CHAPTER 6

What Is Test Reliability/Precision?

LEARNING OBJECTIVES

After completing your study of this chapter, you should be able to do the following:

- Define reliability/precision, and describe three methods for estimating the reliability/precision of a psychological test and its scores.
- Describe how an observed test score is made up of the true score and random error, and describe the difference between random error and systematic error.
- Calculate and interpret a reliability coefficient, including adjusting a reliability coefficient obtained using the split-half method.
- Differentiate between the KR-20 and coefficient alpha formulas, and understand how they are used to estimate internal consistency.
- Calculate the standard error of measurement, and use it to construct a confidence interval around an observed score.
- Identify four sources of test error and six factors related to these sources of error that are particularly important to consider.
- Explain the premises of generalizability theory, and describe its contribution to estimating reliability.

“My statistics instructor let me take the midterm exam a second time because I was distracted by noise in the hallway. I scored 2 points higher the second time, but she says my true score probably didn’t change. What does that mean?”

“I don’t understand that test. It included the same questions—only in different words—over and over.”

“The county hired a woman firefighter even though she scored lower than someone else on the qualifying test. A man scored highest with a 78, and this woman only scored 77! Doesn’t that mean they hired a less qualified candidate?”

“The psychology department surveyed my class on our career plans. When they reported the results of the survey, they also said our answers were unreliable. What does that mean?”

Copyright ©2016 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.
Have you ever wondered just how consistent or precise psychological test scores are? If a student retakes a test, such as the SAT, can the student expect to do better the second time without extra preparation? Are the scores of some tests more consistent than others? How do we know which tests are likely to produce more consistent scores?

If you have found yourself making statements or asking questions like these, or if you have ever wondered about the consistency of a psychological test or survey, the questions you raised concern the reliability of responses. As you will learn in this chapter, we use the term reliability/precision to describe the consistency of test scores. All test scores—just like any other measurement—contain some error. It is this error that affects the reliability, or consistency, of test scores.

In the past, we referred to the consistency of test scores simply as reliability. Because the term reliability is used in two different ways in the testing literature, the authors of the newly published Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) have revised the terminology, and we follow the revised terminology in this book. When we are referring to the consistency of test scores in general, the term reliability/precision is preferred. When we are referring to the results of the statistical evaluation of reliability, the term reliability coefficient is preferred. In the past, we used the single term reliability to indicate both concepts.

In this chapter, we describe the three categories of reliability coefficients and the methods we use to estimate them. We first consider test–retest coefficients, which estimate the reliability/precision of test scores when the same people take the same test form on two separate occasions. Next, we cover alternate-form coefficients, which estimate the reliability/precision of test scores when the same people take a different but equivalent form of a test on two independent testing sessions. Finally, we discuss internal consistency coefficients, which estimate the reliability/precision of test scores by looking at the relationships between different parts of the same test given on a single occasion. This category of coefficients also enables us to evaluate scorer reliability or agreement when raters use their subjective judgment to assign scores to test taker responses.

In the first part of this chapter, we define what each of these categories and methods is in more detail. We then show you how each of the three categories of reliability coefficients is calculated. We also discuss how to calculate an index of error called the standard error of measurement (SEM), and a measure of rater agreement called Cohen's kappa. Finally, we discuss factors that increase and decrease a test’s reliability/precision.

What Is Reliability/Precision?

As you are aware, psychological tests are measurement instruments. In this sense, they are no different from yardsticks, speedometers, or thermometers. A psychological test measures how much the test taker has of whatever skill or quality the test measures. For instance, a driving test measures how well the test taker drives a car, and a self-esteem test measures whether the test taker’s self-esteem is high, low, or average when compared with the self-esteem of similar others.

The most important attribute of a measurement instrument is its reliability/precision. A yardstick, for example, is a reliable measuring instrument over time because each time it measures an object (e.g., a room), it gives approximately the same answer. Variations in the measurements of the room—perhaps a fraction of an inch from time to time—can be referred to as measurement error. Such errors are probably due to random mistakes or inconsistencies of the person using the yardstick or because the smallest
increment on a yardstick is often a quarter of an inch, making finer distinctions difficult. A yardstick also has internal consistency. The first foot on the yardstick is the same length as the second foot and third foot, and the length of every inch is uniform.

Reliability/precision is one of the most important standards for determining how trustworthy data derived from a psychological test are. A **reliable test** is one we can trust to measure each person in approximately the same way every time it is used. A test also must be reliable if it is used to measure attributes and compare people, much as a yardstick is used to measure and compare rooms. Although a yardstick can help you understand the concept of reliability, you should keep in mind that a psychological test does not measure physical objects as a yardstick does, and therefore a psychological test cannot be expected to be as reliable as a yardstick in making a measurement.

Keep in mind that just because a test has been shown to produce reliable scores, that does not mean the test is also **valid**. In other words, evidence of reliability/precision does not mean that the inferences that are being made from the scores on the test are correct or that it is being used properly. (We explain the concept of validity in Chapters 7, 8, and 9.)

### Three Categories of Reliability Coefficients

If you measured a room but you were unsure whether your measurement was correct, what would you do? Most people would measure the room a second time using either the same or a different tape measure.

Psychologists use the same strategies of remeasurement to check psychological measurements. These strategies establish evidence of the reliability/precision of test scores. Some of the methods that we will discuss require two administrations of the same (or very similar) test forms, while other methods can be accomplished in a single administration of the test. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) recognize three categories of reliability coefficients used to evaluate the reliability/precision of test scores. Each category uses a different method for estimating the reliability/precision of a test. The methods are (a) the test–retest method, (b) the alternate-forms method, and (c) the internal consistency method (split-half, coefficient alpha methods, and methods that evaluate scorer reliability or agreement). Each of these methods takes into account various conditions that can produce inconsistencies in test scores. Not all methods are used for all tests. The method chosen to determine reliability/precision depends on the test itself and the conditions under which the test user plans to administer the test. Each method produces a numerical reliability coefficient, which enables us to evaluate the reliability/precision of the test.

### Test–Retest Method

To estimate how reliable a test is using the **test–retest method**, a test developer gives the same test to the same group of test takers on two different occasions. The scores from the first and second administrations
are then compared using correlation—a statistical technique that was discussed in Chapter 5. This method of estimating reliability allows us to examine the stability of test scores over time and provides an estimate of the test’s reliability/precision.

The interval between the two administrations of the test may vary from a few hours up to several years. As the interval lengthens, test–retest reliability will decline because the number of opportunities for the test takers or the testing situation to change increases over time. For example, if we give a math achievement test to a student today and then again tomorrow, there probably is little chance that the student’s knowledge of math will change overnight. However, if we give a student a math achievement test today and then again in 2 months, it is very likely that something will happen during the 2 months that will increase (or decrease) the student’s knowledge of math. When test developers or researchers report test–retest reliability, they must also state the length of time that elapsed between the two test administrations.

Using test–retest reliability, the assumption is that the test takers have not changed between the first administration and the second administration in terms of the skill or quality measured by the test. On the other hand, changes in test takers’ moods, levels of fatigue, or personal problems from one administration to another can affect their test scores. The circumstances under which the test is administered, such as the test instructions, lighting, or distractions, must be alike. Any differences in administration or in the individuals themselves will introduce error and reduce reliability/precision.

The test developer makes the first estimates of the reliability/precision of a test’s scores. A good example of estimating reliability/precision using the test–retest method can be seen in the initial reliability testing of the Personality Assessment Inventory (PAI). The PAI, developed by Leslie Morey, is used for clinical diagnoses, treatment planning, and screening for clinical psychopathology in adults. To initially determine the PAI’s test–retest reliability coefficient, researchers administered it to two samples of individuals not in clinical treatment. (Although the test was designed for use in a clinical setting, using a clinical sample for estimating reliability would have been difficult because changes due to a disorder or to treatment would have confused interpretation of the results of the reliability studies.) The researchers administered the PAI twice to 75 normal adults. The second administration followed the first by an average of 24 days. The researchers also administered the PAI to 80 normal college students, who took the test twice with an interval of 28 days. In each case, the researchers correlated the set of scores from the first administration with the set of scores from the second administration. The two studies yielded similar results, showing acceptable estimates of test–retest reliability for the PAI. (Later in this chapter, we discuss what an “acceptable level of reliability” means.)

An important limitation in using the test–retest method of estimating reliability is that the test takers may score differently (usually higher) on the test because of practice effects. Practice effects occur when test takers benefit from taking the test the first time (practice), which enables them to solve problems more quickly and correctly the second time. (If all test takers benefited the same amount from practice, it would not affect reliability; however, it is likely that some will benefit from practice more than others will.) Therefore, the test–retest method is appropriate only when test takers are not likely to learn something the first time they take the test that can affect their scores on the second administration or when the interval between the two administrations is long enough to prevent practice effects. In other words, a long time between administrations can cause test takers to forget what they learned during the first administration. However, short intervals between testing implementations may be preferable when the test measures an attribute that may change in an individual over time due to learning or maturation, or when the possibility that changes in the testing environment that occur over time may affect the scores.

More detail about the PAI can be found in Test Spotlight 6.1 in Appendix A.
Alternate-Forms Method

To overcome problems such as practice effects, psychologists often give two forms of the same test—designed to be as much alike as possible—to the same people. This strategy requires the test developer to create two forms of the test that are referred to as alternate forms or parallel forms. Again, the sets of scores from the two tests are compared using correlation. This method of estimating reliability/precision provides a test of equivalence. The two forms (Form A and Form B) are administered as close in time as possible—usually on the same day. To guard against any order effects—changes in test scores resulting from the order in which the tests were taken—half of the test takers may receive Form A first and the other half may receive Form B first.

An example of the use of alternate forms in testing can be seen in the development of the Test of Nonverbal Intelligence, Fourth Edition (TONI-4; PRO-ED, n.d.). The TONI-4 is the fourth version of an intelligence test that was designed to assess cognitive ability in populations that have language difficulties due to learning disabilities, speech problems, or other verbal problems that might result from a neurological deficit or developmental disability. The test does not require any language to be used in the administration of the test or in the responses of the test takers. The items are carefully drawn graphics that represent problems with four to six possible solutions. The test takers can use any mode of responding that the test administrator can understand to indicate their answers, such as nodding, blinking, or pointing. Because this test is often used in situations in which there is a need to assess whether improvement in functioning has occurred, two forms of the test needed to be developed—one to use as a pretest and another to use as a posttest. After the forms were developed, the test developers assessed the alternate-forms reliability by giving the two forms to the same group of subjects in the same testing session. The results demonstrated that the correlation between the test forms (which is the reliability coefficient) across all ages was .81, and the mean score difference between the two forms was one half of a score point. This is good evidence for alternate-forms reliability of the TONI-4.

The greatest danger when using alternate forms is that the two forms will not be truly equivalent. Alternate forms are much easier to develop for well-defined characteristics, such as mathematical ability, than for personality traits, such as extroversion. For example, achievement tests given to students at the beginning and end of the school year are alternate forms. Although we check the reliability of alternate forms by administering them at the same time, their practical advantage is that they can also be used as pre- and posttests if desired.

Internal Consistency Method

What if you can give the test only once? How can you estimate the reliability/precision? As you recall, test–retest reliability provides a measure of the test’s reliability/precision over time, and that measure can be taken only with two administrations. However, we can measure another type of reliability/precision, called internal consistency, by giving the test once to one group of people. Internal consistency is a measure of how related the items (or groups of items) on the test are to one another. Another way to think about this is whether knowledge of how a person answered one item on the test would give you information that would help you correctly predict how he or she answered another item on the test. If you can (statistically) do that across the entire test, then the items must have something in common with each other. That commonality is usually related to the fact that they are measuring a similar attribute, and therefore we say that the test is internally consistent. Table 6.1 shows two pairs of math questions. The first pair has more commonality for assessing ability to do math calculations than the second pair does.

More detail about the TONI-4 can be found in Test Spotlight 6.2 in Appendix A.
Can you see why this is so? The problems in Pair A are very similar; both involve adding single-digit numbers. The problems in Pair B, however, test different arithmetic operations (addition and multiplication), and Pair A uses simpler numbers than Pair B does. In Pair A, test takers who can add single digits are likely to get both problems correct. However, test takers who can add single digits might not be able to multiply three-digit numbers. The problems in Pair B measure different kinds of math calculation skills, and therefore they are less internally consistent than the problems in Pair A, which both measure the addition of single-digit numbers. Another way to look at the issue is that if you knew that a person correctly answered Question 1 in Pair A, you would have a good chance of being correct if you predicted that the person also would answer Question 2 correctly. However, you probably would be less confident about your prediction about a person answering Question 1 in Pair B correctly also answering Question 2 correctly.

Table 6.1 Internally Consistent Versus Inconsistent Test Questions

| A. Questions with higher internal consistency for measuring math calculation skill: |
| Question 1: 7 + 8 = ? | Question 2: 8 + 3 = ? |
| B. Questions with lower internal consistency for measuring math calculation skill: |
| Question 1: 4 + 5 = ? | Question 2: 150 × 300 = ? |

Statisticians have developed several methods for measuring the internal consistency of a test. One traditional method, the split-half method, is to divide the test into halves and then compare the set of individual test scores on the first half with the set of individual test scores on the second half. The two halves must be equivalent in length and content for this method to yield an accurate estimate of reliability. The best way to divide the test is to use random assignment to place each question in one half or the other. Random assignment is likely to balance errors in the score that can result from order effects (the order in which the questions are answered), difficulty, and content.

When we use the split-half method to calculate a reliability coefficient, we are in effect correlating the scores on two shorter versions of the test. Shortening a test decreases its reliability. Therefore, when using the split-half method, we must mathematically adjust the reliability coefficient to compensate for the impact of splitting the test into halves. We will discuss this adjustment—using an equation called the Spearman–Brown formula—later in the chapter.

An even better way to measure internal consistency is to compare individuals’ scores on all possible ways of splitting the test into halves. This method compensates for any error introduced by any unintentional lack of equivalence that splitting a test in the two halves might create. Kuder and Richardson (1937, 1939) first proposed a formula, KR-20, for calculating internal consistency of tests whose questions can be scored as either right or wrong (such as multiple-choice test items). Cronbach (1951) proposed a formula called coefficient alpha that calculates internal consistency for questions that have more than two possible responses such as rating scales. We also discuss these formulas later in this chapter.

Estimating reliability using methods of internal consistency is appropriate only for homogeneous tests—measuring only one trait or characteristic. With heterogeneous tests—measuring more than one trait or characteristic—estimates of internal consistency are likely to be lower. For example, a test for people who are applying for the job of accountant may measure knowledge of accounting principles,
calculation skills, and ability to use a computer spreadsheet. Such a test is heterogeneous because it measures three distinct factors of performance for an accountant.

It is not appropriate to calculate an overall estimate of internal consistency (e.g., coefficient alpha, split-half) when a test is heterogeneous. Instead, the test developer should calculate and report an estimate of internal consistency for each homogeneous subtest or factor. The test for accountants should have three estimates of internal consistency: one for the subtest that measures knowledge of accounting principles, one for the subtest that measures calculation skills, and one for the subtest that measures ability to use a computer spreadsheet. In addition, Neal Schmitt (1996) stated that the test developer should report the relationships or correlations between the subtests or factors of a test.

Furthermore, Schmitt (1996) emphasized that the concepts of internal consistency and homogeneity are not the same. Coefficient alpha describes the extent to which questions on a test or subscale are interrelated. Homogeneity refers to whether the questions measure the same trait or dimension. It is possible for a test to contain questions that are highly interrelated, even though the questions measure two different dimensions. This can happen when there is some third “common” factor that may be related to all the other attributes that the test measures. For instance, we described a hypothetical test for accountants that contained subtests for accounting skills, calculation skills, and use of a spreadsheet. Even though these three subtests may be considered heterogeneous dimensions, all of them may be influenced by a common dimension that might be named general mathematical ability. Therefore, people who are high in this ability might do better across all three subtests than people lower in this ability. As a result, coefficient alpha might still be high even though the test measures more than one dimension. Therefore, a high coefficient alpha is not “proof” that a test measures only one skill, trait, or dimension.

The developers of the PAI also conducted studies to determine its internal consistency. Because the PAI requires test takers to provide ratings on a response scale that has five options (false, not at all true, slightly true, mainly true, and very true), they used the coefficient alpha formula. The developers administered the PAI to three samples: a sample of 1,000 persons drawn to match the U.S. census, another sample of 1,051 college students, and a clinical sample of 1,246 persons. The coefficient alpha formula ideally requires large numbers of respondents to estimate internal consistency because violation of assumptions underlying the formula causes coefficient alpha to overestimate or underestimate the population reliability/precision when the number of respondents is small (Zimmerman, Zumbo, & Lalonde, 1993). However, coefficient alpha is still routinely calculated even when the number of respondents is much lower than was used in the PAI example.

Table 6.2 shows the estimates of internal consistency for the scales and subscales of the PAI. Again, the studies yielded levels of reliability/precision considered to be acceptable by the test developer for most of the scales and subscales of the PAI. Two scales on the test—Inconsistency and Infrequency—yielded low estimates of internal consistency. However, the test developer anticipated lower alpha values because these scales measure the care used by the test taker in completing the test, and careless responding could vary during the testing period. For instance, a test taker might complete the first half of the test accurately but then become tired and complete the second half haphazardly.

**Scorer Reliability**

What about errors made by the person who scores the test? An individual can make mistakes in scoring, which add error to test scores, particularly when the scorer must make judgments about whether an answer is right or wrong. When scoring requires making judgments, two or more persons should score the test. We then compare the judgments that the scorers make about each answer to see how much they
agree. The methods we have already discussed pertain to whether the test itself yields consistent scores, but scorer reliability and agreement pertain to how consistent the judgments of the scorers are.

Some tests, such as those that require the scorer to make judgments, have complicated scoring schemes for which test manuals provide the explicit instructions necessary for making these scoring judgments. Deviation from the scoring instructions or a variation in the interpretation of the instructions introduces error into the final score. Therefore, scorer reliability or interscorer agreement—the
amount of consistency among scorers’ judgments—becomes an important consideration for tests that require decisions by the administrator or scorer.

A good example of estimating reliability/precision using scorer reliability can be seen in the Wisconsin Card Sorting Test (WCST). This test was originally designed to assess perseveration and abstract thinking, but it is currently one of the most widely used tests by clinicians and neurologists to assess executive function (cognitive abilities that control and regulate abilities and behaviors) of children and adults. Axelrod, Goldman, and Woodard (1992) conducted two studies on the reliability/precision of scoring the WCST using adult psychiatric inpatients. In these studies, one person administered the test and others scored the test. In the first study, three clinicians experienced in neuropsychological assessment scored the WCST data independently according to instructions given in an early edition of the test manual (Heaton, 1981). Their agreement was measured using a statistical procedure called intraclass correlation, a special type of correlation appropriate for comparing responses of more than two raters or of more than two sets of scores. The scores that each clinician gave each individual on three subscales correlated at .93, .92, and .88—correlations that indicated very high agreement. The studies also looked at intrascorer reliability—whether each clinician was consistent in the way he or she assigned scores from test to test. Again, all correlations were greater than .90.

In the second study, six novice scorers, who did not have previous experience scoring the WCST, scored 30 tests. The researchers divided the scorers into two groups. One group received only the scoring procedures in the test manual (Heaton, 1981), and the other group received supplemental scoring instructions as well as those in the manual. All scorers scored the WCST independently. The consistency level of these novices was high and was similar to the results of the first study. Although there were no significant differences between groups, those receiving the supplemental scoring material were able to score the WCST in a shorter time period. Conducting studies of scorer reliability for a test, such as those of Axelrod and colleagues (1992), ensures that the instructions for scoring are clear and unambiguous so that multiple scorers arrive at the same results.

We have discussed three methods for estimating the reliability/precision of a test: test–retest, alternate forms, and internal consistency, which included scorer reliability. Some methods require only a single administration of the test, while others require two. Again, each of these methods takes into account various conditions that could produce differences in test scores, and not all strategies are appropriate for all tests. The strategy chosen to determine an estimate of reliability/precision depends on the test itself and the conditions under which the test user plans to administer the test.

Some tests have undergone extensive reliability/precision testing. An example of such a test is the Bayley Scales of Infant and Toddler Development, a popular and interesting test for children that has extensive evidence of reliability. According to Dunst (1998), the standardization and the evidence of reliability/precision and validity of this test far exceed generally accepted guidelines. (We discuss such guidelines later in this chapter. Of interest here is the challenge the Bayley Scales faced in gathering evidence of test–retest reliability.)

The test developer should report the reliability method as well as the number and characteristics of the test takers in the reliability study along with the associated reliability coefficients. For some tests, such as the PAI, the WCST, and the Bayley Scales, more than one method may be appropriate. Each method provides evidence that the test is consistent under certain circumstances. Using more than one method provides strong corroborative evidence that the test is reliable.
The next section describes statistical methods for calculating reliability coefficients, which estimate the reliability/precision of a test. As you will see, the answer to how reliable a test's scores are may depend on how you decide to measure it. Test–retest, alternate forms, and internal consistency are concerned with the test itself. Scorer reliability involves an examination of how consistently the person or persons scored the test. That is why test publishers may need to report multiple reliability coefficients for a test to give the test user a complete picture of the instrument.

### INTERIM SUMMARY 6.1

**THREE TYPES OF RELIABILITY/PRECISION**

- The test–retest method compares the scores of the same test takers taking the same test at two different times. This method is appropriate when the test takers have not changed on the ability, trait, or attitude that the test measures.
- The alternate-forms method compares the scores of the same test takers on two equivalent forms.
- The internal consistency method compares the scores of test takers on two halves of the test taken at the same time (split-half method). Scores of test takers on all possible halves can be calculated using the coefficient alpha or KR-20 formula when the test questions are homogeneous.
- Scorer reliability is a method for estimating reliability/precision or accuracy of those making judgments when scoring a test. Two types of scorer reliability are important: the comparison of judgments among scorers (interscorer) and the consistency of judgments that each scorer makes for all tests (intrascorer). We use correlation to calculate the interscorer reliability for interval data such as numerical test scores.

### Classical Test Theory

No measurement instrument is perfectly reliable or consistent. Many clocks are slow or fast—even if we measure their errors in nanoseconds. Unfortunately, psychologists are not able to measure psychological qualities with the same precision that engineers have for measuring speed or physicists have for measuring distance.

For instance, did you ever stop to think about the fact that when you give a test to a group of people, their scores will vary, that is, they will not all obtain the same score? The obvious reason for this is that the people to whom you give the test differ in the amount of the attribute the test measures, and the variation in test scores simply reflects this fact. Now think about the situation in which you retest the same people the next day using the same test. Would each individual score exactly the same on the second testing as on the first? The answer is that they most likely wouldn’t. The scores would probably be close to the scores they obtained on the first testing, but they would not be exactly the same. Some people would score higher on the second testing, while some people would score lower. But assuming that the amount of the attribute that the test measures has stayed the same in each person (after all, it’s only 1 day later), why should the obtained test scores have changed? It turns out that according to classical test theory, any test score that a person obtains can be considered to be made up of two independent parts. The first part is a measure of the amount of the attribute that the test measures. This is known as the person’s true score ($T$). The second part of an obtained test score consists of random errors that occur anytime a person takes a test ($E$). It is this random error that causes a person’s test score to change from
one administration of a test to the next (assuming that his or her true score hasn’t changed). Because this type of error is a random event, sometimes it causes an individual’s test score to go up on the second administration, and sometimes it causes it to go down. So if you could know what a person’s true score was on a test, and also know the amount of random error, you could easily determine what the person’s actual observed score on the test would be. Likewise, error in measurement can be defined as the difference between a person’s observed score and his/her true score. Formally, classical test theory expresses these ideas by saying that any observed test score ($X$) is made up of two parts: a true score ($T$) and random error ($E$). Therefore,

$$X = T + E.$$ 

**True Score**

An individual’s true score ($T$) on a test is a value that can never really be known or determined. It represents the score that would be obtained if that individual took a test an infinite number of times and then the average score across all the testings were computed. As we discuss in a moment, random errors that may occur in any one test occasion will actually cancel themselves out over an infinite number of testing occasions. Therefore, if we could average all the scores together, the result would represent a score that no longer contained any random error. This is the true score on the test and represents the amount of the attribute the person who took the test actually possesses without any random measurement error.

One way to think about a true score is to think about choosing a member of your competitive video gaming team. You could choose a person based on watching him or her play a single game. But you would probably recognize that that single score could have been influenced by a lot of factors (random error) other than the person’s actual skill playing video games (the true score). Perhaps the person was just plain lucky in that game, and the observed score was really higher than his or her actual skill level would suggest. So perhaps you might prefer that the person play three games so that you could take the average score to estimate his or her true level of video gaming ability. Intuitively, you may understand by asking the person to play multiple games that some random influences on performance might even out because sometimes these random effects will cause the observed score to be higher than the true score, and sometimes it will cause the observed score to be lower. This is the nature of random error. So you can probably see that if somehow you could get the person to play an infinite number of games and average all the scores, the random error would cancel itself out entirely and you would be left with a score that represents the person’s true score in video gaming.

**Random Error**

Random error is defined as the difference between a person’s actual score on a test (the obtained score) and that person’s true score ($T$). As we described above, because this source of error is random in nature, sometimes a person’s obtained score will be higher than his or her true score and sometimes the obtained score will be lower than his or her true score. Unfortunately, in any single test administration, we can never know whether random error has led to an obtained score that is higher or lower than the true score. An important characteristic of this type of measurement error is that, because it is random, over an infinite number of testings the error will increase and decrease a person’s score by exactly the same amount. Another way of saying this is that the mean or average of all the error scores over an infinite number of testings will be zero. That is why random error actually cancels itself out over repeated
testings. Another important characteristic of measurement error is that it is normally distributed (see Chapter 5 for a discussion of normal distributions). Clearly, we can never administer a test an infinite number of times in an attempt to fully cancel out the random error component. The good news is that we don’t have to. It turns out that making a test longer also reduces the influence of random error on the test score for the same reason—the random error component will be more likely to cancel itself out (although never completely in practice). We will have more to say about this when we discuss how reliability coefficients are actually computed.

**Systematic Error**

Systematic error is another type of error that obscures the true score. When a single source of error always increases or decreases the true score by the same amount, we call it *systematic error*. For instance, if you know that the scale in your bathroom regularly adds 3 pounds to anyone’s weight, you can simply subtract 3 pounds from whatever the scale says to get your true weight. In this case, the error your scale makes is predictable and systematic. The last section of this chapter discusses how test developers and researchers can identify and reduce systematic error in test scores.

Let us look at an example of the difference between random error and systematic error proposed by Nunnally (1978). If a chemist uses a thermometer that always reads 2 degrees warmer than the actual temperature, the error that results is *systematic*, and the chemist can predict the error and take it into account. If, however, the chemist is nearsighted and reads the thermometer with a different amount and direction of inaccuracy each time, the readings will be wrong and the inconsistencies will be unpredictable, or *random*.

Systematic error is often difficult to identify. However, two problems we discussed earlier in this chapter—practice effects and order effects—can add systematic error as well as random error to test scores. For instance, if test takers learn the answer to a question in the first test administration (practice effect) or can derive the answer from a previous question (order effect), more people will get the question right. Such occurrences raise test scores systematically. In such cases, the test developer can eliminate the systematic error by removing the question or replacing it with another question that will be unaffected by practice or order.

Another important distinction between random error and systematic error is that random error lowers the reliability of a test. Systematic error does not; the test is reliably inaccurate by the same amount each time. This concept will become apparent when we begin calculating reliability/precision using correlation.

**The Formal Relationship Between Reliability/Precision and Random Measurement Error**

Building on the previous discussion of true score and error, we can provide a more formal definition of reliability/precision. Recall that according to classical test theory, any score that a person makes on a test (his or her observed score, \(X\)) is composed of two components, his or her true score, \(T\), and random measurement error, \(E\). This is expressed by the formula \(X = T + E\). Now for the purposes of this discussion, let’s assume that we could build two different forms of a test that measured the exactly the same construct in exactly the same way. We would say that these alternate forms of the test were strictly parallel. As we discussed earlier, if we gave these two forms of the test to the same group of people, we would still not expect that everyone would score exactly the same on the second administration of the...
test as they did on the first. This is because there will always be some measurement error that influences everyone's scores in a random, nonpredictable fashion. Of course, if the tests were really measuring the same concepts in the same way, we would expect people's scores to be very similar across the two testing sessions. And, the more similar the scores are, the better the reliability/precision of the test would be.

Now for a moment, let's imagine an alien world where there was no measurement error (either random or systematic). With no measurement error, we would expect that everyone's scores on the two parallel tests would be exactly the same. In effect, both sets of test scores would simply be a measure of each individual's “true score” on the construct the test was measuring. If this were the case, the correlation between the two sets of test scores (i.e., the reliability coefficient) would be a perfect 1.0, and we would say that the test is perfectly reliable. It would also be the case that if the two groups of test scores were exactly the same for all individuals, the variance of the scores of each test (see Chapter 5 for the formula for calculating variance) would be exactly the same as well. This also makes intuitive sense. If two tests really were measuring the same concept in the same way and there were no measurement error, then nobody's score would “vary” or change from the first test to the second. So, the variance of both groups of scores would likewise be identical. In other words, we would be measuring only the true scores, which would not change across administrations of two strictly parallel tests.

Now let's move back to our real world, where there is measurement error. Random measurement error affects each individual's score in an unpredictable and different fashion. Sometimes the overall measurement error will cause an individual's test score to go up, sometimes it will go down, and sometimes it may remain unchanged. But, you can never predict the impact that the error will have on an individual's test score, and it will be different for each person as well. That is the nature of random error. However, there is one thing that you can predict. The presence of random error will always cause the variance of a set of scores to increase over what it was if there were no measurement error. A simple example will help make this point clear.

Suppose your professor administered a test and everyone in the class scored between a 79 and 81. The variance of this group of scores would be quite small because there was such a narrow range of scores. Now let's presume that your professor was unhappy about that outcome and wanted to change it so that the range of scores (the variance) was a little larger. So, he decided to add or subtract some points to everyone's score. In order that he not be accused of any favoritism in doing this, he generates a list of random numbers that range between –2 and +2. Starting at the top of his class roster and at the top of the list of random numbers, he adjusts each student's test score by the number that is in the same position on random number list as the student in on the class roster. Each student would now have had a random number of points added or subtracted from his or her score, and the test scores would now vary between 77 and 83 instead between of 79 and 81. You can immediately see that if you calculated the variance on this adjusted group of test scores, it would be higher than the original group of scores. But now, your professor has obscured the true scores on the test by adding “random error” into everyone's test score. As a result, the reliability/precision of the test scores would be reduced because the scores on the test now would contain random error. This makes the observed scores different from the students’ “true scores” by an amount equal to random error added to the score. If the students complained and were given the option to take the test a second time and no “random error” was added to the results on the second testing, the scores on the two testing occasions might be similar, but they would not be the same. The presence of the random error which the professor “added” to the first test has distorted the results. The presence of the random error that was added would reduce the estimate of reliability/precision of the test.
More formally, reliability is about estimating the proportion of variability in a set of observed test scores that is attributable to true scores. In classical test theory, reliability is defined as true-score variance divided by total observed-score variance:

$$\frac{\sigma_t^2}{\sigma_x^2},$$

where

$$\sigma_t^2 = \text{true-score variance}$$
$$\sigma_x^2 = \text{observed-score variance}$$

In words, this means that reliability reflects the proportion of the total observed variance in the test scores that is attributable to the variance of true scores. Recall that according to classical test theory, observed-score ($X$) variance is itself composed of two parts. Part of the variance in the observed scores will be attributable to the variance in the true scores ($T$), and part will be attributable to the variance added by measurement error ($E$). Therefore, if observed-score variance $\sigma_x^2$ were equal to true-score variance $\sigma_t^2$, this would mean that there is no measurement error, and so, using the above formula, the reliability coefficient in this case would be 1.00. But any time observed-score variance is greater than true-score variance (which is always the case because of the presence of measurement error), the reliability coefficient will be less than 1.

We wanted to make this idea more concrete for you, so we simulated what would occur if a professor actually changed a set test scores the way we described above. We started by first creating a group of 100 test scores that varied between 79 and 81. For the purposes of this example, we will pretend that we have moved back to the alien world that has no measurement error, so we will consider these simulated scores to be the usually unknowable true scores. We then calculated the variance of these "true" scores and found it to be 1.00. (It was a simply a coincidence that the variance of these simulated true scores turned out to be 1.00. True-score variance can be any positive number.) We will call this variance our true-score variance, $\sigma_t^2$. We then generated 100 normally distributed random numbers that ranged from –2 to +2 and had a mean of 0. These random numbers represent our random measurement error. (Remember, random error as defined by classical test theory is considered to be normally distributed with a mean of 0.) To create our simulated “observed scores” we added one of these random numbers to each of the simulated true scores, just like our hypothetical professor did. This created a new distribution in which the scores now ranged from 77 to 83. Because the addition of the simulated random error made the scores vary more than they did before, the calculated variance would also increase. When we calculated the variance of these simulated observed scores, we found that the variance did indeed increase, to 2.00. We can call this our observed-score variance, $\sigma_x^2$. So based on our formal definition of reliability/precision—true-score variance divided by observed-score variance ($\sigma_t^2/\sigma_x^2$)—we would calculate the reliability coefficient of these simulated test scores to be 1.00/2.00, or .5. If we had not added simulated random error to the scores in our simulation, the true-score variance and observed-score variance would have been the same, so the reliability coefficient would have been 1.00. Finally, we calculated the reliability coefficient in the usual manner, by computing the correlation coefficient between the two groups of simulated scores. It was also exactly .5, as expected. From this simulation, you should be able to see exactly how the presence of random measurement error acts to decrease the reliability coefficient of test scores by increasing the variance of the observed scores while the true-score variance remains unchanged.
Earlier we told you that you can never really determine a person’s true score on any measure. Remember, the definition of a true score is the score that a person would get if he or she took a test an infinite number of times and we averaged the all the results, which is something we can never actually do. And because we cannot ever know what a person’s true score actually is, we can never exactly calculate a reliability coefficient as we did with our simulated example. The best that we can do is to estimate it using the methods we have described in this chapter. That is why throughout this chapter, we have always spoken about reliability coefficients as being “estimates” of reliability/precision. In the next section, we show you how estimates of reliability/precision and related statistics are actually computed.

The Reliability Coefficient

We need to be as accurate as possible in describing our estimates of the reliability/precision of test scores. Therefore, we use correlation (discussed in Chapter 5) to provide an index of the strength of the relationship between two sets of test scores. To calculate the reliability coefficient using the test–retest method, we correlate the scores from the first and second test administrations; in the case of the alternate-forms and split-half methods, we correlate the scores of the first test and the second test.

The symbol that stands for a correlation coefficient is \( r \). To show that the correlation coefficient represents a reliability coefficient, we add two subscripts of the same letter, such as \( r_{xx} \) or \( r_{aa} \). Often authors omit the subscripts in the narrative texts of journal articles and textbooks when the text is clear that the discussion involves reliability, and we follow that convention in this chapter. Remember that a reliability coefficient is simply a Pearson product–moment correlation coefficient applied to test scores.

Adjusting Split-Half Reliability Estimates

As we mentioned earlier, the number of questions on a test is directly related to reliability; the more questions on the test, the higher the reliability, provided that the test questions are equivalent in content and difficulty. This is because the influence of random measurement error due to the particular choice of questions used to represent the concept is reduced when a test is made longer. Other sources of measurement error can still exist, such as inconsistency in test administration procedures or poorly worded test instructions. When a test is divided into halves and then the two halves of the test are correlated to estimate its internal consistency, the test length is reduced by half. Therefore, researchers adjust the reliability coefficient (obtained when scores on each half are correlated) using the formula developed by Spearman and Brown. This formula is sometimes referred to as the prophecy formula because it designed to estimate what the reliability coefficient would be if the tests had not been cut in half, but instead were the original length. We typically use this formula when adjusting reliability coefficients derived by correlating two halves of one test. Other reliability coefficients, such as test–retest and coefficient alpha, should not be adjusted in this fashion. For Your Information Box 6.1 provides the formula developed by Spearman and Brown and shows how to calculate an adjusted reliability coefficient.

The Spearman–Brown formula is also helpful to test developers who wish to estimate how the reliability/precision of a test would change if the test were made either longer or shorter. As we have said, the length of the test influences the reliability of the test; the more homogeneous questions (questions about the same issue or trait) the respondent answers, the more information the test yields about the concept the test is designed to measure. This increase yields more distinctive information about each
SECTION II: PSYCHOMETRIC PRINCIPLES

Using the Spearman–Brown Formula

The Spearman–Brown formula below represents the relationship between reliability and test length. It is used to estimate the change in reliability/precision that could be expected when the length of a test is changed. It is often used to adjust the correlation coefficient obtained when using the split-half method for estimating the reliability coefficient, but it is also used by test developers to estimate how the reliability/precision of a test would change if a test were made longer or shorter for any reason.

\[ r_{xx} = \frac{nr}{1 + (n - 1)(r)} \]

where

- \( r_{xx} \) = estimated reliability coefficient of the longer or shorter version of the test
- \( n \) = number of questions in the revised (often longer) version divided by the number of questions in the original (shorter) version of the test
- \( r \) = calculated correlation coefficient between the two short forms of the test

Suppose that you calculated a split-half correlation coefficient of .80 for a test split randomly in half. You are interested in knowing what the estimated reliability coefficient of the full-length version of the test would be. Because the whole test is twice as long as each split half, \( n \) is 2. You can then follow these steps to adjust the coefficient obtained and estimate the reliability of the test.

**Step 1:** Substitute values of \( r \) and \( n \) into the equation:

\[ r_{xx} = \frac{2(.80)}{1 + (2 - 1)(.80)} \]

**Step 2:** Complete the algebraic calculations:

\[ r_{xx} = .89. \]

Our best estimate of the reliability coefficient of the full-length test is .89.

Other Methods of Calculating Internal Consistency

As you recall, a more precise way to measure internal consistency is to compare individuals’ scores on all possible ways of splitting the test in halves (instead of just one random split of test items into two halves). This method compensates for error introduced by any lack of equivalence in the two halves. The two formulas researchers use for estimating internal consistency are KR-20 and coefficient alpha.
Researchers use the KR-20 formula (Kuder & Richardson, 1937, 1939) for tests whose questions, such as true/false and multiple choice, can be scored as either right or wrong. (Note that although multiple-choice questions have a number of possible answers, only one answer is correct.) Researchers use the coefficient alpha formula (Cronbach, 1951) for test questions, such as ratings scales, that have more than one correct answer. Coefficient alpha may also be used for scales made up of questions with only one right answer because the formula will yield the same result as does the KR-20.

How do most researchers and test developers estimate internal consistency? Charter (2003) examined the descriptive statistics for 937 reliability coefficients for various types of tests. He found an increase over time in the use of coefficient alpha and an associated decrease in the use of the split-half method for estimating internal consistency. This change is probably due to the availability of computer software that can calculate coefficient alpha. Charter also reported that the median reliability coefficient in his study was .85. Half of the coefficients examined were above what experts recommend, and half were below what experts recommend. For Your Information Box 6.2 provides the formulas for calculating KR-20 and coefficient alpha.

**FYI**

**FOR YOUR INFORMATION BOX 6.2**

**Formulas for KR-20 and Coefficient Alpha**

Two formulas for estimating internal reliability are KR-20 and coefficient alpha. KR-20 is used for scales that have questions that are scored either right or wrong, such as true/false and multiple-choice questions. The formula for coefficient alpha is an expansion of the KR-20 formula and is used when test questions have a range of possible answers, such as a rating scale. Coefficient alpha may also be used for scales made up of questions with only one right answer:

\[
\hat{r}_{KR20} = \left( \frac{k}{k-1} \right) \left( 1 - \frac{\sum pq}{\sigma^2} \right),
\]

where

- \( r_{KR20} \) = KR-20 reliability coefficient
- \( k \) = number of questions on the test
- \( p \) = proportion of test takers who gave the correct answer to the question
- \( q \) = proportion of test takers who gave an incorrect answer to the question
- \( \sigma^2 \) = variance of all the test scores

The formula for coefficient alpha (\( \alpha \) is the Greek symbol for alpha) is similar to the KR-20 formula and is used when test takers have a number of answers from which to choose their response:

\[
\hat{r}_\alpha = \left( \frac{k}{k-1} \right) \left( 1 - \frac{\sum \sigma_i^2}{\sigma^2} \right),
\]

where

- \( r_\alpha \) = coefficient alpha estimate of reliability
- \( k \) = number of questions on the test
- \( \sigma_i^2 \) = variance of the scores on one question
- \( \sigma^2 \) = variance of all the test scores
Calculating Scorer Reliability/Precision and Agreement

We can calculate scorer reliability/precision by correlating the judgments of one scorer with the judgments of another scorer. When there is a strong positive relationship between scorers, scorer reliability will be high.

When scorers make judgments that result in nominal or ordinal data, such as ratings and yes/no decisions, we calculate interrater agreement—an index of how consistently the scorers rate or make decisions. One popular index of agreement is Cohen's kappa (Cohen, 1960). For Your Information Box 6.3 describes kappa and demonstrates how to calculate it.

When one scorer makes judgments, the researcher also wants assurance that the scorer makes consistent judgments across all tests. For example, when a teacher scores essay exams, we would like the teacher to judge the final essays graded in the same way that he or she judged the first essays. We refer to this concept as intrarater agreement. (Note that inter refers to “between,” and intra refers to “within.”) Calculating intrarater agreement requires that the same rater rate the same thing on two or more occasions. In the example mentioned above, a measure of intrarater agreement could be computed if a teacher graded the same set of essays on two different occasions. This would provide information on how consistent (i.e., reliable) the teacher was in his or her grading. One statistical technique that is used to evaluate intrarater reliability is called the intraclass correlation coefficient, the discussion of which goes beyond the scope of this text. Shrout and Fleiss, (1979) provided an in-depth discussion of this topic. Table 6.3 provides an overview of the types of reliability we have discussed and the appropriate formula to use for each type.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Administration</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test–retest reliability</td>
<td>Administer the same test to the same people at two points in time.</td>
<td>Pearson product–moment correlation</td>
</tr>
<tr>
<td>Alternate forms or</td>
<td>Administer two forms of the test to the same people.</td>
<td>Pearson product–moment correlation</td>
</tr>
<tr>
<td>parallel forms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internal consistency</td>
<td>Give the test in one administration, and then split the test into two halves for scoring.</td>
<td>Pearson product–moment correlation corrected for length by the Spearman–Brown formula</td>
</tr>
<tr>
<td>Internal consistency</td>
<td>Give the test in one administration, and then compare all possible split halves.</td>
<td>Coefficient alpha or KR-20</td>
</tr>
<tr>
<td>Interrater reliability</td>
<td>Give the test once, and have it scored (interval- or ratio-level data) by two scorers or two methods.</td>
<td>Pearson product–moment correlation</td>
</tr>
<tr>
<td>Interrater agreement</td>
<td>Create a rating instrument, and have it completed by two judges (nominal- or ordinal-level data).</td>
<td>Cohen's kappa</td>
</tr>
<tr>
<td>Intrarater agreement</td>
<td>Calculate the consistency of scores for a single scorer A single scorer rates or scores the same thing on more than one occasion.</td>
<td>Intraclass correlation coefficient</td>
</tr>
</tbody>
</table>
Using Computer Software to Calculate Reliability

When you begin developing or using tests, you will not want to calculate reliability by hand. All statistical software programs and many spreadsheet programs will calculate the Pearson product-moment correlation coefficient. You simply enter the test scores for the first and second administrations (or halves) and choose the correlation menu command. If you calculate the correlation coefficient to estimate split-half reliability, you will probably need to adjust the correlation coefficient by hand using the Spearman–Brown formula, because most software programs do not make this correction.

Computing coefficient alpha and KR-20 are more complicated. Spreadsheet software programs usually do not calculate coefficient alpha and KR-20, but the formulas are available in the larger, better known statistical packages such as SAS and SPSS. Consult your software manual for instructions on how to enter your data and calculate internal consistency. Likewise, some statistical software programs calculate Cohen's kappa; however, you may prefer to use the matrix method demonstrated in For Your Information Box 6.3.

Interpreting Reliability Coefficients

As you recall from Chapter 5, we look at a correlation coefficient in two ways to interpret its meaning. First, we are interested in its sign—whether it is positive or negative. The sign tells us whether the two variables increase or decrease together (positive sign) or whether one variable increases as the other decreases (negative sign).

Second, we look at the number itself. As you also recall, correlation coefficients range from –1.00 (a perfect negative correlation) to +1.00 (a perfect positive correlation). Most often, the coefficient’s number will fall in between. Therefore, if a test’s reliability coefficient is +.91, we know that its sign is positive; people who made high scores on the first administration made similarly high scores on the second, and people who made low scores on the first administration made similarly low scores on the second. Furthermore, the coefficient .91 is very close to +1.00 or perfect agreement, so the test appears to be very reliable. Likewise, a correlation can be negative. A correlation coefficient of −.91 would also be very reliable, but the interpretation would be different. In a negative correlation, those who scored high on one test would score low on the second test, and those who scored low on the first test would consistently score high on the second test. While correlations can range from −1.00 to +1.00, reliability coefficients are considered to range from 0.00 to 1.00. Tests can range from not at all reliable ($r_{xx} = 0.00$) to perfectly reliable ($r_{xx} = 1.00$). To better understand the amount of error in a test score, we use the reliability coefficient to calculate another statistic called the standard error of measurement.

Calculating the Standard Error of Measurement

Psychologists use the standard error of measurement (SEM) as an index of the amount of inconsistency or error expected in an individual’s observed test score. In other words, the SEM is an estimate of how much the individual’s observed test score ($X$) might differ from the individual’s true test score ($T$). As you recall, the true test score is the theoretical score that a person would obtain if there were no measurement errors. For Your Information Box 6.4 shows how to calculate the SEM.
Cohen’s Kappa

Cohen’s kappa provides a nonparametric index for scorer agreement when the scores are nominal or ordinal data (Cohen, 1960). For example, pass/fail essay questions and rating scales on personality inventories provide categorical data that cannot be correlated. Kappa compensates and corrects interobserver agreement for the proportion of agreement that might occur by chance. Cohen developed the following formula for kappa (κ):

\[ \kappa = \frac{p_o - p_c}{1 - p_c}, \]

where

- \( p_o \) = observed proportion
- \( p_c \) = expected proportion

An easier way to understand the formula is to state it using frequencies (\( f \)):

\[ \kappa = \frac{f_o - f_e}{N - f_e}, \]

where

- \( f_o \) = observed frequency
- \( f_e \) = expected frequency
- \( N \) = overall total of data points in the frequency matrix

Many researchers calculate Cohen’s kappa by arranging the data in a matrix in which the first rater’s judgments are arranged vertically and the second rater’s judgments are arranged horizontally. For example, assume that two scorers rate nine writing samples on a scale of 1 to 3, where 1 indicates very poor writing skills, 2 indicates average writing skills, and 3 indicates excellent writing skills. The scores that each rater provided are shown below:

Scorer 1: 3, 3, 2, 2, 3, 1, 2, 3, 1
Scorer 2: 3, 2, 3, 2, 3, 2, 2, 3, 1

As you can see, Scorers 1 and 2 agreed on the first writing sample, did not agree on the second sample, did not agree on the third sample, and so on. We arrange the scores in a matrix by placing a check mark in the cell that agrees with the match for each writing sample. For example, the check for the first writing sample goes in the bottom right cell, where excellent for Scorer 1 intersects with excellent for Scorer 2, the check for the second writing sample goes in the middle right cell where excellent for Scorer 1 intersects with average for Scorer 2, and so on.
### Chapter 6: What Is Test Reliability/Precision?

To calculate kappa, each cell in the matrix must contain at least one agreement. Unfortunately, our $N$ of 9 is too small. As you can see, our nine writing samples do not fill all of the cells in the matrix. The following is another matrix containing data for 36 writing samples:

<table>
<thead>
<tr>
<th></th>
<th>Poor (1)</th>
<th>Average (2)</th>
<th>Excellent (3)</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scorer 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor (1)</td>
<td>9</td>
<td>3</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>Average (2)</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>Excellent (3)</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Column totals</td>
<td>15</td>
<td>12</td>
<td>9</td>
<td>36</td>
</tr>
</tbody>
</table>

In this matrix for 36 writing samples, Scorers 1 and 2 agreed a total of 23 times (the sum of the diagonal cells). The sum of the row totals ($S$ rows) is 36, and the sum of the column totals ($S$ columns) is 36, in agreement with the overall total of 36.

To calculate the expected frequency ($f_c$) for each diagonal, we use the following formula:

$$f_c = \frac{\text{Row Total} \times \text{Column Total}}{\text{Overall Total}},$$

where

- $f_c$ for the first cell in the diagonal = $(13 \times 15)/36 = 5.42$
- $f_c$ for the second cell in the diagonal = $4.67$
- $f_c$ for the third cell in the diagonal = $2.25$

Now we can calculate the sum of the expected frequencies of the diagonals ($\Sigma f_c$):

$$\Sigma f_c = 5.42 + 4.67 + 2.25 = 12.34.$$

When we plug the sum of the expected frequencies of the diagonals into the frequencies formula for kappa, we can calculate the value of kappa:

$$\kappa = \frac{\Sigma f_o - \Sigma f_c}{N - \Sigma f_c} = \frac{23 - 12.34}{36 - 12.34} = .45.$$

In this example, kappa ($\kappa$) equals .45.

Kappa ranges from $-1.00$ to $+1.00$. The higher the value of kappa, the stronger the agreement among the judges or raters. The scorers of the 36 writing samples are in moderate agreement. They should discuss how they are making their judgments so that they can increase their level of agreement.
Calculating the Standard Error of Measurement

The formula for calculating the standard error of measurement is

\[ \text{SEM} = \sigma \sqrt{1 - r_{xx}} \]

where

- \( \text{SEM} \) = standard error of measurement
- \( s \) = standard deviation of one administration of the test scores
- \( r_{xx} \) = reliability coefficient of the test

For this example, we will use the data in Table 6.4, which provides data on two administrations of the same test for 10 test takers. The calculated reliability coefficient \( r_{xx} \) for this test is .91. The standard deviation \( (\sigma) \) for the first administration of the test is 14.327.

With \( \sigma = 14.327 \) and \( r_{xx} = .91 \), you can calculate the SEM by substituting these values into the equation and completing the algebraic calculations as follows:

\[ \text{SEM} = 14.327 \sqrt{1 - .91} \]
\[ \text{SEM} = 4.2981 \text{ or } 4.3 \]

The SEM can be used to construct a confidence interval around a test score to provide a better estimate of the range in which the test taker’s true score is likely to fall. This process is demonstrated in For Your Information Box 6.5.

Interpreting the Standard Error of Measurement

To understand what the SEM means, we must apply it to an individual’s test score. As you now know, if an individual took a particular test two times, the scores on the first and second administrations of the test would likely be different because of random errors in measurement. If the person took the test 10 times, we would probably observe 10 similar but not identical scores. Remember, we are assuming the person’s true score has not changed across the administrations, but rather the observed differences in scores are due to random measurement error. The important point to understand is that a person’s observed score on a test is really only an estimate of his or her true score on the construct that the test was designed to measure.

Also recall that random error is assumed to be normally distributed. What this means is that each time a person takes a test, the amount of influence that measurement error will have on that person’s observed score can vary. Sometimes measurement error can create a large difference between a person’s observed and true scores; sometimes the difference will be small. It depends on the magnitude of the measurement error present in the test. And because random error is normally distributed (if graphed it would look like a normal curve), its influence on the observed score will vary from one test
administration to another. The SEM enables us to quantify the amount of variation in a person's observed score that measurement error would most likely cause.

Because of the characteristics of the normal distribution, we can assume that if the individual took the test an infinite number of times, the following would result:

- Approximately 68% of the observed test scores ($X$) would be within ±1 SEM of the true score ($T$).
- Approximately 95% of the observed test scores ($X$) would be within ±2 SEM of the true score ($T$).
- Approximately 99.7% of the observed test scores ($X$) would be within ±3 SEM of the true score ($T$).

(To understand this assumption, refer to our discussion of the properties of the normal curve in Chapter 5.)

**Confidence Intervals**

We can then use the preceding information to construct a confidence interval—a range of scores that we feel confident will include the test taker's true score. For Your Information Box 6.5 shows how to calculate a confidence interval for an observed score.

Confidence intervals are important because they give us a realistic estimate of how much error is likely to exist in an individual's observed score, that is, how big the difference between the individual's observed score and his or her (unobservable) true score is likely to be. The wider the confidence interval, the more measurement error is present in the test score.

Understanding confidence intervals is important any time we make decisions based on people's test scores, such as whether to hire them or admit them to a special educational program or whether they may be at risk for a particular medical disorder. The presence of error in the test scores could cause the decision to be incorrect. The more confident we are that the observed score on a test is really close to the person's true score, the more comfortable we can be that we are making a correct decision about the meaning of the score.

For Your Information Box 6.5 shows you how to calculate a confidence interval that is likely to contain an individual's true score using the data presented in Table 6.4. That table presents the observed test scores for 10 candidates on two administrations.

<table>
<thead>
<tr>
<th>Test Taker</th>
<th>First Administration</th>
<th>Second Administration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams</td>
<td>90</td>
<td>95</td>
</tr>
<tr>
<td>Butler</td>
<td>70</td>
<td>75</td>
</tr>
<tr>
<td>Chavez</td>
<td>50</td>
<td>65</td>
</tr>
<tr>
<td>Davis</td>
<td>100</td>
<td>95</td>
</tr>
<tr>
<td>Ellis</td>
<td>90</td>
<td>80</td>
</tr>
<tr>
<td>Franks</td>
<td>70</td>
<td>75</td>
</tr>
<tr>
<td>Garrison</td>
<td>60</td>
<td>65</td>
</tr>
<tr>
<td>Hart</td>
<td>75</td>
<td>80</td>
</tr>
<tr>
<td>Isaacs</td>
<td>75</td>
<td>80</td>
</tr>
<tr>
<td>Jones</td>
<td>85</td>
<td>80</td>
</tr>
</tbody>
</table>
scores for 10 people who took the same test on two occasions. The reliability coefficient of the test is .91, the standard deviation (σ) for the first administration of the test is 14.327, and the SEM is 4.3 points. When we calculate the 95% confidence interval for the true scores on the first administration of the test, it is ±8.4 points of the observed score. This means that we are 95% sure that a person who has an observed score of 75 on the test actually has a true score between 66.6 and 83.4, a relatively wide interval of about 17 points. Let’s see what the implications of this are in practice. If we calculated the 95% confidence interval for a person who had an observed score of 70 on the test, we see that we can be 95% confident that the person’s true score is between 61.6 and 78.4. Can you see the potential problem this creates? Assume that we had set the passing score on the test at 73. Without knowledge of the SEM and confidence interval, we would conclude that the person who scored 75 passed, and the person who scored 70 did not. But based on the 95% confidence interval, the true score of the person with the score of 75 could be as low as 66.6, and the true score of the person who scored 70 could be as high as 78.4. So it is possible that the person with the observed score of 70 might have really passed (based on his or her true score), while the person with the observed score of 75 might have actually failed. Unfortunately, as we have stated before, there is no way to know the precise true score. So, when making a judgment about

---

**FOR YOUR INFORMATION BOX 6.5**

**Calculating a 95% Confidence Interval Around an Observed Test Score**

The formula for calculating a 95% confidence interval around a score is

\[
95\% \text{ CI} = X \pm 1.96(\text{SEM}),
\]

where

- \(95\% \text{ CI}\) is the 95% confidence interval
- \(X\) is an individual’s observed test score
- \(\pm 1.96\) are the 2 points on the normal curve that include 95% of the scores
- \(\text{SEM}\) is the standard error of measurement for the test

For this example, we will use the data in Table 6.4 for the first administration of a test. The calculated SEM is 4.3 (see For Your Information Box 6.4). If we wanted to calculate the 95% confidence interval for an observed score of 90 on that first administration, the calculation is performed as follows:

\[
95\% \text{ CI} = X \pm 1.96(\text{SEM})
\]

\[
95\% \text{ CI} = 90 - (1.96 \times 4.3) \text{ and } 90 + (1.96 \times 4.3)
\]

\[
= (90 - 8.428) \text{ and } (90 + 8.428)
\]

\[
= 81.572 \text{ and } 98.428
\]

\[
95\% \text{ CI} = 81.572 \text{ to } 98.428.
\]

Therefore, we would say that there is a 95% chance that this confidence interval will contain the true test score \((T)\), which falls between 81.572 and 98.428.
the meaning of two different observed scores, it is important to evaluate the confidence intervals to see whether they overlap like they do in this case. When the true-score confidence intervals for two different observed scores overlap, it means that you cannot be sure that the observed scores’ differences reflect equivalent differences in true scores. In that case, the two observed scores should be treated as if they are the same score. While our example used a 95% confidence interval, it is not uncommon to use the 68% confidence interval when dealing with test scores. Using the 68% confidence interval will produce a narrower band of test scores than the 95% confidence interval does, but statistically we will be less certain that the person’s true score falls within the interval.

An applied example of how the estimation of measurement error is used occurs when a political poll suggests that one candidate will win an election by 2%, but the stated margin of error in the poll is 3%. In this case, the race would be considered to be a statistical tie despite the fact that the poll showed that one candidate was ahead by 2%, because the estimated 2% difference is smaller than the margin of error.

One of the issues that is usually not mentioned in textbooks on psychological testing when confidence intervals around true scores are discussed is that the calculated confidence interval is almost always centered on an observed score—not a true score. We also follow that practice in this book. As you now know, any observed score is only an estimate of a true score that will be more or less precise depending on the amount of measurement error present in the test. Some authors, such as Nunnally and Bernstein (1994), have suggested that the observed score around which the confidence interval is to be constructed should be statistically adjusted to account for measurement error before the confidence interval is calculated. By doing so, the confidence interval for the true scores will be a more precise estimate because it will be centered on an estimated true score, not the original observed score. However, other authors, such as Harvill (1991), have indicated that centering the confidence interval on an unadjusted observed score will provide a satisfactory estimate so long as the reliability/precision of the test is reasonably high and the observed score is not an extreme score relative to the mean score on the test.

Finally, it is important to mention that the standard error of measurement as we have presented it here is an average across all the observed scores on a test. But it can be shown that the SEM may not be exactly the same at all score levels on a test. Raw (untransformed) scores near the mean of the score distribution tend to have a larger SEM than very high or very low scores, but scaled scores that have been transformed from the raw scores for easier interpretation (see Chapter 5) can sometimes show the opposite pattern (Brennan & Lee, 1999). This becomes a very important consideration when test scores are used to make any kind of selection or placement decision. As you have learned, when confidence intervals around the true scores overlap, you may not be sure that differences in observed test scores actually correspond to differences in true scores. In those cases, you might have to consider the two different observed scores equivalent for decision-making purposes. You also have learned that the width of the confidence interval is dependent upon the SEM. So if the SEM differs at different observed scores, the confidence interval around the true scores will also differ. If you are using a predetermined passing or cut score for selection or classification of individuals, it is important to calculate the SEM at the passing score when possible, as it might be different than the SEM averaged across all the scores. An SEM calculated at a specific score is known as a conditional standard error of measurement, because its value is conditioned upon, or calculated at, a particular observed score. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) suggest that where possible, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error of measurement is constant across a wide range of scores. The calculation of a conditional standard error of measurement requires some advanced statistical techniques that we will not be able consider here.

As a general statement, remember that when the reliability/precision of the test scores is high, the SEM is low. This is because high reliability/precision implies low random measurement error. As that
reliability/precision decreases, random measurement error increases and the SEM increases. Although high reliability/precision is always important, it is especially so when test users use test scores to distinguish among individuals. For instance, when hiring, the answer to whether one candidate really had a lower test score than another can be found by using the SEM to calculate a 95% confidence interval around each candidate’s score. Often there will be a substantial overlap of confidence intervals for observed scores that are close to each other, suggesting that although there is a difference in observed scores, there might not be a difference in true scores of candidates.

Next we discuss how the reliability estimate—and thus the reliability of the test scores—may be increased or decreased.

**INTERIM SUMMARY 6.2**

**CALCULATING AND INTERPRETING THE RELIABILITY COEFFICIENT**

- A reliability coefficient ($r_{xx}$) is calculated by correlating the scores of test takers on two administrations of a test.
- A reliability coefficient is interpreted by examining its sign (positive or negative) and its proximity to 1.00. Reliability coefficients should be positive and very close to 1.00.
- The standard error of measurement provides a measure of how much an individual’s score is likely to differ from his or her true score.
- The standard error of measurement may not be the same at all levels of observed scores.
- A conditional standard error is one that is calculated at a specific value of observed score and is important to calculate when the test scores are used for selection or placement.
- Using the standard error of measurement, we can calculate a confidence interval that is likely to contain the true score.
- Cohen’s kappa is a measure of agreement for use with nominal or ordinal data such as ratings and decisions. The closer kappa is to 1.00, the stronger the agreement.

**Factors That Influence Reliability**

Because reliability is so important to accurate measurement, we need to consider several factors that may increase or decrease the reliability of the test scores. Error that can increase or decrease individual scores, and thereby decrease reliability, comes from four sources:

- The test itself can generate error by being poorly designed; by containing trick questions, ambiguous questions, or poorly written questions; or by requiring a reading level higher than the reading level of the test takers. (Chapter 10 provides information on designing survey and test questions that yield a minimal amount of error.)
- The test administration can generate error when administrators do not follow instructions for administration in the test manual or allow disturbances to occur during the test period. For example, the test administrator might misread the instructions for the length of the test period; answer test takers’ questions inappropriately; allow the room to be too hot, cold, or noisy; or display attitudes that suggest the test is too difficult or unimportant.
- The test scoring can generate error if it is not conducted accurately and according to the directions in the test manual. For example, scorers might make errors in judgment or in calculating test scores. Although computer scoring is likely to decrease scoring errors, it is important to enter the correct scoring scheme into the computer software.
Test takers themselves also can contribute to test error. Fatigue, illness, or exposure to test questions before taking the test can change test scores. In addition, test takers who do not provide truthful and honest answers introduce error into their test scores.

Six factors related to these sources of error—test length, homogeneity of questions, test–retest interval, test administration, scoring, and cooperation of test takers—stand out as particularly important and worthy of consideration in detail. Test developers and administrators focus on these factors to increase the reliability and accuracy of the test scores.

**Test Length**

As a rule, adding more questions that measure the same trait or attribute can increase a test's reliability. Each question on a test serves as an observation that indicates the test taker's knowledge, skill, ability, or trait being measured. The more observations there are on the construct that the test is designed to measure, the less random error will contribute to the observed scores and the more accurate the measure is likely to be.

Adding more questions to a test is similar to adding finer distinctions to a measuring tape, for example, adding indications for each 16th of an inch to a tape that previously had indications only for each 8th of an inch. Likewise, shortening a test by skipping or dropping questions causes the test to lose reliability. An extreme example is the test that has only one question—a most unreliable way to measure any trait or attitude.

As you recall, the Spearman–Brown formula adjusts the reliability estimate for test length. Test developers can also use the Spearman–Brown formula to estimate the number of questions to add to a test so as to increase its reliability to the desired level.

Embretson (1996) pointed out an important exception to this rule when using adaptive tests (e.g., the computer-based version of the GRE). A short adaptive test can be more reliable than a longer version. In an adaptive test, the test taker responds to questions selected based on his or her skill or aptitude level, and therefore the SEM decreases. As a result, the test taker answers fewer questions without sacrificing reliability. This circumstance, however, does not suggest that a test made up of one question or only a few questions would be reliable.

**Homogeneity**

Another important exception to the rule that adding questions increases reliability is that an increase in test questions will increase reliability only when the questions added are homogeneous (very much alike) with other questions on the test. That is, to increase reliability, the test developer must add questions that measure the same attribute as the other questions on the test. Heterogeneous (very different or diverse) tests can be expected to have lower reliability coefficients. As you recall, estimating reliability by calculating internal consistency is not appropriate for heterogeneous tests. If you have ever taken a test in which it seemed you were asked the same questions a number of times in slightly different ways, you have experienced a test that is homogeneous and probably very reliable.

**Test–Retest Interval**

The longer the interval between administrations of a test, the lower the reliability coefficient is likely to be. A long interval between test administrations provides more opportunity for test takers to change in terms of the factor being measured. Such changes cause a change in individuals' true scores. In addition,
the longer time increases the possibility of error through changes in test administration, environment, or personal circumstances. A long interval may lessen practice effects; however, a better way to decrease practice effects would be to use alternate forms.

**Test Administration**

Proper test administration affects the reliability estimate in three ways. First, carefully following all of the instructions for administering a test ensures that all test takers experience the same testing situation each time the test is given. In other words, test takers hear the same instructions and take the test under the same physical conditions each time. Treating all test takers in the same way decreases error that arises from creating differences in the way individuals respond. Second, constancy between two administrations decreases error that arises when testing conditions differ. Third, effective testing practices (discussed in detail in Chapter 11) decrease the chance that test takers’ scores will be contaminated with error due to poor testing conditions or poor test instructions.

**Scoring**

Even tests scored by computer are subject to incorrect scoring. Test users must be careful to use the correct scoring key, to check questions that have unusually large numbers of correct or incorrect answers for mistakes in scoring, and to exercise considerable care when scoring tests that require judgments about whether an answer is right or wrong. Frequent checks of computations—including those made by computers—also decrease the chance of scoring errors. Scorers who will make qualitative judgments when scoring tests, such as using a rating scale, must receive training together to calibrate their judgments and responses.

**Cooperation of Test Takers**

Some tests, such as the PAI, have a built-in method for determining whether test takers guessed, faked, cheated, or in some other way neglected to answer questions truthfully or to the best of their ability. Many times, however, it is up to the test administrator to observe and motivate respondents to cooperate with the testing process. For instance, test administrators need to be aware of individuals who complete the test in an unusually short amount of time. These individuals might have checked answers without reading the questions or skipped whole pages either deliberately or by mistake. Although respondents cannot be forced to participate honestly, their tests can be dropped from the group of tests used to calculate reliability when there are doubts about the truthfulness of their answers.

**Generalizability Theory**

Up to now in this chapter, we have used classical test theory to describe the processes for measuring a test’s consistency or reliability. Another approach to estimating reliability/precision is generalizability theory, proposed by Cronbach, Gleser, Nanda, and Rajaratnam (1972). This theory concerns how well and under what conditions we can generalize an estimation of reliability/precision of test scores from one test administration to another. In other words, the test user can predict the reliability/precision of test scores obtained under different circumstances, such as administering a test in various plant locations or school systems. Generalizability theory proposes separating sources of systematic error from random error to eliminate systematic error.
Why is the separation of systematic error and random error important? As you recall, we can assume that if we were able to record the amount of random error in each measurement, the average error would be zero, and over time random error would not interfere with obtaining an accurate measurement. However, systematic error does affect the accuracy of a measurement; therefore, using generalizability theory, our goal is to eliminate systematic error.

For example, if you weigh yourself once a week in the gym, your weight will consist of your true weight and measurement error. One possible source of measurement error would be random error in the scale or in your precision in reading the scale. But another source of the measurement error could be the weight of your clothes and shoes. Another source might be the time of day when you weigh yourself; generally speaking, you will weigh less in the morning than you will later in the day. These sources of error would not be random, but would be more systematic because each time they occurred, they would have the same influence on the measurement.

Using generalizability theory, you could look for systematic or ongoing predictable error that occurs when you weigh yourself. For instance, the weight of your clothes and shoes will vary systematically depending on the weather and the time of the year. Likewise, your weight will be greater later in the day. On the other hand, variations in the measurement mechanism and your ability to read the scale accurately vary randomly. We would predict, therefore, that if you weighed yourself at the same time of day wearing the same clothes (or, better yet, none at all), you would have a more accurate measurement of your weight. When you have the most accurate measurement of your weight, you can confidently assume that changes in your weight from measurement to measurement are due to real weight gain or loss and not to measurement error.

Researchers and test developers identify systematic error in test scores by using the statistical procedure called analysis of variance. As you recall, we discussed four sources of error: the test itself, test administration, test scoring, and the test taker. Researchers and test developers can set up a generalizability study in which two or more sources of error (the independent variables) can be varied for the purpose of analyzing the variance of the test scores (the dependent variable) to find systematic error.

As an example, consider the situation in which 20 employees participate in three business simulations all designed to measure the same set of leadership skills. The employees are all observed and scored by the same two raters. So we have the scores of each of two raters scoring 20 employees on three simulations, or 120 scores. As you would expect, these scores will not all be the same, but rather they will vary. The question becomes, Why do the scores vary? Well, intuitively you probably realize that employees’ scores might vary because of the differing levels of leadership skills present in the employee group. This is the object of our measurement. But is the level of leadership skills possessed by each employee the only reason why the scores on the simulations might vary? Probably not in this example.

Another reason that could cause the scores to vary is that although the simulations were all designed to measure the same leadership skills, perhaps the simulations are not equally difficult. Or perhaps one of the simulations is easier for employees who happen to have a background in finance, while another of the simulations is easier for employees with a background in sales. Another possibility is that one of the raters might be systematically more lenient or stringent than the other rater across all the simulations when rating the performance of the employees. Finally, a combination of conditions could contribute to the variance in the scores, as would happen if a particular rater tended to give employees evaluated earlier in the day higher ratings than those evaluated later in the day.

The beauty of generalizability theory is that it allows you to actually quantify each of these (and other) possible sources of variation so that you can determine whether the results you obtain are likely to generalize (thus the name) to a different set of employees evaluated by different raters on different occasions. Using this approach you would be able to tell the degree to which each of the facets (simulations, raters, and their interaction) contributed to the variations in the leadership skill scores of the employees. In this
case, we would hope that the main contributor to the variation in scores was the skill level of the employees because that is the focus or object of our measurement. In other cases, we might be more interested in the variation in scores attributable to the simulations themselves or the consistency of the raters.

As you have learned, at the heart of the concept of reliability/precision is the idea of consistency of measurement. If the same employees went through the same set of simulations a second time, we would like to expect that their scores would be similar to what they were in the first administration. If they were not, we might conclude that the simulations were not reliable measures of the leadership skills they were designed to measure. However, if the reason why the scores were different on the second administration was the fact that we used a different set of raters who differed in the way they scored the simulations from the original set of raters, it would be incorrect to conclude that the simulations were unreliable. The actual source of the unreliability in this case would be error caused by scoring differences between the first and second sets of raters. Using generalizability would enable us to separate the variance in the employee's scores that was attributable to the raters from the variance in the scores that was due to the simulations themselves.

This approach is conceptually different from the classical measurement of the reliability of a test, because classical reliability measurement focuses on the amount of random measurement error and cannot separately evaluate error that may be systematic. The actual calculations are somewhat complicated and beyond the scope of this book, but we wanted to give you an idea of another approach that can be used to evaluate the reliability of a measure.

---

**INTERIM SUMMARY 6.3**

**FACTORS THAT INFLUENCE RELIABILITY**

- Errors that increase or decrease individual scores and change the reliability estimate result from four sources: the test itself, test administration, test scoring, and test takers.
- Errors resulting from poor test design include trick, ambiguous, or poorly worded questions and reading levels that are too high for the test takers.
- Test administration can generate error when administrators do not follow instructions for test administration or when the testing environment is uncomfortable or distracting.
- Tests must be scored accurately and according to the instructions in the test manual.
- Test takers can contribute to test error by being fatigued or ill, by cheating, or by providing dishonest answers.
- Reliability is related to test length—the longer the test, the more reliable it is likely to be—provided that the questions on the test are homogeneous.
- Although classical test theory remains the most popular method for sorting true scores from error, generalizability theory is also useful. Generalizability theory analyzes the many causes of inconsistency error in test scores and seeks to separate sources of systematic error from random error.

---

**Chapter Summary**

Psychological tests are measurement instruments. An important attribute of a measurement instrument is its reliability/precision or consistency. We need evidence that the test yields the same score each time a person takes the test unless the test taker has actually changed. When we know a test is reliable, we can
conclude that changes in a person’s score really are due to changes in that person. Also, we can compare the scores of two or more people on a reliable test.

Test developers use three methods for checking reliability. Each takes into account various conditions that could produce differences in test scores. Using the test–retest method, a test developer gives the same test to the same group of test takers on two different occasions. The scores from the first and second administrations are then correlated to obtain the reliability coefficient. The greatest danger in using the test–retest method of estimating reliability/precision is that the test takers will score differently (usually higher) on the test because of practice effects. To overcome practice effects and differences in individuals and the test administration from one time to the next, psychologists often give two forms of the same test—alike in every way—to the same people at the same time. This method is called alternate or parallel forms.

If a test taker can take the test only once, researchers divide the test into halves and correlate the scores on the first half with the scores on the second half. This method, called split-half reliability, includes using the Spearman–Brown formula to adjust the correlation coefficient for test length. A more precise way to measure internal consistency is to compare individuals’ scores on all possible ways of splitting the test into halves. The KR-20 and coefficient alpha formulas allow researchers to estimate the reliability of the test scores by correlating the answer to each test question with the answers to all of the other test questions.

The reliability of scoring is also important. Tests that require the scorer to make judgments about the test takers’ answers and tests that require the scorer to observe the test takers’ behavior may have error contributed by the scorer. We estimate scorer reliability by having two or more persons score the same test and then correlating their scores to see whether their judgments are consistent or have a single person score two occasions of the same test.

No measurement instrument is perfectly reliable or consistent. We express this idea by saying that each observed test score \( X \) contains two parts: a true score \( T \) and error \( E \). Two types of error appear in test scores: random error and systematic error. The more random error present in a set of test scores, the lower the reliability coefficient will be. Another way of saying the same thing is that the higher the proportion of true score variance is of the observed scores, the higher the reliability coefficient will be.

To quantify a test’s reliability/precision estimate, we use a reliability coefficient, which is another name for the correlation coefficient when it estimates reliability/precision. This statistic quantifies the estimated relationship between two forms of the test. The statistical procedure we use most often to calculate the reliability coefficient is the Pearson product–moment correlation. All statistical software programs and many spreadsheet programs will calculate the Pearson product–moment correlation. Coefficient alpha and KR-20, both of which also use correlation, are available in statistical packages only.

To interpret the meaning of the reliability coefficient, we look at its sign and the number itself. Reliability coefficients range from –0.00 (a completely unreliable test) to +1.00 (a perfectly reliable test). Psychologists have not set a fixed value at which reliability can be interpreted as satisfactory or unsatisfactory.

Psychologists use the standard error of measurement (SEM) as an index of the amount of inconsistency or error expected in an individual’s test score. We can then use the SEM to construct a confidence interval—a range of scores that most likely includes the true score. Confidence intervals provide information about whether individuals’ observed test scores are statistically different from each other. Six factors—test length, homogeneity of questions, the test–retest interval, test administration, scoring, and cooperation of test takers—are important factors that influence the reliability of the test scores.

Another approach to estimating reliability is generalizability theory, which concerns how well and under what conditions we can generalize an estimation of reliability from one test to another or on the
same test given under different circumstances. Generalizability theory seeks to identify sources of systematic error that classical test theory would simply label as random error. Using analysis of variance, researchers and test developers can identify systematic error and then take measures to eliminate it, thereby increasing the overall reliability of the test.

### Engaging in the Learning Process

Learning is the process of gaining knowledge and skills through schooling or studying. Although you can learn by reading the chapter material, attending class, and engaging in discussion with your instructor, more actively engaging in the learning process may help you better learn and retain chapter information. To help you actively engage in the learning process, we encourage you to access our new supplementary student workbook. The workbook contains a variety of information and learning activities to help you understand and apply information, and help you make progress toward learning and retaining material. If you do not have a copy of the workbook, you can purchase a copy through sagepub.com.

### Key Concepts

After completing your study of this chapter, you should be able to define each of the following terms. These terms are bolded in the text of this chapter and defined in the Glossary.

- alternate forms
- Cohen's kappa
- confidence interval
- correlation
- generalizability theory
- heterogeneous test
- homogeneous test
- internal consistency
- interrater agreement
- interscorer agreement
- intrarater agreement
- intrascorer reliability
- measurement error
- order effects
- parallel forms
- practice effects
- random error
- reliability/precision
- reliable test
- scorer reliability
- split-half method
- standard error of measurement (SEM)
- systematic error
- test–retest method

### Learning Activities

The following are some learning activities you can engage in to support the learning objectives for this chapter.

<table>
<thead>
<tr>
<th>Learning Objectives</th>
<th>Study Tips and Learning Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>After completing your study of this chapter, you should be able to do the following:</td>
<td>The following study tips will help you meet these learning objectives:</td>
</tr>
<tr>
<td>Define reliability/precision, and describe three methods for estimating the reliability/precision of a psychological test and its scores.</td>
<td>• Distinguish among the three methods by writing down who, what, and when for each method.</td>
</tr>
</tbody>
</table>
### Learning Objectives

<table>
<thead>
<tr>
<th>Describe how an observed test score is made up of the true score and random error, and describe the difference between random error and systematic error.</th>
</tr>
</thead>
</table>

### Study Tips and Learning Activities

<table>
<thead>
<tr>
<th>• Weigh yourself several times on your bathroom scale. Do you get the same weight each time? If not, your scale may have random error. Now try another scale. Compare your weights on the first and second scales. Does your first scale weigh too heavy or too light? If so, your scale may have systematic error.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Calculate and interpret a reliability coefficient, including adjusting a reliability coefficient obtained using the split-half method.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>• The learning activities that follow are designed to help you understand how to calculate and interpret reliability coefficients. Complete the activities. If you have questions, be sure to ask your instructor.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Differentiate between the KR-20 and coefficient alpha formulas, and understand how they are used to estimate internal consistency.</th>
</tr>
</thead>
</table>

| • Write a short answer in complete sentences that answers these questions:
  - What kinds of questions require the KR-20 formula?
  - What kinds of questions require the coefficient alpha formula? |
|---|

<table>
<thead>
<tr>
<th>Calculate the standard error of measurement, and use it to construct a confidence interval around an observed score.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>• Weigh yourself again three times in a row. Do not adjust the scale if your weight appears to vary. Calculate the standard error of measurement of the three weights, and construct a confidence interval around your mean weight. Interpret your weight in terms of the confidence interval.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Identify the four sources of test error and six factors related to these sources of error that are particularly important to consider.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>• Try to come up with your own examples for each of these. Check with a classmate and/or your instructor to see whether he or she agrees with your examples.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Explain the premises of generalizability theory, and describe its contribution to estimating reliability.</th>
</tr>
</thead>
</table>

| • Review the example study proposed in the section on generalizability theory. Then propose your own study for detecting systematic error in test results. Share your proposed study with your classmates or your instructor. |

---

**SAGE edge**

Sharpen your skills with SAGE edge at [edge.sagepub.com/miller5e](http://edge.sagepub.com/miller5e)

**SAGE edge** for Students provides a personalized approach to help you accomplish your coursework goals in an easy-to-use learning environment.