instrument" (*Standards*, 1999, p. 25).

We conclude, therefore, that every test score consists of two components: the *true score* plus some *error of measurement*. As noted previously, this error component may be caused by any one, or a combination, of a number of factors associated with variations within the examinee from time to time or with the test and its administration.

The reliability of a test is expressed mathematically as the best estimate of what proportion of the total variance of scores on the test is true variance. As we explained in Chapter 6, variance is an index of the spread of a set of scores. If you administer a test to a group of students, some of the spread (variance) of the students' scores is due to true differences among the group and some of the spread (variance) is due to errors of measurement.

The idea of error component and true component in a single test score may be represented mathematically by Formula 9.1:

$$X = T + E \qquad (9.1)$$

where

X = observed score
T = true score component
E = error-of-measurement component

The true score component may be defined as the score an individual would obtain under conditions in which a perfect measuring device is used. The error-of-measurement component can be either positive or negative. If it is positive, the individual's true score will be overestimated by the observed score; if it is negative, the person's true score will be underestimated. Because researchers assume that an error of measurement is just as likely to be positive as it is to be negative, they can conclude that the sum of the errors and the mean of the errors would both be 0 if the same measuring instrument or an equivalent form of the instrument were administered an infinite number of times to a subject. Under these conditions, the true component would be defined as the individual's mean score on an infinite number of measurements. The true score is a theoretical concept because an infinite number of administrations of a test to the same subject is not feasible.

In the usual research situation, the investigator has one measure on each of a group of subjects, a single set of scores, to consider. Each observed score has a true score component and an error score component. It has been shown mathematically that the variance of the observed scores of a large group of subjects (σ_x^2) is equal to the variance of their true scores (σ_t^2) plus the variance of their errors of measurement (σ_e^2) or

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 \tag{9.2}$$

Reliability may be defined theoretically as the ratio of the true score variance to the observed score variance in a set of scores, as expressed by the following formula:

$$r_{xx} = \frac{\sigma_t^2}{\sigma_x^2} \tag{9.3}$$

where

r_{xx} = reliability of the test
σ_t^2 = variance of the true scores
σ_x^2 = variance of the observed scores

Reliability is the proportion of the variance in the observed scores that is free of error. This notion can be expressed in the following formula, derived from Formulas 9.2 and 9.3:

$$r_{xx} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \tag{9.4}$$

The **coefficient of reliability** r_{xx} can range from 1, when there is no error in the measurement, to 0, when the measurement is all error. When there is no error in the measurement, σ_e^2 in the reliability formula is 0 and $r_{xx} = 1$.

$$r_{xx} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \quad r_{xx} = 1 - \frac{0}{\sigma_x^2} = 1 - 0 = 1$$

If measurement is all error, $\sigma_e^2 = \sigma_x^2$ and $r_{xx} = 0$.

$$r_{xx} = 1 - \frac{\sigma_e^2}{\sigma_x^2} = 1 - 1 = 0$$

The extent of error is indicated by the degree of departure of the reliability coefficient from 1. A coefficient of .80 on a test, for example, indicates the best estimate is that 80 percent of the observed variance in the scores is true variance and 20 percent is error. Thus, the greater the error, the more the reliability coefficient is depressed below 1 and the lower the reliability. Conversely, if the reliability coefficient is near 1.00, the instrument has relatively little error and high reliability.

APPROACHES TO RELIABILITY

A test is reliable to the extent that the scores made by an individual remain nearly the same in repeated measurements. That is, individuals will have the same, or nearly the same, rank on the repeated administrations. There are two ways to express the consistency of a set of measurements.

1. The first method indicates the amount of variation to be expected within a set of repeated measurements of a *single* individual. If it were possible to weigh an individual on 200 scales, you would get a frequency distribution of scores to represent his or her weight. This frequency distribution would have an average value, which you could consider the "true" weight. It would also have a standard deviation, indicating the spread. This standard deviation is called the **standard error of measurement** because it is the standard deviation of the "errors" of measuring the weight for one person. With psychological or educational data, researchers do not often make repeated measurements on an individual. Time would not permit such repetition; in addition, the practice and fatigue effects associated with repeated measurement would have an influence on the scores. Thus, instead of measuring one person many times, researchers measure a large group on two occasions. Using the pair of measurements for each individual, they can estimate what the spread of scores would have been for the average person had the measurement been made again and again.
2. The consistency of a set of scores is also indicated by the extent to which each individual maintains the same relative position in the group. With a reliable test, the person who scores highest on a test today should also be one of the highest scorers the next time the same test is given. Each person in the group would stay in approximately the same relative position. The more individuals shift in relative position, the lower the test's reliability. You can compute a coefficient of correlation between two administrations of the same test to determine the extent to which the individuals maintain the same relative position. This coefficient is called a **reliability coefficient** (r_{xx}). A reliability coefficient of 1.00 indicates that each individual's relative position on the two administrations remained exactly the same and the test would be perfectly reliable.

Thus, the consistency of a measure is indicated by (1) its standard error of measurement or (2) its reliability coefficient. We discuss standard error of measurement later in the chapter. Let us now consider the various reliability coefficients.

RELIABILITY COEFFICIENTS

There are three broad categories of reliability coefficients used with norm-referenced tests: (1) coefficients derived from correlating individuals' scores on the same test administered on different occasions (test–retest coefficients), (2) coefficients derived from correlating individuals' scores on different sets of equivalent items (equivalent-forms coefficients), and (3) coefficients based on the relationship among scores derived from individual items or subsets of items within a test (internal-consistency coefficients). The internal-consistency coefficient requires only a single administration of a test, whereas the other coefficients require two administrations.

Test–Retest Reliability

An obvious way to estimate the reliability of a test is to administer it to the same group of individuals on two occasions and correlate the two sets of scores. The correlation coefficient obtained by this procedure is called a **test–retest reliability coefficient.** For example, a physical fitness test may be given to a class during one week and the same test given again the following week. If the test has good reliability, each individual's relative position on the second administration of the test will be near his or her relative position on the first administration of the test.

The test–retest reliability coefficient, because it indicates consistency of subjects' scores over time, is sometimes referred to as a **coefficient of stability.** A high coefficient tells you that you can generalize from the score a person receives on one occasion to a score that person would receive if the test had been given at a different time. A test–retest coefficient assumes that the characteristic being measured by the test is stable over time, so any change in scores from one time to another is caused by random error. The error may be caused by the condition of the subjects themselves or by testing conditions. The test–retest coefficient also assumes there is no practice effect or memory effect. For example, students may learn something just from taking a test and thus will react differently on the second taking of the test. These practice effects from the first testing will not likely be the same across all students, thus lowering the reliability estimate. If the interval of time is short, there may also be a memory effect; students may mark a question the same way they did previously just because they remember marking it that way the first time. This memory effect tends to inflate the reliability estimate, but it can be controlled somewhat by increasing the time between the first test and the retest. However, if the time between testings is too long, differential learning may be a problem—that is, students will learn different amounts during the interval, which would affect the reliability coefficient. Thus, the period of time between the two administrations is an issue that must be considered.

Because of these problems, the test–retest procedure is not usually appropriate for tests in the cognitive domain. Use of this procedure in schools is largely restricted to measures of physical fitness and athletic prowess.

Equivalent-Forms Reliability

Researchers use the **equivalent-forms technique** of estimating reliability, which is also referred to as the **alternate-forms technique** or **parallel-forms technique,** when it is probable that subjects will recall their responses to the test items. Here, rather than correlating the scores from two administrations of the same test to the same group, the researcher correlates the results of alternate (equivalent) forms of the test administered to the same individuals. If the two forms are administered at essentially the same time (in immediate succession), the resulting reliability coefficient is called the **coefficient of equivalence.** This measure reflects variations in performance from one specific set of items to another. It indicates whether you can generalize a student's score to what the student would receive if another form of the same test had been given. The question is, To what extent does the student's performance depend on the particular set of items used in the test? If subjects are tested with one form on one occasion and with an equivalent form on a second occasion and their scores on the two forms are correlated, the resulting coefficient is called the **coefficient of stability and equivalence.** This coefficient reflects two aspects of test reliability: variations in performance from one time to another and variations from one form of the test to another. A high coefficient of stability and equivalence indicates that the two forms are measuring the same skill and measuring consistently over time. This is the most demanding and the most rigorous measure available for determining the reliability of a test.

Designing alternate forms of a test that are truly equivalent is a challenge with this technique of estimating reliability. If a successful design is not achieved, then the variation in scores from one form to another could not be considered error variance. Alternate forms of a test are independently constructed tests that must meet the same specifications—that is, they must have the same number of items, instructions, time limits, format, content, range, and level of difficulty—but the actual questions are not the same. Ideally, you should have pairs of equivalent items and assign one of each pair to each form. In a world geography test, for example, form A might ask, "On what continent is the Nile River?" whereas form B asks, "On what continent is the Amazon River?" Form A might ask, "What is the capital of Italy?" and form B, "What is the capital of France?" The distribution of the test scores must also be equivalent.

The alternate-forms technique is recommended when you want to avoid the problem of recall or practice effect and in cases in which you have available a large number of test items from which to select equivalent samples. Researchers generally consider that the equivalent-forms procedure provides the best estimate of the reliability of academic and psychological measures.

Internal-Consistency Measures of Reliability

Other reliability procedures are designed to determine whether all the items in a test are measuring the same thing. These are called the **internal-consistency procedures** and require only a single administration of one form of a test.

Split-Half Reliability The simplest of the internal-consistency procedures, known as the *split-half*, artificially splits the test into two halves and correlates the individuals' scores on the two halves. Researchers administer the test to a group and later divide the items into two halves, obtain the scores for each individual on the two halves, and calculate a coefficient of correlation. This **split-half reliability coefficient** is like a coefficient of equivalence because it reflects fluctuations from one sample of items to another. If each subject has a very similar position on the two halves, the test has high reliability. If there is little consistency in positions, the reliability is low. The method requires only one form of a test, there is no time lag involved, and the same physical and mental influences will be operating on the subjects as they take the two halves. A problem with this method is in splitting the test to obtain two comparable halves. If, through item analysis, you establish the difficulty level of each item, you can place each item into one of the two halves on the basis of equivalent difficulty and similarity of content. The most common procedure, however, is to correlate the scores on the odd-numbered items of the test with the scores on the even-numbered items. However, the correlation coefficient computed between the two halves systematically underestimates the reliability of the entire test because the correlation between the 50 odd-numbered and 50 even-numbered items on a 100-item test is a reliability estimate for a 50-item test, not a 100-item test. To transform the split-half correlation into an appropriate reliability estimate for the entire test, the **Spearman–Brown prophecy formula** is employed:

$$r_{xx} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}} \tag{9.5}$$

where

r_{xx} = estimated reliability of the entire test
$r_{\frac{1}{2}\frac{1}{2}}$ = Pearson *r* correlation between the two halves

For example, if we find a correlation coefficient of .65 between two halves of a test, the estimated reliability of the entire test, using the Spearman–Brown formula, would be

$$r_{xx} = \frac{(2)(.65)}{1 + .65} = .79$$

The Spearman–Brown procedure is based on the assumption that the two halves are parallel. Because this assumption is seldom exactly correct, in practice, the split-half technique with the Spearman–Brown correction tends to overestimate the reliability that would be obtained with test–retest or equivalent-forms procedures. Bear this in mind when evaluating the reliabilities of competing tests.

Split-half reliability is an appropriate technique to use when time-to-time fluctuation in estimating reliability is to be avoided and when the test is relatively long. For short tests the other techniques, such as test–retest or equivalent-forms, are more appropriate. The split-half procedure is not appropriate to use with speed tests because it yields spuriously high coefficients of equivalence in such tests. A speed test is one that purposefully includes easy items so that the scores mainly depend on the speed with which subjects can respond. Errors are minor, and most of the items are correct up to the point where time is called. If a student responds to 50 items, his or her split-half score is likely to be 25–25; if another student marks 60 items, his or her split-half score is likely to be 30–30, and so on. Because individuals' scores on odd- and even-numbered items are very nearly identical, within-individual variation is minimized and the correlation between the halves would be nearly perfect. Thus, other procedures are recommended for use with speed tests.*

*There are computer programs for calculating all the reliability formulas in this chapter. We included the formulas and worked examples so you can see *how* the procedures work.

Homogeneity Measures Other internal-consistency measures of reliability do not require splitting the test into halves and scoring each half separately. These procedures assess the interitem consistency, or **homogeneity,** of the items. They reflect two sources of error: (1) the content sampling as in split-half and (2) the heterogeneity of the behavior domain sampled. The more heterogeneous the domain, the lower the interitem consistency; conversely, the more homogeneous the domain, the higher the interitem consistency.

Kuder–Richardson Procedures Kuder and Richardson (1937) developed procedures that have been widely used to determine homogeneity or internal consistency. Probably the best known index of homogeneity is the **Kuder–Richardson formula** 20 (K–R 20), which is based on the proportion of correct and incorrect responses to each of the items on a test and the variance of the total scores:

$$r_{xx} = \frac{K}{K-1}\left(\frac{s_x^2 - \Sigma pq}{s_x^2}\right) \qquad \text{K–R 20} \quad (9.6)$$

where

r_{xx} = reliability of the whole test
K = number of items on the test
s_x^2 = variance of scores on the total test (squared standard deviation)
p = proportion of correct responses on a single item
q = proportion of incorrect responses on the same item

The product pq is computed for each item, and the products are summed over all items to give Σpq. K–R 20 is applicable to tests whose items are scored dichotomously (0 or 1); thus, it is useful with test items that are scored as true/false or right/wrong. Many machine-scoring procedures for tests routinely provide a K–R 20 coefficient along with a split-half coefficient.

Another formula, **Kuder–Richardson 21,** is computationally simpler but requires the assumption that all items in the test are of equal difficulty. This assumption is often unrealistic:

$$r_{xx} = \frac{Ks_x^2 - \overline{X}(K - \overline{X})}{s_x^2(K-1)} \qquad \text{K–R 21} \quad (9.7)$$

where

r_{xx} = reliability of the whole test
K = number of items in the test
s_x^2 = variance of the scores
$\overline{X}$ = mean of the scores

This method is by far the least time-consuming of all the reliability estimation procedures. It involves only one administration of a test and employs only easily available information. As such, it can be recommended to teachers for classroom use if the test is not machine scored and the K–R 20 cannot be calculated by computer.

For example, suppose a teacher has administered a 50-item test to a class and has computed the mean as 40 and the standard deviation as 6. Applying Formula 9.7, the reliability could be estimated as follows:

$$r_{xx} = \frac{(50)6^2 - 40(50-40)}{6^2(50-1)} = \frac{1800 - 400}{1764} = .79$$

Because the Kuder–Richardson procedures stress the equivalence of all the items in a test, they are especially appropriate when the intention of the test is to measure a single trait. For a test with homogeneous content (e.g., math test covering fractions), the reliability estimate will be similar to that provided by the split-half. For a test designed to measure several traits, the Kuder–Richardson reliability estimate is usually lower than reliability estimates based on a correlational procedure.

Analysts have shown through deductive reasoning that the Kuder–Richardson reliability for any test is mathematically equivalent to the mean of the split-half reliability estimates computed for every possible way of splitting the test in half. This fact helps explain the relationship between the two procedures. If a test is of uniform difficulty and is measuring a single trait, any one way of splitting that test in half is as likely as any other to yield similar half scores. Therefore, the Spearman–Brown and Kuder–Richardson methods will yield similar estimates. If a test has items of varying difficulty and is measuring various traits, the Kuder–Richardson estimate is expected to be lower than the split-half estimate. For example, suppose a secretarial skills test samples typing, shorthand, spelling, and English grammar skills. In applying the split-half method, the test maker would assign equal numbers of items from each subtest to each half of the test. If the test is doing a good job of measuring this combination of skills, the split-half reliability will be high. The Kuder–Richardson method, which assesses the extent to which all the items are equivalent to one another, would yield a considerably lower reliability estimate.

Coefficient Alpha Another widely used measure of homogeneity is **coefficient alpha,** also called **Cronbach alpha** after Lee Cronbach, who developed it in 1951. Coefficient alpha has wider applications than the K–R 20 formula. When items are scored dichotomously, it yields the same result as the K–R 20, but it can also be used when items are not scored dichotomously. The formula for alpha is as follows:

$$\alpha = \left(\frac{K}{K-1}\right)\left(\frac{s_x^2 - \Sigma s_i^2}{s_x^2}\right) \tag{9.8}$$

where

K = number of items on the test
Σs_i^2 = sum of variances of the item scores
s_x^2 = variance of the test scores (all K items)

The formula for alpha is similar to the K–R 20 except that the Σpq is replaced by Σs_i^2, the sum of the variances of item scores. To calculate, you determine the variance of all the scores for *each* item and then add these variances across all items to get s_x^2.

Researchers use Cronbach alpha when measures have items that are not scored simply as right or wrong, such as attitude scales or essay tests. The item score may take on a range of values; for example, on a Likert attitude scale the individual may receive a score from 1 to 5 depending on which option was

Table 9.3 Summary of Reliability Coefficients

		Number of Test Forms Required	
		One	Two
Number of Administrations Required	One	Split-half K–R 20 Coefficient alpha	Equivalent forms (no time lapse)
	Two	Test–retest	Equivalent-forms (time lapse)

chosen. Similarly, on essay tests a different number of points may be assigned to each answer. Many computer programs for reliability, such as the one included in SPSS, provide a coefficient alpha as the index of reliability.

If the test items are heterogeneous—that is, they measure more than one trait or attribute—the reliability index as computed by either coefficient alpha or K–R 20 will be lowered. Furthermore, these formulas are not appropriate for timed tests because item variances will be accurate only if each item has been attempted by every person.

Table 9.3 presents a summary of the different types of reliability coefficients arranged according to the number of forms and number of administrations required.

INTERPRETATION OF RELIABILITY COEFFICIENTS

The interpretation of a reliability coefficient should be based on a number of considerations. Certain factors affect reliability coefficients, and unless these factors are taken into account, any interpretation of reliability will be superficial.

1. *The reliability of a test is in part a function of the length of the test*. Other things being equal, the longer the test, the greater its reliability. A test usually consists of a number of sample items that are, theoretically, drawn from a universe of test items. You know from what you have studied about sampling that the greater the sample size, the more representative it is expected to be of the population from which it is drawn. This is also true of tests. If it were possible to use the entire universe of items, the score of a person who takes the test would be his or her true score. A theoretical universe of items consists of an infinite number of questions and is obviously not a practical possibility. You therefore construct a test that is a sample from such a theoretical universe. The greater the number of items included in the test, the more representative it should be of the true scores of the people who take it. Because reliability is the extent to which a test represents the true scores of individuals, the longer the test, the greater its reliability, provided that all the items in the test belong in the universe of items.
2. *Reliability is in part a function of group heterogeneity*. The reliability coefficient increases as the spread, or heterogeneity, of the subjects who take the test increases. Conversely, the more homogeneous the group is with respect to the trait being measured, the lower will be the reliability coefficient. One

explanation of reliability is that it is the extent to which researchers can place individuals, relative to others in their groups, according to certain traits. Such placement is easier when you are dealing with individuals who are more heterogeneous than homogeneous on the trait being measured. It does not take a sensitive device to determine the placement of children in a distribution according to their weights when the age range of these children is from 5 to 15 years. In fact, this placement is possible with some degree of accuracy even without using any measuring device. It does take a sensitive device, however, to carry out the same placement if all those who are to be compared and placed in the distribution are 5 years old. Thus, the heterogeneity of the group with whom a measuring instrument is used is a factor that affects the reliability of that instrument. The more heterogeneous the group used in the reliability study, the higher the reliability coefficient. Keep this fact in mind when selecting a standardized test. The publisher may report a high reliability coefficient based on a sample with a wide range of ability. However, when the test is used with a group having a much narrower range of ability, the reliability will be lower.

3. *The reliability of a test is in part a function of the ability of the individuals who take that test.* A test may be reliable at one level of ability but unreliable at another level. The questions in a test may be difficult and beyond the ability level of those who take it—or the questions may be easy for the majority of the subjects. This difficulty level affects the reliability of the test. When a test is difficult, the subjects are guessing on most of the questions and a low reliability coefficient will result. When it is easy, all subjects have correct responses on most of the items, and only a few difficult items are discriminating among subjects. Again, we would expect a low reliability. There is no simple rule by which you can determine how difficult, or how easy, a test should be. That depends on the type of test, the purpose, and the population with which it will be used.

4. *Reliability is in part a function of the specific technique used for its estimation.* Different procedures for estimating the reliability of tests result in different reliability coefficients. The alternate forms with time lapse technique gives a lower estimation of reliability than either test–retest or split-half procedures because in this technique form-to-form as well as time-to-time fluctuation is present. The split-half method, in contrast, results in higher reliability coefficients than do its alternatives because of the speed element in most tests. Thus, in evaluating the reliability of a test, you would give preference to a test whose reliability coefficient has been estimated by the alternate-forms technique, rather than by other techniques, when the reported reliabilities are similar. Standardized test manuals report reliability coefficients based on test–retest and alternate-forms techniques, but teachers generally do not use these procedures for estimating reliability. Repeated testing and alternate forms are not feasible in most classroom situations. Instead, teachers use the split-half, the Kuder–Richardson, or one of the other measures of internal consistency as a measure of reliability.

5. *Reliability is in part a function of the nature of the variable being measured.* Some variables of interest to researchers yield consistent measures more

often than do other variables. For instance, because academic achievement is relatively easy to measure, most established tests of academic achievement have quite high reliability (coefficients of .90 or higher). Aptitude tests that are designed to predict future behavior—a more difficult task—have somewhat lower reliability (.80 or lower). Reliable measures of personality variables are most difficult to obtain; thus, these measures typically have only moderate reliability (.60 to .70).

6. *Reliability is influenced by the objectivity of the scoring*. Inconsistent scoring introduces error that reduces the reliability of a test. The potential unreliability of the scoring of essay tests, for example, means that essay tests are generally considered to be not as reliable as multiple-choice and other types of selected-response tests.

Table 9.4 summarizes the factors affecting reliability.

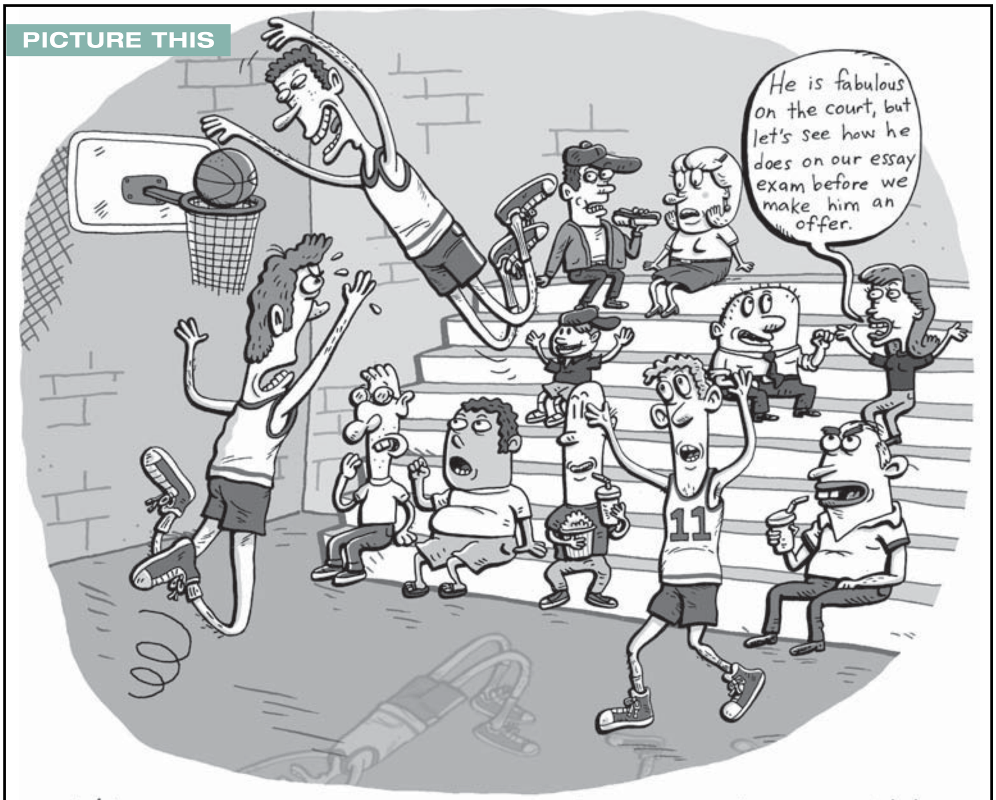
What is the minimum reliability acceptable for an instrument? Perhaps the best response to this question is that a good reliability is one that is as good as or better than the reliability of competing measures. A spelling achievement test with a reliability of .80 is unsatisfactory if competing tests have reliability coefficients of .90 or better. A coefficient of .80 for a test of creativity would be judged excellent if other tests of the same construct have reliabilities of .60 or less.

The degree of reliability you need in a measure depends to a great extent on the use you will make of the results. The need for accurate measurement increases as the consequences of decisions and interpretation become more important. If the measurement results are to be used for making a decision about a group or for research purposes, or if an erroneous initial decision can be easily corrected, scores with modest reliability (coefficients in the range of .50 to .60) may be acceptable. However, if the results are to be used as a basis for making decisions about individuals, especially important or irreversible decisions (e.g., rejection or admission of candidates to a professional school or the placement of children in special education classes), only instruments with the highest reliability are acceptable. Measurement experts state that in such situations a reliability of .90 is the minimum that should be tolerated, and a reliability of .95 should be the desired standard.

Table 9.4 Factors Affecting Reliability of a Test

Factor	Potential Effect
1. Length of the test	The longer the test, the greater the reliability.
2. Heterogeneity of group	The more heterogeneous the group, the greater the reliability.
3. Ability level of group	A test that is too easy or too difficult for a group results in lower reliability.
4. Techniques used to estimate reliability	Test–retest and split-half give higher estimates. Equivalent forms give lower estimates.
5. Nature of the variable	Tests of variables that are easier to measure yield higher reliability estimates.
6. Objectivity of scoring	The more objective the scoring, the greater the reliability.

PICTURE THIS


He is fabulous on the court, but let's see how he does on our essay exam before we make him an offer.
11

Adding an irrelevant measure will lower reliability and validity.

Joe Rocco

THINK ABOUT IT 9.2

Indicate the type of reliability coefficient illustrated in each of the following exercises:

a. A teacher prepares two forms of a math achievement test, administers the two forms to a group of students on consecutive days, and correlates the students' scores from the two administrations.

b. A college professor administers a 40-item multiple-choice test in educational psychology. The scoring office provides the professor a reliability index found by dividing the test into two forms and calculating the correlation between the students' scores on the two.

c. A teacher questions the results of a verbal aptitude test administered to her English class. She decides to have the students take the same test on the following day. She then correlates the two sets of scores and finds a coefficient of .90.

d. A commercial test developed two forms of a standardized reading test and administered the two forms of the test to a representative sample of elementary school students in the fall and again in the spring.

e. A teacher wanted a reliability estimate of an essay test in history administered at the end of the semester. She used a computer program that calculated the variance of all the scores for each item and then plugged the total variances across all items into a formula.

Answers

a. Alternate forms (coefficient of equivalence)

b. Split-half reliability coefficient

c. Test–retest (coefficient of stability)
d. Alternate forms (coefficient of stability and equivalence)
e. Coefficient alpha

STANDARD ERROR OF MEASUREMENT

The reliability coefficient looks at the consistency of test scores for a group, but it does not tell us anything about the amount of error in individual test scores. Suppose you had an aptitude test score of 105 for an individual child. If we retested, we would probably not obtain that same score. How much variability could we expect in the child's score on retesting? Recall that measurement theory states that any obtained score is made up of the true score plus an error score: $X = T + E$. Because of error, the obtained score is sometimes higher than the true score and sometimes lower than the true score. Returning to the example of the aptitude test, you would expect with repeated administration to obtain a number of different scores for the same individual. In fact, you would have a frequency distribution of aptitude scores. The mean of this distribution of scores would be the best approximation of the child's true score, and the standard deviation would be an indicator of the errors of measurement. Because this standard deviation is the standard deviation of the errors of measurement, it is called the standard error of measurement. Test theory tells us that the distribution of error scores approximates a normal distribution, and we can use the normal distribution to represent it. Measurement errors are normally distributed with a mean of zero. There may be many small errors, but there will be few very large ones. The standard deviation of this distribution of errors (standard error of measurement, s_M) would give an estimate of how frequently errors of a given size might be expected to occur when the test is administered many times.

In practice, you usually do not have repeated measures for an individual but you can get an estimate of the standard error of measurement from one group administration of a test. The formula for standard error of measurement is

$$s_M = s_x\sqrt{1 - r_{xx}}$$

where

$$\begin{aligned} s_M &= \text{standard error of measurement} \\ s_x &= \text{standard deviation of test scores} \\ r_{xx} &= \text{reliability coefficient} \end{aligned}$$

Thus, using the standard deviation of the obtained scores and the reliability of the test, we can estimate the amount of error in individual scores. If the aptitude test has a reliability coefficient of .96 and a standard deviation of 15, then

$$s_M = 15\sqrt{1 - .96} = 15\sqrt{.04} = 3$$

What does the standard error of measurement tell us? It tells us something about how accurate an individual's score is on a test. We can use what we know about a normal distribution to make statements about the percentage of scores that fall between different points in a distribution. Given a student's obtained score, you use the s_M to determine the range of score values that will, with a given probability, include the individual's true score. This range of scores is referred to as a **confidence band.** Assuming that the errors of measurement are normally distributed about a given score and equally distributed throughout the score range, you could be 68 percent confident that a person's true score (the score if there were no errors of measurement) lies within one s_M on either side of the observed score. For example, if a subject has an observed score of 105 on an aptitude test where the standard error of measurement is 3, you could infer at the 68 percent confidence level that the subject's true score lies somewhere between 102 and 108. Or you can state at the 95 percent confidence level that the true score will fall within 1.96 (or rounded to 2) s_M of the obtained score (between 99 and 111). You can also use the standard error of measurement to determine how much variability could be expected on retesting the individual. If the subject could be retested on the same aptitude test a number of times, you could expect that in approximately two-thirds of the retests the scores would fall within a range of 6 points of the observed score, and in 95 percent of retests the scores would fall within a range of 12 points. Figure 9.3 shows (a) the distribution of error scores

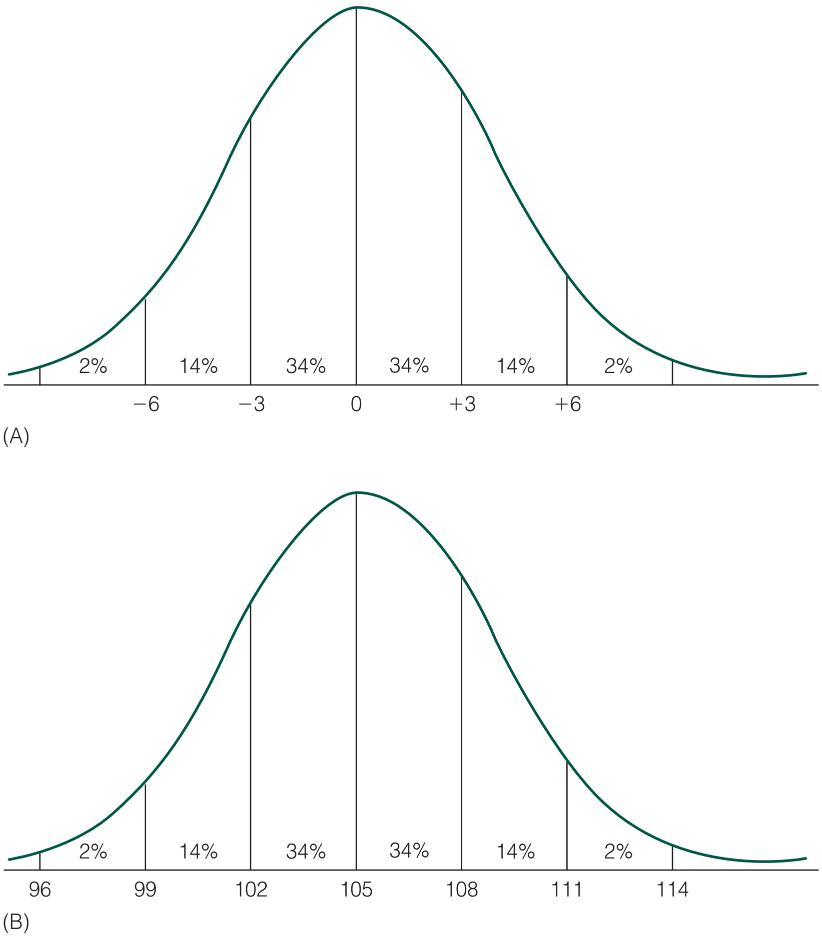


Figure 9.3 (A) The Distribution of Error Scores When $s_M = 3.00$ and (B) the Distribution around an Obtained Score of 105 with $s_M = 3.00$

(standard error of measurement of the test) and (b) the distribution of errors around an obtained score of 105 with $s_M = 3$.

The standard error of measurement (s_M) and the reliability coefficient (r_{xx}) are alternative ways of expressing how much confidence we can place in an observed score. The reliability coefficients provide an indicator of the consistency of a group of scores or items making up a test. The standard error of measurement provides an estimate of the consistency of an individual's performance on a test. How accurate or precise an estimate of the true score any observed score will provide is indicated by the size of these two indexes of reliability. As the reliability coefficient increases, the standard error of measurement decreases; as reliability decreases, the standard error of measurement increases. Look for a low standard error of measurement or a high reliability coefficient as an indicator of the stability of test scores. No one method of estimating reliability is optimal in all situations. The standard error of measurement is recommended for use when interpreting individual scores, and the reliability coefficient is recommended for use when comparing the consistency of different tests. You always want scores that are sufficiently consistent to justify anticipated uses and interpretations.

It is meaningless, however, to make a general statement that a test is "reliable." You must report the methods used to estimate the reliability index, the nature of the group from which data were derived, and the conditions under which data were obtained. Potential users of a test then must take responsibility for determining how the reliability data would apply to their population.

THINK ABOUT IT 9.3

a. A standardized test has a reported reliability coefficient of .84 and a standard deviation of 8. Calculate the standard error of measurement for this test.

b. Mary had a score of 100 on this test. Calculate the band within which Mary's true score is likely to fall. (Use the 95 percent confidence level.)

Answers

a. $s_M = s_x\sqrt{1 - r_{xx}}$ $\quad s_M = 8\sqrt{1 - .84} = 8\sqrt{.16} = 8(.4) = 3.2$

b. You can state at the 95 percent confidence level that Mary's true score is between 94 and 106 [$100 \pm (1.96)(3) \approx 100 \pm 6 = 94$ and 106].

RELIABILITY OF CRITERION-REFERENCED TESTS

The traditional methods used to determine the reliability of norm-referenced tests require sets of scores with considerable variability. Thus, these methods are not appropriate for criterion-referenced tests in which the scores are limited to 1, mastery, or 0, nonmastery. Several procedures have been suggested for estimating the reliability of criterion-referenced tests.

Agreement Coefficient (ρ)

The **agreement coefficient (ρ)** involves administering two equivalent forms of a criterion-referenced test, or the same test on two occasions, and determining the consistency of the decisions reached. The consistency is expressed as the percentage of people for whom the same decision (mastery or nonmastery) is made on both forms. This index of reliability is referred to as the agreement coefficient (ρ).

For example, the results displayed in Table 9.5 were obtained when two equivalent forms of a criterion-referenced test were administered to a sample of 100 students. In this case, 70 students were consistently classified as masters on both forms and 14 students were consistently classified as nonmasters.

The agreement coefficient (ρ) is the proportion of the total people consistently classified on the two forms, or

$$\rho = \frac{b+c}{N} \tag{9.9}$$

where

ρ = agreement coefficient
b = number classified as masters on both forms
c = number classified as nonmasters on both forms
N = total number of subjects

$$= \frac{70+14}{100} = \frac{84}{100} = .84$$

Thus, 84 percent of the subjects were classified consistently, and .84 is the agreement coefficient of this test. If classifications as master or nonmaster are consistent for all examinees on both administrations of the test, the agreement coefficient equals 1, the maximum value.

Some agreement in classifications as master or nonmaster between two forms is expected merely by chance; that is, even if classifications were made randomly, some individuals would be expected to fall in cells (b) and (c) in Table 9.5. Therefore, we suggest using a statistic proposed by Cohen (1960) that takes *chance agreement* into consideration.

Table 9.5 Decisions Based on Forms 1 and 2 of a Criterion-Referenced Test

		Form 1		
		Nonmaster	Master	
Form 2	Master	(a) 10	(b) 70	80
	Nonmaster	(c) 14	(d) 6	20
		24	76	100 (*N*)

b = number classified as masters on both forms
c = number classified as nonmasters on both forms
a = number classified as nonmasters on form 1 but masters on form 2
d = number classified as masters on form 1 but nonmasters on form 2
N = total number of students who have taken both form.

Kappa Coefficient

Cohen's **kappa coefficient**, κ, refers to the proportion of consistent classifications observed *beyond* that expected by chance alone. The rationale of the kappa coefficient is straightforward. First, calculate the percentage of cases expected to have consistent classification even if there were no genuine relationship between the forms—that is, if the classification on the two forms were completely independent. This index is referred to as the *expected chance agreement* (ρ_c). The expected chance agreement is subtracted from the observed agreement ($\rho_o - \rho_c$) to obtain the actual increase over chance consistency; this quantity is then divided by $1 - \rho_o - \rho_c$, the maximum possible increase in decision consistency beyond chance, to yield κ, the kappa coefficient.

Thus, the expected chance agreement is shown by the following formula:

$$\rho_c = \frac{(a + b)(a + c) + (c + d)(b + d)}{N^2} \tag{9.10}$$

where

$$\rho_c = \text{proportion of agreement expected by chance}$$

$$\kappa = \frac{\rho_o - \rho_c}{1 - \rho_c} \tag{9.11}$$

where

$$\begin{aligned} \kappa &= \text{proportion of agreement } \textit{above} \text{ that expected by chance} \\ \rho_o &= \text{observed agreement coefficient} \\ \rho_c &= \text{expected chance agreement} \end{aligned}$$

Using the data in the precedin(g example,

$$\begin{aligned} \rho_c &= \frac{(80)(24) + (20)(76)}{100^2} & \kappa &= \frac{.84 - .34}{1 - .34} \\ &= \frac{1920 + 1520}{10{,}000} & &= \frac{.50}{.66} \\ &= .34 & &= .76 \end{aligned}$$

You can see that the kappa coefficient (.76), which adjusts for expected chance agreement, provides a lower estimate of reliability than the agreement coefficient (.84). This is always the case, except when agreement is perfect ($\rho = 1.00$), because kappa begins with the observed agreement and then adjusts it for expected chance agreement. Because kappa is not inflated by chance agreements, it is considered a better indicator of reliability than the agreement coefficient. The agreement coefficient and kappa require two administrations of a test. There are techniques available for estimating the reliability of a criterion-referenced test from a single test administration, but we do not discuss them in this text.

Phi Coefficient

Another coefficient that is not inflated by chance agreement and thus yields results similar to kappa is phi (ϕ): The **phi coefficient,** a mathematical simplification of the Pearson *r* when all scores are either 1 or 0, is a useful measure of reliability for criterion-referenced measures.

$$\phi = \frac{bc - ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{9.12}$$

Using the data in Table 9.5,

$$\phi = \frac{(70)(14) - (10)(6)}{\sqrt{(10 + 70)(14 + 6)(10 + 14)(10 + 6)}}$$

$$= \frac{980 - 60}{\sqrt{(80)(20)(24)(76)}} = \frac{920}{\sqrt{2{,}918{,}400}}$$

$$= \frac{920}{1708.33} = .54$$

Note how close phi (.54) and kappa (.53) are. The phi coefficient is interpreted in the same way as the other forms of the Pearson *r*. It ranges from −1.00 (all disagreement) through 0 (no consistency) to +1.00 (all agreement). Note that phi (.54) is near the square of kappa $(.76^2 = .57)$.

RELIABILITY OF OBSERVATIONAL DATA

Reliability is also important in measuring instruments that require ratings or observations of individuals by other individuals. The researcher in these cases must determine the reliability of the ratings—whether different judges/observers have given similar scores or ratings to the same behaviors. A simple way to determine the reliability of ratings is to have two or more observers independently rate the same behaviors and then correlate the observers' ratings. The resulting correlation is called the **interrater** or **interobserver reliability.** If the behaviors to be observed are well defined and the observers well trained, the reliability of the observations should be positive and quite high (approximately .90).

Take the case of two individuals who have rated several students in a performance assessment in which the ratings range from 1 (very poor) to 10 (excellent). Here, reliability can be assessed through correlational procedures in the same way these procedures are used in test–retest or alternate-forms reliability. The second observer serves the same function as a retest or an alternate form in a paper-and-pencil test. When the scores are only 1 or 0 (behavior occurred versus behavior did not occur), the kappa (Formula 9.12) can be used to assess the reliability of the observers' scores. These procedures are also useful when training observers. Trainees watch and score a videotape that has been scored by an experienced observer, and the agreement coefficient, or kappa, indicates the correspondence between a trainee and the experienced observer. The trainer can go through the tape with the trainee to determine when and why the trainee misclassified observations.

The phi coefficient may also be used to assess the agreement of observers scoring 0 and 1. Suen and Ary (1989) provide an extensive discussion of reliability procedures in behavioral observations.

VALIDITY AND RELIABILITY COMPARED

Validity is a more important and comprehensive characteristic than reliability. Because it is more difficult to measure systematic error than random error, evaluating validity is more challenging. Validity is not obtained as directly as reliability. Assessing validity involves accumulating a great deal of evidence to support the proposed interpretations of scores. The conceptual framework indicates the kinds of evidence that you need to collect to support the meaning and interpretation of test scores. You must answer questions about the appropriateness of test content, the adequacy of criteria, the definitions of human traits, the specification of the behavioral domain, the theory behind the test content, and so forth. All these matters involve judgment and the gathering of data from many sources. You will find that published research studies typically report much more reliability data than validity data.

Reliability, in contrast, can be investigated directly from the test data; no data external to the measure are required. The basic issues of reliability lend themselves easily to mathematical analysis, and reasonable conclusions about the amount of error can be stated in mathematical terms. Figure 9.4 illustrates the difference between reliability and validity.

If a measure is to yield valid score-based interpretations, it must first be reliable. The reliability of an instrument determines the upper limit of its validity. Scores on a test with zero reliability are entirely random and therefore cannot correlate with any criterion. The possible correlation of an instrument with a criterion (validity coefficient) increases as the reliability of the instrument increases.

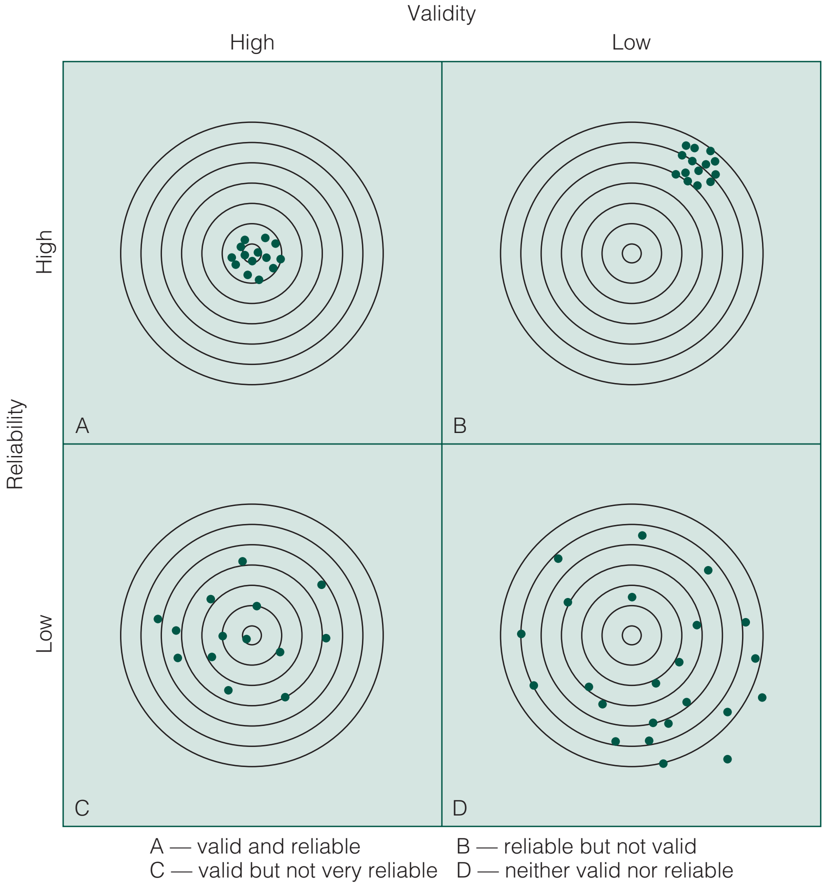

Validity
High
Low
Reliability
High
Low
A
B
C
D
A — valid and reliable
B — reliable but not valid
C — valid but not very reliable
D — neither valid nor reliable

Figure 9.4 Four Rifles Tested by Aiming at Bull's-Eye and Pulling Trigger

Source: From *Contemporary Advertising*, 8th edition, by William Arens. Copyright © 2002 by McGraw-Hill Companies, Inc. Reproduced by permission.

Table 9.6 Example of a Test Review from Mental Measurements Yearbook

TECHNICAL. The norming sample included $n = 1{,}898$ students from 23 states. For the most part, the sample appears to be representative of nationwide statistics as reported in U.S. Census information with regard to geographic region, gender, family income, educational level of parents, and exceptionality status. The test yields a single raw score. Standard scores, percentile ranks, age, and grade equivalents are provided on easy-to-interpret tables.

RELIABILITY. Evidence of reliability is provided using alternate forms (immediate and delayed), test–retest, and interrater scoring. The reliability coefficients for alternate forms (immediate administration) ranged from .82 to .89 by age level and from .76 to .96 for selected subgroups. The reliability coefficients for test–retest with a 2-week interval ranged from .82 to .95. Evidence of interrater reliability for scorers was high.

VALIDITY. In terms of content validity, the format of the test is more analogous to games like hidden puzzles than reading actual text. As such, it is not a strong reflection of what the authors intend. Some evidence of content validity might be inferred in that the test uses sentences adapted from two well-established reading tests.

For evidence of criterion-related validity, the TOSCRF was compared to archival scores on the Woodcock–Johnson III, the Gray Oral Reading Test (GORT-4), and the Stanford Achievement Test Series 9, and with the Test of Silent Word Reading Fluency (TOSWRF) administered at the time. Average uncorrected correlations across all forms of the TOSCRF ranged from .48 with the GORT-4 to .76 with the TOSWRF. The authors also compare standardized scores from the TOSCRF to a global measure generated from a combination of the other measures, via an independent samples *t*-test. The findings show that the means of the standard scores are similar. The authors interpret this as evidence of validity, but this support seems weak at best.

COMMENTARY. The TOSCRF is acceptable as a quick screening measure for students, as one part of a testing program. Its evidence of reliability and validity need more substantiation. Interpretation of scores for some areas of the U.S. and some subgroups not represented in the norming group should be made with caution.

Source: Geisinger, K., Spies, R., Carlson, J., & Plake, B. (Eds.) (2007) ***The seventeenth mental measurements yearbook*** (pp. 797–800). Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Remember, however, that a measure can have reliability without providing valid interpretations; it can consistently measure the wrong thing. Feldt and Brennan (1989) emphasize the primacy of validity in evaluating the adequacy of an educational measure by stating, “No body of reliability data, regardless of the elegance of the methods used to analyze it, is worth very much if the measure to which it applies is irrelevant or redundant” (p. 143).

Table 9.6 is an excerpt from the *Seventeenth Mental Measurements Yearbook* (Geisinger, Spies, Carlson, & Plake, 2007) showing the kind of validity and reliability data available on published tests. In this case, the instrument is the Test of Silent Contextual Reading Fluency (TOSCRF) designed as “a quick and accurate method of assessing silent reading ability” for individuals from ages 7 to 18 years.

SUMMARY

Choosing from the multiplicity of measuring instruments available to the researcher requires the use of criteria for the evaluation of these instruments. The two most important criteria for measuring devices are validity and reliability. Validity is the extent to which theory and evidence support the proposed interpretations of test scores for an intended purpose. In the process of assessing validity, the researcher gathers various types of supporting evidence from many sources. Three types of evidence are gathered: (1) content-related evidence, which assesses how well the instrument samples the content domain being measured; (2) criterion-related evidence, which assesses how well the instrument correlates with other measures of the variable of interest; and (3) construct-related evidence, which assesses how well the instrument represents the construct of interest.

The researcher must also ask, How consistently does the test measure whatever it does measure? This is the issue of reliability. No test can permit meaningful interpretations unless it measures consistently—that is, unless it is reliable. Reliability refers to the extent to which the test is consistent in measuring whatever it does measure. Specifically, reliability refers to the extent to which an individual scores nearly the same in repeated measurements, as indicated by a high reliability coefficient. Reliability coefficients can be computed in various ways, depending on the source of error being considered. The reliability coefficient shows the extent to which random errors of measurement influence scores on the test. The standard error of measurement, another index of reliability, enables researchers to employ the normal curve to estimate the limits within which a subject's true score can be expected to lie.

Validity and reliability procedures appropriate for criterion-referenced tests were discussed in this chapter. Procedures are also available for determining the reliability of observations.

KEY CONCEPTS

agreement coefficient
alternate-forms technique
coefficient (Cronbach) alpha
coefficient of equivalence
coefficient of reliability
coefficient of stability
coefficient of stability and equivalence
concurrent validity evidence
confidence band
construct-irrelevant variance
construct-related evidence of validity
construct underrepresentation
content-related validity evidence
convergent evidence of validity
criterion-related validity evidence
discriminant evidence of validity
divergent evidence
equivalent-forms technique
evidence based on internal structure
evidence based on response processes
evidence based on test content
face validity
factor analysis
homogeneity measures
internal-consistency procedures
interobserver reliability
interrater reliability
kappa coefficient
known-groups technique
Kuder–Richardson formulas
multitrait–multimethod matrix
observed score
parallel-forms technique
phi coefficient
predictive validity evidence
random errors of measurement
reliability
reliability coefficient
Spearman–Brown prophecy formula
split-half reliability coefficient
standard error of measurement
systematic errors of measurement
test–retest reliability coefficient
true score
validity
validity coefficient

EXERCISES

1. Compare *validity* and *reliability* with respect to the following:
   a. The meaning of each concept
   b. The relative importance of each concept
   c. The extent to which one depends on the other
2. Explain the following statement: A measuring device may be reliable without being valid, but it cannot be valid without being reliable.
3. How would you propose to gather evidence to support the use of a new scholastic aptitude test that had been developed for use with high school seniors?
4. You have been asked to assess the validity of an instrument designed to measure a student's academic self-concept (i.e., the way he or she views himself or herself as a student). How would you go about this task?

5. What source of evidence supporting the proposed interpretation of test scores is indicated in each of the following situations?

a. The high school language proficiency test scores of college dropouts and college persisters are compared in order to determine whether the test data correlated with the subjects' college status.

b. A new scholastic aptitude test is found to have a correlation of .93 with the SAT, which has been used to predict college success.

c. A new intelligence test has been developed. The author argues that the mental processes required by the test are congruent with the Z theory of intelligence. Furthermore, he shows that among children the average score on the test increases with each year of age.

d. A teacher carefully examines a standardized achievement test to determine if it covers the knowledge and skills that are emphasized in the class.

e. The mean difference between the rankings of members of the Ku Klux Klan and members of the Americans for Democratic Action on the liberalism scale was found to be highly significant.

f. A mathematics test is judged by a group of teachers to be an adequate and representative sample of the universe of test items.

g. Students are asked to verbalize how they solve mathematics problem-solving items.

6. Identify the type of procedure for estimating reliability that is illustrated in each of the following:

a. The same test was given twice to a certain group. The correlation between the scores on the two administrations of the test was .90.

b. The group's scores on the odd items of a test were correlated with their scores on the even items of the same test: $r_{xx} = .95$.

c. Alternate forms of the test were administered after 1 month, and results of the two administrations were correlated: $r_{xx} = .85$.

d. The variance, the mean, and the number of items are used to estimate reliability.

7. How would you account for the differences in the reliability coefficients in Exercise 6, assuming that the groups tested were the same?

8. How would you gather evidence for the validity of a reading readiness test?

9. What can you do to increase reliability when constructing a test?

10. Indicate the source of evidence that might be most relevant for assessing validity of the following types of tests:

a. A classroom history test

b. An instrument to measure achievement motivation

c. A measure designed to identify potential dropouts

d. A group intelligence test

e. A reading readiness test

11. Explain how a mathematics achievement test could be judged to have high validity in one mathematics class and low validity in another mathematics class.

12. Criticize the following statement: The reliability of the intelligence test is .90. Therefore, you can assume that the test scores can be interpreted as measuring intelligence.

13. Determine the standard error of measurement for a test with a standard deviation of 16 and a reliability coefficient of $r_{xx} = .84$. How would you interpret this standard error of measurement?

14. Select a standardized achievement test that you might use in a research study and obtain the necessary validity data on this test. (You may use *Mental Measurements Yearbook* and the manual that accompanies the test you select.)

15. Check the test manual for the achievement test being used in your school. What type of reliability data are reported there?

16. The following data were obtained when two forms of a criterion-referenced test in mathematics were given to a group of elementary school children. There were 50 items on each form. To pass, a student had to get 80 percent correct on each form. Express the reliability of this test in terms of the kappa coefficient (κ).

Examinee	Form 1	Form 2
1	45	47
2	43	48
3	45	31
4	39	39
5	39	48
6	34	37
7	46	46
8	48	49
9	43	38
10	36	46
11	45	48
12	38	39
13	44	45
14	31	34
15	42	48

17. Criticize the following procedures used to gather validity evidence:

a. A high school English teacher developed a writing test for identifying talented high school students and administered the test to her senior English classes. On the basis of high scores, students were permitted to enroll in an English class at the local university. At the end of the semester, the teacher correlated the original test scores with the grades the students earned in the college English class. The teacher was surprised to find a negligible correlation. What was the problem?

b. A school counselor developed a scale to measure need for academic achievement in elementary school children. The scale was administered to two classes of elementary school children, and the results were given to the teachers of these children. The teachers were asked to observe these children carefully for one semester, after which they were asked to rate the children on their need for achievement. The teachers' ratings were then correlated with the scores the children received on the scale. The correlation was quite high, so the counselor concluded that the scale had high validity for measuring need for achievement. Do you agree with the counselor's conclusion?

18. Assume that you wanted to investigate teacher "burnout." Suggest some indicators of this construct that you might use in developing a scale for this purpose.

19. What type of reliability estimate would be most appropriate for the following measuring instruments?

a. A multiple-choice achievement test will be used as the dependent variable in an experimental study.

b. A researcher will study changes in attitude and will administer one form of an attitude scale as both the premeasure and the postmeasure.

c. A researcher has two forms of an achievement test; she administered one form at the beginning of the study and the other at the conclusion of the study. She wants to determine the reliability of the test.

20. A 100-item test was split into two halves, and the split-half coefficient of correlation was found to be .60. Calculate the reliability coefficient for the full-length test.

	Judge 1	Judge 2
Kata	10	9
Ashok	8	7
Mary	7	10
Kwaku	9	8
Anil	6	5
Ester	4	3

21. Using a 10-point scale, two judges gave the following ratings to the essays written by a group of students. Calculate an index that indicates the reliability of this rating procedure.

22. Indicate whether each of the following practices would increase or decrease reliability?

a. The teacher decides to give a weekly quiz instead of one major test at the end of the grading period.

b. Jane Smith brags about her difficult tests where a large percentage of students fail.

c. On Friday afternoon, Miss Jones postponed the major exam until the following Monday after she heard about the football game scheduled for after school that day.

d. The teacher decided to add 10 easy test items that everyone could answer correctly.

e. The teacher wrote items having a wide range of difficulty, with most items answered correctly by 40 to 70 percent of students.

f. To save time, Ms. White had the students do only two of the subtests from a standardized test instead of taking the complete test.

g. The teacher decided to give 25 spelling words on the weekly test instead of 10.

23. The following are some comments often heard from students following exams. To what test characteristic are the comments most directly related?

a. The test measured minute details, not the important concepts emphasized in class.

b. The test was too long for the time available.

c. That material was not even covered in class.

d. The reading level was so complex that the test was really a measure of reading comprehension, not math.

e. Many students were observed to be cheating.

f. What does this test have to do with choosing students for the advanced chemistry class?

Test	Mean	Reliability	SD
A	50	.75	6
B	50	.91	15

24. You have the following technical information from two tests:
On which test would a student's score be expected to fluctuate the most on repeated administrations?

25. What are the sources of error that affect the reliability of a test? Give an example of each.

26. The following types of reliability coefficients were calculated for a test. Which coefficient do you think would be highest? Explain why.

a. Test–retest (1 month)

b. Parallel forms (1 week)

c. Split-half

ANSWERS

1. Validity is the extent to which an instrument measures what it is designed to measure. Reliability is the extent to which an instrument is consistent in measuring whatever it is measuring. Validity is considered a more important aspect than reliability because lack of validity implies lack of meaning. However, an instrument cannot be valid without first being reliable.

2. A measure may produce consistent scores (reliability) but may bear no relationship to other accepted measures of the construct or not be able to predict behavior associated with the construct (validity). Scores on a test with zero reliability are entirely random and therefore cannot correlate with any criterion. The extent of reliability sets an upper limit on possible validity.

3. You first must define what is meant by *aptitude*. If you wish to measure general academic ability, gather evidence about its content by examining the test items for representativeness. Do they assess the basic academic skills of reading, spelling, math, and so on? Gather evidence about the correlation between the test scores and senior year GPA, college freshman GPA, and other criteria. Correlation with other validated aptitude test scores could also be done.

4. The items of the scale or questionnaire would need to cover aspects of the student behavior that would logically be a part of the construct *academic self-concept* (e.g., I intend to go to college). Criterion measures could be personal interviews with students or independent assessment by teachers. Assuming academic self-concept is related to achievement, self-concept scores could be correlated with GPA and/or achievement test scores.

5. a. Evidence based on correlation with other variables

b. Evidence based on correlation with other variables

c. Evidence is construct related

d. Evidence based on content

e. Evidence based on known-groups technique

f. Evidence based on content

g. Evidence based on response processes (construct related)

6. a. Test–retest reliability
   b. Split-half reliability
   c. Alternate forms with time lapse reliability
   d. Internal consistency (Kuder–Richardson formula 21)

7. Split-half reliabilities tend to be higher than test–retest reliabilities because subject variability due to maturation, increase in testing skill, and other random factors is less. Equivalent-forms reliability is lower than same-test reliability because (a) it is impossible to construct exactly equivalent forms and (b) there is an added source of variability when nonidentical forms are used. The internal-consistency reliability will be depressed if the test is not homogeneous.

8. You would first identify which specific skills (e.g., letter recognition and left-to-right orientation) comprise reading readiness and then determine if the test incorporated these skills in appropriate proportions. When subjects who have taken the test have begun their reading programs, you would determine how scores on the test and on subtests correlate with reading test scores, teachers' ratings, and other criteria.

9. Rewriting ambiguous items, using items of appropriate difficulty, and clarifying instructions will increase reliability. Making a test longer by including additional items drawn from the same universe increases reliability, as does testing on a more heterogeneous group.

10. a. Evidence based on content
    b. Evidence based on internal structure of test, correlation with other criteria of achievement motivation, and performance of contrasted groups
    c. Evidence based on relationship with some criteria
    d. Evidence based on internal structure of the test and relationships with appropriate criteria
    e. Evidence based on internal structure of the test and relationship with appropriate criteria of reading achievement

11. A mathematics test that covered only computation would have little validity in a class that stressed concepts and reasoning. If content and emphasis of a different class match the content and emphasis of the test, the test will have high validity in that class.

12. A test can be reliable without measuring what it intends to measure. To determine validity, you need to look at content, constructs, and relations with other measures of the same construct as well as relations with measures of behavior assumed to be correlated with the construct.

13. By Formula 9.9, you interpret the standard error of measurement as a standard deviation. Thus, you can say that there are two chances in three that the individual's true score will fall in the range of 66.4 score points from the observed score.

$$\begin{aligned} s_M &= s_x\sqrt{1-r_{xx}} \\ &= 16\sqrt{1-.84} \\ &= 16\,(.4) \\ &= 6.4 \end{aligned}$$

14. Answers will vary.

15. Answers will vary.

16. A score of 40 represents mastery (50 × .80 = 40).

		Form 1		
		Nonmaster	Master	
Form 2	Master	(b) 2	(a) 7	9
	Nonmaster	(d) 4	(c) 2	6
		6	9	15

$$\rho_0 = \frac{7+4}{15} = \frac{11}{15} = .73$$

(73% of the students were classified consistently)

$$\rho_c = \frac{(9)(9)+(6)(6)}{15^2} = \frac{81+36}{225} = \frac{117}{225} = .52$$

$$\kappa = \frac{.73-.52}{1-.52} = \frac{.21}{.48} = .44$$

17. a. Selecting just high scorers restricts the variability. The restricted variability lowered the coefficient of correlation.

b. There was criterion contamination. Letting the teachers see the results of the original measurement of need for achievement contaminated their ratings of the children on need for achievement.

18. There are a number of possible indicators of teacher burnout. You could look at absenteeism from school, lower evaluations by supervisors, incidences of hostility toward students or supervisors, and incidences of challenging of school policies. You might also develop a scale to measure attitudes toward their work; from teachers' own responses to appropriate questions, you might infer the presence of burnout.

19. a. You would be interested in the internal consistency of this one form of the test. A split-half, alpha, or Kuder–Richardson reliability coefficient would be appropriate.

b. With one form to be used as both a pre- and postmeasure, you would compute a coefficient of stability.

c. With two forms and two administrations, you would compute the coefficient of stability and equivalence.

20. $r_{xx} = \dfrac{2(.60)}{1 + .60} = .75$

21. Pearson $r = .78$

22. a. Increase
b. Decrease
c. Increase
d. No effect
e. Increase
f. Decrease
g. Increase

23. a. Validity
b. Reliability
c. Validity
d. Validity
e. Reliability
f. Validity

24. Fluctuation would be greater on test B because the standard error of measurement is larger. (Calculate the standard error of measurement for each test.)

25. (a) The test itself (too short or ambiguous items); (b) administration of the test (poor directions and distractions during test); and (c) test taker (illness, fatigue, and lack of motivation).

26. Split-half because it measures only fluctuation from one-half of the test to another. There is no time lapse; thus, there is only one source of error.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological tests*. Washington, DC: Author.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20*(4), 6–18.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice, 22*(3), 5–11.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Doll, E. A. (1935, 1949, 1965). *Vineland social maturity scale*. Circle Pines, MN: American Guidance Service.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. Brennan (Ed.), *Educational measurement*. New York: American Council on Education and Measurement. [Reprinted 2006 by Greenwood Publishing]

Geisinger, K., Spies, R., Carlson, J., & Plake, B. (Eds.). (2007). *The Seventeenth Mental Measurements Yearbook*. Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, *2*, 151–160.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741–749.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. Upper Saddle River, NJ: Pearson.