CHAPTER

4 Reliability for Teachers

It is the user who must take responsibility for determining whether or not scores are sufficiently trustworthy to justify anticipated uses and interpretations.

—AERA et al., 1999, p. 31

CHAPTER HIGHLIGHTS

Errors of Measurement
Methods of Estimating Reliability
The Standard Error of Measurement
Reliability: Practical Strategies for Teachers

LEARNING OBJECTIVES

After reading and studying this chapter, students should be able to

1. Define and explain the importance of reliability in educational assessment.
2. Define and explain the concept of measurement error.
3. Explain classical test theory and its importance to educational assessment.
4. Describe the major sources of measurement error and give examples.
5. Identify the major methods for estimating reliability and describe how these analyses are performed.
6. Identify the sources of measurement error that are reflected in different reliability estimates.
7. Explain how multiple scores can be combined in a composite to enhance reliability.
8. Describe the factors that should be considered when selecting a reliability coefficient for a specific assessment application.
9. Explain the factors that should be considered when evaluating the magnitude of reliability coefficients.
10. Describe steps that can be taken to improve reliability.
11. Discuss special issues in estimating reliability such as estimating the reliability of speed tests and mastery testing.
12. Define the standard error of measurement (SEM) and explain its importance.
13. Explain how SEM is calculated and describe its relation to reliability.
14. Explain how confidence intervals are calculated and used in educational and psychological assessment.
15. Describe and apply shortcut procedures for estimating the reliability of classroom tests.

In simplest terms, in the context of measurement, *reliability* refers to consistency or stability of assessment results.

Most dictionaries define *reliability* in terms of dependability, trustworthiness, or having a high degree of confidence in something. Reliability in the context of educational and psychological measurement is concerned to some extent with these same factors, but is extended to such concepts as stability and consistency. In simplest terms, in the context of measurement, **reliability** refers to consistency or stability of assessment results. Although it is common for people to refer to the "reliability of a test," in the *Standards for Educational and Psychological Testing* (AERA et al., 1999) reliability is considered to be a characteristic of scores or assessment results, not tests themselves.

Consider the following example: A teacher administers a 25-item math test in the morning to assess the students' skill in multiplying two-digit numbers. If the test had been administered in the afternoon rather than the morning, would Susie's score on the test have been the same? Because there are literally thousands of two-digit multiplication problems, if the teacher had used a different group of 25 two-digit multiplication problems, would Susie have received the same score? What about the ambulance that went by, its siren wailing loudly, causing Johnny to look up and watch for a few seconds? Did this affect his score, and did it affect Susie's, who kept working quietly? Jose wasn't feeling well that morning but came to school because he felt the test was so important. Would his score have been better if he had waited to take the test when he was feeling better? Would the students have received the same scores if another teacher had graded the test? All of these questions involve issues of reliability. They all ask if the test produces consistent scores.

As you can see from these examples, numerous factors can affect reliability. The time the test is administered, the specific set of questions included on the test, distractions due to external (e.g., ambulances) or internal (e.g., illness) events, and the person grading the test are just a few of these factors. In this chapter you will learn to take many of the sources of unreliability into account when selecting or developing assessments and evaluating scores. You will also learn to estimate the degree of reliability in test scores with a method that best fits your particular situation. First, however, we will introduce the concept of measurement error as it is essential to developing a thorough understanding of reliability.

Errors of Measurement

Some degree of *measurement error* is inherent in all measurement.

Some degree of error is inherent in all measurement. Although **measurement error** has largely been studied in the context of psychological and educational tests, measurement error clearly is not unique to this context. In fact, as Nunnally and Bernstein (1994) point out, measurement in other scientific disciplines has as much, if not more, error than that in psychology and education. They give the example of physiological blood pressure measurement, which is considerably less reliable than many educational tests. Even in situations in which we generally believe measurement is exact, some error is present. If we asked a dozen people to time a 440-yard race using the same brand of stopwatch, it is extremely unlikely that they would all report precisely the same time. If we had a dozen people and a measuring tape graduated in millimeters and required each person to measure independently

the length of a 100-foot strip of land, it is unlikely all of them would report the same answer to the nearest millimeter. In the physical sciences the introduction of more technologically sophisticated measurement devices has reduced, but not eliminated, measurement error.

Different theories or models have been developed to address measurement issues, but possibly the most influential is classical test theory (also called true score theory). According to this theory, every score on a test is composed of two components: the **true score** (i.e., the score that would be obtained if there were no errors, if the score were perfectly reliable) and the **error score:** Obtained Score = True Score + Error. This can be represented in a very simple equation:

$$X_i = T + E$$

Here we use X_i to represent the observed or **obtained score** of an individual; that is, X_i is the score the test taker received on the test. The symbol T is used to represent an individual's true score and reflects the test taker's true skills, abilities, knowledge, attitudes, or whatever the test measures, assuming an absence of measurement error. Finally, E represents measurement error.

Measurement error reduces the usefulness of measurement. It limits the extent to which test results can be generalized and reduces the confidence we have in test results (AERA et al., 1999). Practically speaking, when we administer a test we are interested in knowing the test taker's true score. Due to the presence of measurement error we can never know with absolute confidence what the true score is. However, if we have information about the reliability of measurement, we can establish intervals around an obtained score and calculate the probability that the true score will fall within the interval specified. We will come back to this with a more detailed explanation when we discuss the standard error of measurement later in this chapter. First, we will elaborate on the major sources of measurement error. It should be noted that we will limit our discussion to random measurement error. Some writers distinguish between random and systematic errors. Systematic error is much harder to detect and requires special statistical methods that are generally beyond the scope of this text; however, some special cases of systematic error are discussed in Chapter 16. (Special Interest Topic 4.1 provides a brief introduction to Generalizability Theory, an extension of classical reliability theory.)

Measurement error limits the extent to which test results can be generalized and reduces the confidence we have in test results (AERA et al., 1999).

Sources of Measurement Error

Because measurement error is so pervasive, it is beneficial to be knowledgeable about its characteristics and aware of the methods that are available for estimating its magnitude. As educational professionals, we should also work to identify sources of measurement error and minimize their impact to the extent possible. Generally, whenever you hear a discussion of reliability or read about the reliability of test scores, it is the relative freedom from measurement error that is being discussed. Reliable assessment results are relatively free from measurement error whereas less reliable

As educational professionals, we should work to identify sources of measurement error and minimize their impact to the extent possible.

SPECIAL INTEREST TOPIC 4.1

Generalizability Theory

Lee Cronbach and colleagues developed an extension of classical reliability theory known as "generalizability theory" in the 1960s and 1970s. Cronbach was instrumental in the development of the general theory of reliability discussed in this chapter during and after World War II. The basic focus of generalizability theory is to examine various conditions that might affect the reliability of a test score. In classical reliability theory there are only two sources for variation in an observed test score: true score and random error. Suppose, however, that for different groups of people the scores reflect different things. For example, boys and girls might respond differently to career interest items. When the items for a particular career area are then grouped into a scale, the reliability of the scale might be quite different for boys and girls as a result. This gender effect becomes a limitation on the generalizability of the test's functioning with respect to reliability.

Generalizability theory extends the concept of reliability as the ratio of true score variance to total score variance by adding other possible sources of true score variation to both the numerator and the denominator of the reliability estimate. Because gender is a reliable indicator, if there is significant gender variation on a test scale due to gender differences, this additional variation will change the original true score variation. What originally appeared to be high true score variation might instead be a modest true score variation and large gender variation. In some instances the true score variation may be high within the boys' group but near zero within the girls' group. Thus, in this study, gender limits the generalizability of the test with respect to reliability.

The sources of variation that might be considered are usually limited to theoretically relevant characteristics of population, ecology, or time. For example, population characteristics may include gender, ethnicity, or region of residence. These are usually discussed as fixed sources, because typically all characteristics will be present in the data analysis (male and female, all ethnic groups of interest, or all regions to be considered). Sources are considered random when only a sample of the possible current or future population is involved in the analysis. Common random sources include raters or observers; classrooms, schools, or districts; clinics or hospitals; or other organized groupings in which respondents are placed. In a nursing school, for example, students may be evaluated in a series of activities that they are expected to have mastered (administering an injection, adjusting a drip, and determining medication levels). Several instructors might rate each activity. Because we are usually interested in how reliable the ratings are for raters like those in the study, but not just those specific raters, the rater source is considered a random source. Random sources are always included only in the denominator of the reliability ratio (discussed in the next major section). That is, the variance associated with raters will be added to the error variance and true score variance only to the total variance term.

Although calculating generalizability coefficients is beyond the scope of this text, the general procedure is to use a statistical analysis program such as Statistical Package for the Social Sciences (SPSS) or Statistical Analysis System (SAS). These statistical programs have analysis options that will estimate the variance components (i.e., the variances of each source of variation specified by the analyst). A numerical value for each source is obtained, and the generalizability value is calculated from specific rules of computation that have been derived over the last several decades. Some psychometricians advocate simply examining the magnitude of the variances. For example, if true score variance in the gender study mentioned is 10, but gender variance is 50, while error variance per item is 2, it is clear that most of the apparent reliability of the test is due to gender differences rather than individual differences. Boy and girl studies might be conducted separately at this point. In the nursing study, if rater variance is 3 and individual true score variance is 40, it is clear without further study that raters will have little effect on the reliability of the nursing assessments.

results are not. A number of factors may introduce error into test scores and even though all cannot be assigned to distinct categories, it may be helpful to group these sources in some manner and to discuss their relative contributions. The types of errors that are our greatest concern are errors due to content sampling and time sampling.

Content Sampling Error. Tests rarely, if ever, include every possible question or evaluate every possible relevant behavior. Let's revisit the example we introduced at the beginning of this chapter. A teacher administers a math test designed to assess skill in multiplying two-digit numbers. We noted that there are literally thousands of two-digit multiplication problems. Obviously it would be impossible for the teacher to develop and administer a test that includes all possible items. Instead, a universe or domain of test items is defined based on the content of the material to be covered. From this domain a sample of test questions is taken. In this example, the teacher decided to select 25 items to measure students' ability. These 25 items are simply a sample and, as with any sampling procedure, may not be representative of the domain from which they are drawn. The error that results from differences between the sample of items (i.e., the test) and the domain of items (i.e., all the possible items) is referred to as **content sampling error.** In reading other sources, you might see this type of error referred to as domain sampling error. Domain sampling error and content sampling error are the same. Content sampling error typically is considered the largest source of error in test scores and therefore is the source that concerns us most. Fortunately, content sampling error is also the easiest and most accurately estimated source of measurement error.

The amount of measurement error due to content sampling is determined by how well we sample the total domain of items. If the items on a test are a good sample of the domain, the amount of measurement error due to content sampling will be relatively small. If the items on a test are a poor sample of the domain, the amount of measurement error due to content sampling will be relatively large. Measurement error resulting from content sampling is estimated by analyzing the degree of statistical similarity among the items making up the test. In other words, we analyze the test items to determine how well they correlate with one another and with the test taker's standing on the construct being measured. We will explore a variety of methods for estimating measurement errors due to content sampling later in this chapter.

If the items on a test are a good sample of the domain, the amount of measurement error due to content sampling will be relatively small.

Time Sampling Error. Measurement error also can be introduced by one's choice of a particular time to administer the test. If Eddie did not have breakfast and the math test was just before lunch, he might be distracted or hurried and not perform as well as if he took the test after lunch. But Michael, who ate too much at lunch and was up a little late last night, was a little sleepy in the afternoon and might not perform as well on an afternoon test as he would have on the morning test. If during the morning testing session a neighboring class was making enough noise to be disruptive, the class might have performed better in the afternoon when the neighboring class was relatively quiet. These are all examples of situations in which random changes over time in the test taker (e.g., fatigue, illness, anxiety) or the testing environment (e.g., distractions, temperature) affect

Measurement error due to time sampling reflects random fluctuations in performance from one situation to another and limits our ability to generalize test scores across different situations.

Metho

performance on the test. This type of measurement error is referred to as **time sampling error** and reflects random fluctuations in performance from one situation or time to another and limits our ability to generalize test results across different situations. Some assessment experts refer to this type of error as temporal instability. As you might expect, testing experts have developed methods of estimating error due to time sampling.

Other Sources of Error. Although errors due to content sampling and time sampling account for the major proportion of random error in testing, administrative and scoring errors that do not affect all test takers equally will also contribute to the random error observed in scores. Clerical errors committed while adding up a student's score or an administrative error on an individually administered test are common examples. When the scoring of a test relies heavily on the subjective judgment of the person grading the test or involves subtle discriminations, it is important to consider differences in graders, usually referred to as inter-scorer or **inter-rater differences.** That is, would the test taker receive the same score if different individuals graded the test? For example, on an essay test would two different graders assign the same scores? These are just a few examples of sources of error that do not fit neatly into the broad categories of content or time sampling errors.

Methods of Estimating Reliability

You will note that we are referring to reliability as being *estimated.* This is because the absolute or precise reliability of assessment results cannot be known. Just as we always have some error in test scores, we also have some error in our attempts to measure reliability. Earlier in this chapter we introduced the idea that test scores are composed of two components, the true score and the error score. We represented this with the equation:

$$X_i = T + E$$

As you remember, X_i represents an individual's obtained score, T represents the true score, and E represents random measurement error. This equation can be extended to incorporate the concept of variance. This extension indicates that the variance of test scores is the sum of the **true score variance** plus the **error variance,** and is represented in the following equation:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

Here, σ_X^2 represents the variance of the total test, σ_T^2 represents true score variance, and σ_E^2 represents the variance due to measurement error. True score variance reflects differences in test takers due to real differences in skills, abilities, knowledge, attitudes, and so on, whereas the total score variance is made up of true score variance plus variance due to all the sources of random error we have previously described.

The general symbol for the reliability of assessment results associated with content or domain sampling is r_{xx} and is referred to as the **reliability coefficient.** We estimate the reliability of a test score as the ratio of true score variance to total score variance. Mathematically, reliability is written

$$r_{xx} = \sigma_T^2 / \sigma_X^2$$

This equation defines the reliability of test scores as the proportion of test score variance due to true score differences. The reliability coefficient is considered to be the summary mathematical representation of this ratio or proportion.

Reliability coefficients can be classified into three broad categories (AERA et al., 1999). These include (1) coefficients derived from the administration of the same test on different occasions (i.e., test-retest reliability), (2) coefficients based on the administration of parallel forms of a test (i.e., alternate-form reliability), and (3) coefficients derived from a single administration of a test (internal consistency coefficients). A fourth type, inter-rater reliability, is indicated when scoring involves a significant degree of subjective judgment. The major methods of estimating reliability are summarized in Table 4.1. Each of these approaches produces a reliability coefficient (r_{xx}) that can be interpreted in terms of the proportion or percentage of test score variance attributable to true variance. For example, a reliability coefficient of 0.90 indicates that 90% of the variance in test scores is attributable to true variance. The remaining 10% reflects error variance. We will now consider each of these methods of estimating reliability.

Reliability can be defined as the proportion of test score variance due to true score differences.

TABLE 4.1 Major Types of Reliability

Type of Reliability Estimate	Common Symbol	Number of Test Forms	Number of Testing Sessions	Summary
Test-Retest	r_{12}	One form	Two sessions	Administer the same test to the same group at two different sessions.
Alternate forms				
Simultaneous administration	r_{ab}	Two forms	One session	Administer two forms of the test to the same group in the same session.
Delayed administration	r_{ab}	Two forms	Two sessions	Administer two forms of the test to the same group at two different sessions.
Split-half	r_{oe}	One form	One session	Administer the test to a group one time. Split the test into two equivalent halves, typically correlating scores on the odd-numbered items with scores on the even-numbered items.
Coefficient alpha or KR-20	r_{xx}	One form	One session	Administer the test to a group one time. Apply appropriate procedures.
Inter-rater	r	One form	One session	Administer the test to a group one time. Two or more raters score the test independently.

Test-Retest Reliability

Probably the most obvious way to estimate the reliability of a test is to administer the same test to the same group of individuals on two different occasions. With this approach the reliability coefficient is obtained by simply calculating the correlation between the scores on the two administrations. For example, we could administer our 25-item math test one week after the initial administration and then correlate the scores obtained on the two administrations. This estimate of reliability is referred to as **test-retest reliability** and is sensitive to measurement error due to time sampling. It is an index of the stability of test scores over time. Because many tests are intended to measure fairly stable characteristics, we expect tests of these constructs to produce stable scores. Test-retest reliability reflects the degree to which test scores can be generalized across different situations or over time.

***Test-retest reliability* is sensitive to measurement error due to time sampling and is an index of the stability of scores over time.**

One important consideration when calculating and evaluating test-retest reliability is the length of the interval between the two test administrations. If the test-retest interval is very short (e.g., hours or days), the reliability estimate may be artificially inflated by memory and practice effects from the first administration. If the test interval is longer, the estimate of reliability may be lowered not only by the instability of the scores but also by actual changes in the test takers during the extended period. In practice, there is no single "best" time interval, but the optimal interval is determined by the way the test results are to be used. For example, intelligence is a construct or characteristic that is thought to be fairly stable, so it would be reasonable to expect stability in intelligence scores over weeks or months. In contrast, an individual's mood (e.g., depressed, elated, nervous) is more subject to transient fluctuations, and stability across weeks or months would not be expected.

One important consideration when calculating and evaluating test-retest reliability is the length of the interval between the two test administrations.

In addition to the construct being measured, the way the test is to be used is an important consideration in determining what is an appropriate test-retest interval. Because the SAT is used to predict performance in college, it is sensible to expect stability over relatively long periods of time. In other situations, long-term stability is much less of an issue. For example, the long-term stability of a classroom achievement test (such as our math test) is not a major concern because it is expected that the students will be enhancing existing skills and acquiring new ones due to class instruction and studying. In summary, when evaluating the stability of test scores, one should consider the length of the test-retest interval in the context of the characteristics being measured and how the scores are to be used.

The test-retest approach does have significant limitations, the most prominent being carryover effects from the first to second testing. Practice and memory effects result in different amounts of improvement in retest scores for different test takers. These carryover effects prevent the two administrations from being independent and as a result the reliability coefficients may be artificially inflated. In other instances, repetition of the test may change either the nature of the test or the test taker in some subtle or even obvious way (Ghiselli, Campbell, & Zedeck, 1981). As a result,

The test-retest approach does have significant limitations, the most prominent being carryover effects from the first to second testing.

only tests that are not appreciably influenced by these carryover effects are suitable for this method of estimating reliability.

Alternate-Form Reliability

Another approach to estimating reliability involves the development of two equivalent or parallel forms of the test. The development of these alternate forms requires a detailed test plan and considerable effort because the tests must truly be parallel in terms of content, difficulty, and other relevant characteristics. The two forms of the test are then administered to the same group of individuals and the correlation is calculated between the scores on the two assessments. In our example of the 25-item math test, the teacher could develop a parallel test containing 25 new problems involving the multiplication of double digits. To be parallel the items would need to be presented in the same format and be of the same level of difficulty. Two fairly common procedures are used to establish **alternate-form reliability.** One is alternate-form reliability based on simultaneous administrations and is obtained when the two forms of the test are administered on the same occasion (i.e., back to back). The other, alternate form with delayed administration, is obtained when the two forms of the test are administered on two different occasions. Alternate-form reliability based on simultaneous administration is primarily sensitive to measurement error related to content sampling. Alternate-form reliability with delayed administration is sensitive to measurement error due to both content sampling and time sampling.

***Alternate-form reliability* based on simultaneous administration is primarily sensitive to measurement error due to content sampling.**

Alternate-form reliability has the advantage of reducing the carryover effects that are a prominent concern with test-retest reliability. However, although practice and memory effects may be reduced using the alternate-form approach, they are often not fully eliminated. Simply exposing test takers to the common format required for parallel tests often results in some carryover effects even if the content of the two tests is different. For example, a test taker given a test measuring nonverbal reasoning abilities may develop strategies during the administration of the first form that alter her approach to the second form, even if the specific content of the items is different. Another limitation of the alternate-form approach to estimating reliability is that relatively few tests, standardized or teacher made, have alternate forms. As we suggested, the development of alternate forms that are actually equivalent is a time-consuming process, and many test developers do not pursue this option. Nevertheless, at times it is desirable to have more than one form of a test, and when multiple forms exist, alternate-form reliability is an important consideration.

Alternate-form reliability based on delayed administration is sensitive to measurement error due to content sampling and time sampling, but cannot differentiate the two types of error.

Internal-Consistency Reliability

Internal-consistency reliability estimates primarily reflect errors related to content sampling. These estimates are based on the relationship between items within a test and are derived from a single administration of the test.

Split-Half Reliability. Estimating **split-half reliability** involves administering a test and then dividing the test into two equivalent halves that are scored independently. The results on the first half of the test are then correlated with results on the other half of the test by calculating the Pearson product-moment correlation. Obviously, there are many ways a test can be divided in half. For example, one might correlate scores on the first half of the test with scores on the second half. This is usually not a good idea because the items on some tests get more difficult as the test progresses, resulting in halves that are not actually equivalent. Other factors, such as practice effects, fatigue, or declining attention that increases as the test progresses, can also make the first and second halves of the test not equivalent. A more acceptable approach would be to assign test items randomly to one half or the other. However, the most common approach is to use an odd–even split. Here all "odd"-numbered items go into one half and all "even"-numbered items go into the other half. A correlation is then calculated between scores on the odd-numbered and even-numbered items.

Split-half reliability **can be calculated from one administration of a test and reflects error due to content sampling.**

Before we can use this correlation coefficient as an estimate of reliability, there is one more task to perform. Because we are actually correlating two halves of the test, the reliability coefficient does not take into account the reliability of the test when the two halves are combined. In essence, this initial coefficient reflects the reliability of only a shortened, half-test. As a general rule, longer tests are more reliable than shorter tests. If we have twice as many test items, then we are able to sample the domain of test questions more accurately. The better we sample the domain the lower the error due to content sampling and the higher the reliability of our test. To "put the two halves of the test back together" with regard to a reliability estimate, we use a correction formula commonly referred to as the **Spearman-Brown formula** (or sometimes the Spearman-Brown prophecy formula since it prophesies the reliability coefficient of the full-length test). To estimate the reliability of the full test, the Spearman-Brown formula is generally applied as:

$$\text{Reliability of Full Test} = \frac{2 \times \text{Reliability of Half Test}}{1 + \text{Reliability of Half Test}}$$

Here is an example. Suppose the correlation between odd and even halves of your midterm in this course was 0.74, the calculation using the Spearman-Brown formula would go as follows:

$$\text{Reliability of Full Test} = \frac{2 \times 0.74}{1 + 0.74}$$

$$\text{Reliability of Full Test} = \frac{1.48}{1.74} = 0.85$$

The reliability coefficient of 0.85 estimates the reliability of the full test when the odd–even halves correlated at 0.74. This demonstrates that the uncorrected split-half reliability coefficient presents an underestimate of the reliability of the full test. Table 4.2 provides examples of half-test coefficients and the corresponding full-test coefficients that were corrected with

the Spearman-Brown formula. By looking at the first row in this table, you will see that a half-test correlation of 0.50 corresponds to a corrected full-test coefficient of 0.67.

Although the odd–even approach is the most common way to divide a test and will generally produce equivalent halves, certain situations deserve special attention. For example, if you have a test with a relatively small number of items (e.g., <8), it may be desirable to divide the test into equivalent halves based on a careful review of item characteristics such as content, format, and difficulty. Another situation that deserves special attention involves groups of items that deal with an integrated problem (this is referred to as a testlet). For example, if multiple questions refer to a specific diagram or reading passage, that whole set of questions should be included in the same half of the test. Splitting integrated problems can artificially inflate the reliability estimate (e.g., Sireci, Thissen, & Wainer, 1991).

An advantage of the split-half approach to reliability is that it can be calculated from a single administration of a test. Also, because only one testing session is involved, this approach reflects errors due only to content sampling and is not sensitive to time sampling errors.

Coefficient Alpha and Kuder-Richardson Reliability. Other approaches to estimating reliability from a single administration of a test are based on formulas developed by Kuder and Richardson (1937) and Cronbach (1951). Instead of comparing responses on two halves of the test as in split-half reliability, this approach examines the consistency of responding to all the individual items on the test. Reliability estimates produced with these formulas can be thought of as the average of all possible split-half coefficients. Like split-half reliability, these estimates are sensitive to measurement error introduced by content sampling. Additionally, they are also sensitive to the heterogeneity of the test content. When we refer to **content heterogeneity,** we are concerned with the degree to which the test items measure related characteristics. For example, our 25-item math test involving multiplying two-digit numbers would probably be more homogeneous than a test designed

***Coefficient alpha* and *Kuder-Richardson reliability* are sensitive to error introduced by content sampling, but also reflect the heterogeneity of test content.**

TABLE 4.2 Half-Test Coefficients and Corresponding Full-Test Coefficients Corrected with the Spearman-Brown Formula

Half-Test Correlation	Spearman-Brown Reliability
0.50	0.67
0.55	0.71
0.60	0.75
0.65	0.79
0.70	0.82
0.75	0.86
0.80	0.89
0.85	0.92
0.90	0.95
0.95	0.97

**If the sc
relies on
it is imp
degree o
differen
test. Thi
*rater reli***

to measure both multiplication and division. An even more heterogeneous test would be one that involves multiplication and reading comprehension, two fairly dissimilar content domains. As discussed later, sensitivity to content heterogeneity can influence a particular reliability formula's use on different domains.

While Kuder and Richardson's formulas and coefficient alpha both reflect item heterogeneity and errors due to content sampling, there is an important difference in terms of application. In their original article Kuder and Richardson (1937) presented numerous formulas for estimating reliability. The most commonly used formula is known as the **Kuder-Richardson formula 20** (KR-20). KR-20 is applicable when test items are scored dichotomously, that is, simply right or wrong, as 0 or 1. **Coefficient alpha** (Cronbach, 1951) is a more general form of KR-20 that also deals with test items that produce scores with multiple values (e.g., 0, 1, or 2). Because coefficient alpha is more broadly applicable, it has become the preferred statistic for estimating internal consistency (Keith & Reynolds, 1990). Tables 4.3 and 4.4 illustrate the calculation of KR-20 and coefficient alpha, respectively.

Inter-Rater Reliability

If the scoring of a test relies on subjective judgment, it is important to evaluate the degree of agreement when different individuals score the test. This is referred to as inter-scorer or **inter-rater reliability.** Estimating inter-rater reliability is a fairly straightforward process. The test is administered one time and two individuals independently score each test. A correlation is then calculated between the scores obtained by the two scorers. This estimate of reliability is not sensitive to error due to content or time sampling, but only reflects differences due to the individuals scoring the test. In addition to the correlational approach, inter-rater agreement can also be evaluated by calculating the percentage of times that two individuals assign the same scores to the performances of students. This approach is illustrated in Special Interest Topic 4.2.

If the scoring of an assessment relies on subjective judgment, it is important to evaluate the degree of agreement when different individuals score the test. This is referred to as *inter-rater reliability.*

On some tests, inter-rater reliability is of little concern. For example, on a test with multiple-choice or true–false items, grading is fairly straightforward and a conscientious grader should produce reliable and accurate scores. In the case of our 25-item math test, a careful grader should be able to determine whether the students' answers are accurate and assign a score consistent with that of another careful grader. However, for some tests inter-rater reliability is a major concern. Classroom essay tests are a classic example. It is common for students to feel that a different teacher might have assigned a different score to their essays. It can be argued that the teacher's personal biases, preferences, or mood influenced the score, not only the content and quality of the student's essay. Even on our 25-item math test, if the teacher required that the students "show their work" and this influenced the students' grades, subjective judgment might be involved and inter-rater reliability could be a concern.

TABLE 4.3 Calculating KR-20

KR-20 is sensitive to measurement error due to content sampling and is also a measure of item heterogeneity. KR-20 is applicable when test items are scored dichotomously, that is, simply right or wrong, as 0 or 1. The following formula is used for calculating KR-20:

$$\text{KR-20} = \frac{k}{k-1}\left(\frac{\text{SD}^2 - \sum p_i \times q_i}{\text{SD}^2}\right)$$

where
k = number of items
SD^2 = variance of total test scores
p_i = proportion of correct responses on item
q_i = proportion of incorrect responses on item

Consider these data for a five-item test administered to six students. Each item could receive a score of either 1 or 0.

	Item 1	Item 2	Item 3	Item 4	Item 5	Total Score
Student 1	1	0	1	1	1	4
Student 1	1	1	1	1	1	5
Student 3	1	0	1	0	0	2
Student 4	0	0	0	1	0	1
Student 5	1	1	1	1	1	5
Student 6	1	1	0	1	1	4
p_i	0.8333	0.5	0.6667	0.8333	0.6667	$\text{SD}^2 = 2.25$
q_i	0.1667	0.5	0.3333	0.1667	0.3333	
$p_i \times q_i$	0.1389	0.25	0.2222	0.1389	0.2222	

Note: When calculating SD^2, n was used in the denominator.

$$\sum p_i \times q_i = 0.1389 + 0.25 + 0.2222 + 0.1389 + 0.2222$$

$$\sum p_i \times q_i = 0.972$$

$$\text{KR-20} = 5/4\left(\frac{2.25 - 0.972}{2.25}\right)$$
$$= 1.25\left(\frac{1.278}{2.25}\right)$$
$$= 1.25(0.568)$$
$$= 0.71$$

Reliability of Composite Scores

Psychological and educational measurement often yields multiple scores that can be combined to form a composite. For example, the assignment of grades in educational settings is often based on a composite of several tests and other assessments administered over a grad-

TABLE 4.4 Calculating Coefficient Alpha

Coefficient alpha is sensitive to measurement error due to content sampling and is also a measure of item heterogeneity. It can be applied to tests with items that are scored dichotomously or that have multiple values. The formula for calculating coefficient alpha is:

$$\text{Coefficient alpha} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma SD_i^2}{SD^2}\right)$$

where k = number of items
SD_i^2 = variance of individual items
SD^2 = variance of total test scores

Consider these data for a five-item test that was administered to six students. Each item could receive a score ranging from 1 to 5.

	Item 1	Item 2	Item 3	Item 4	Item 5	Total Score
Student 1	4	3	4	5	5	21
Student 2	3	3	2	3	3	14
Student 3	2	3	2	2	1	10
Student 4	4	4	5	3	4	20
Student 5	2	3	4	2	3	14
Student 6	2	2	2	1	3	10
SD_i^2	0.8056	0.3333	1.4722	1.5556	1.4722	$SD^2 = 18.81$

Note: When calculating SD_i^2 and SD^2, n was used in the denominator.

$$\text{Coefficient Alpha} = \frac{5}{4}\left(1 - \frac{0.8056 + 0.3333 + 1.4722 + 1.5556 + 1.4722}{18.81}\right)$$
$$= 1.25(1 - 5.63889/18.81)$$
$$= 1.25(1 - 0.29978)$$
$$= 1.25(0.70)$$
$$= 0.875$$

Reliability of *composite scores* is generally greater than the measures that contribute to the composite.

ing period or semester. Many standardized psychological instruments contain several measures that are combined to form an overall **composite score.** For example, the *Wechsler Adult Intelligence Scale—Third Edition* (Wechsler, 1997) is composed of 11 subtests used in the calculation of the Full Scale Intelligence Quotient (FSIQ). Both of these situations involve composite scores obtained by combining the scores on several different tests or subtests. The advantage of composite scores is that the reliability of composites is generally greater than that of the individual scores that contribute to the composite. More precisely, the reliability of a composite is the result of the number of scores in the composite, the reliability of the individual scores, and the correlation between those scores. The more scores in the composite, the higher the correlation between those

SPECIAL INTEREST TOPIC 4.2

Calculating Inter-Rater Agreement

Performance assessments require test takers to complete a process or produce a product in a context that closely resembles real-life situations. For example, a student might engage in a debate, compose a poem, or perform a piece of music. The evaluation of these types of performances is typically based on scoring rubrics that specify what aspects of the student's performance should be considered when providing a score or grade. The scoring of these types of assessments obviously involves the subjective judgment of the individual scoring the performance, and as a result inter-rater reliability is a concern. As noted in the text one approach to estimating inter-rater reliability is to calculate the correlation between the scores that are assigned by two judges. Another approach is to calculate the percentage of agreement between the judges' scores.

Consider an example wherein two judges rated poems composed by 25 students. The poems were scored from 1 to 5 based on criteria specified in a rubric, with 1 being the lowest performance and 5 being the highest. The results are illustrated in the following table:

	Ratings of Rater 1				
Ratings of Rater 2	*1*	*2*	*3*	*4*	*5*
5	0	0	1	2	4
4	0	0	2	3	2
3	0	2	3	1	0
2	1	1	1	0	0
1	1	1	0	0	0

Once the data are recorded you can calculate inter-rater agreement with the following formula:

$$\text{Inter-Rater Agreement} = \frac{\text{Number of Cases Assigned the Same Scores}}{\text{Total Number of Cases}} \times 100$$

In our example the calculation would be:

Inter-Rater Agreement = 12/25 × 100

Inter-Rater Agreement = 48%

This degree of inter-rater agreement might appear low to you, but this would actually be respectable for a classroom test. In fact the Pearson correlation between these judges' ratings is 0.80 (better than many, if not most, performance assessments).

Instead of requiring the judges to assign the exact same score for agreement, some authors suggest the less rigorous criterion of scores being within one point of each other (e.g., Linn & Gronlund, 2000). If this criterion were applied to these data, the modified agreement percent would be 96% because only one of the judges' scores were not within one point of each other (Rater 1 assigned a 3 and Rater 2 a 5).

We caution you not to expect this high a rate of agreement should you examine the inter-rater agreement of your own performance assessments. In fact you will learn later that difficulty scoring performance assessments in a reliable manner is one of the major limitations of these procedures.

scores, and the higher the individual reliabilities, the higher the composite reliability. As we noted, tests are simply samples of the test domain, and combining multiple measures is analogous to increasing the number of observations or the sample size.

Selecting a Reliability Coefficient

Table 4.5 summarizes the sources of measurement error reflected in different reliability coefficients. As we have suggested in our discussion of each approach to estimating reliability, different conditions call for different estimates of reliability. One should consider factors such as the nature of the construct and how the scores will be used when selecting an estimate of reliability. If a test is designed to be given more than one time to the same individuals, test-retest and alternate-form reliability with delayed administration are appropriate because they are sensitive to measurement errors resulting from time sampling. Accordingly, if a test is used to predict an individual's performance on a criterion in the future, it is also important to use a reliability estimate that reflects errors due to time sampling.

One should consider factors such as the nature of the construct being measured and how the scores will be used when selecting an estimate of reliability.

When a test is designed to be administered only one time, an estimate of internal consistency is appropriate. As we noted, split-half reliability estimates error variance resulting from content sampling whereas coefficient alpha and KR-20 estimate error variance due to content sampling and content heterogeneity. Because KR-20 and coefficient alpha are sensitive to content heterogeneity, they are applicable when the test measures a homogeneous domain of knowledge or a unitary characteristic. For example, our 25-item test measuring the ability to multiply double digits reflects a homogeneous domain and coefficient alpha would provide a good estimate of reliability. However, if we have a 50-item test, 25 measuring multiplication with double digits and 25 measuring division, the domain is more heterogeneous and coefficient alpha

TABLE 4.5 Sources of Error Variance Associated with the Major Types of Reliability

Type of Reliability	Error Variance
Test-retest reliability	Time sampling
Alternate-form reliability	
Simultaneous administration	Content sampling
Delayed administration	Time sampling and content sampling
Split-half reliability	Content sampling
Coefficient alpha and KR-20	Content sampling and item heterogeneity
Inter-rater reliability	Differences due to raters/scorers

and KR-20 would probably underestimate reliability. In the situation of a test with heterogeneous content (the heterogeneity is intended and not a mistake), the split-half method is preferred. Because the goal of the split-half approach is to compare two equivalent halves, it would be necessary to ensure that each half has equal numbers of both multiplication and division problems.

We have been focusing on tests of achievement when providing examples, but the same principles apply to other types of tests. For example, a test that measures depressed mood may assess a fairly homogeneous domain, making the use of coefficient alpha or KR-20 appropriate. However, if the test measures depression, anxiety, anger, and impulsiveness, the content becomes more heterogeneous and the split-half estimate would be indicated. In this situation, the split-half approach would allow the construction of two equivalent halves with equal numbers of items reflecting the different traits or characteristics under investigation.

Naturally, if different forms of a test are available, it would be important to estimate alternate-form reliability. If a test involves subjective judgment by the person scoring the test, inter-rater reliability is important. Many contemporary test manuals report multiple estimates of reliability. Given enough information about reliability, one can partition the error variance into its components, as demonstrated in Figure 4.1.

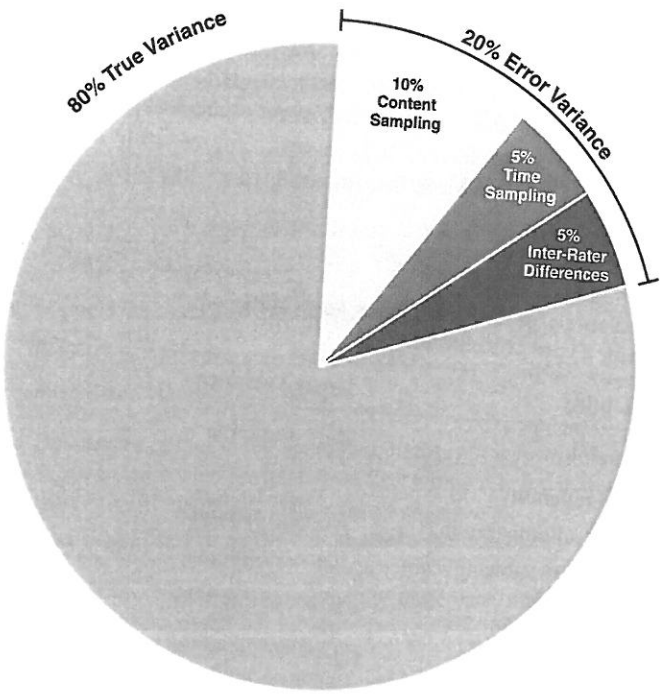

FIGURE 4.1 Partitioning the Variance to Reflect Sources of Variance

Evaluating Reliability Coefficients

Another important question that arises when considering reliability coefficients is "How large do reliability coefficients need to be?" Remember, we said reliability coefficients can be interpreted in terms of the proportion of test score variance attributable to true variance. Ideally we would like our reliability coefficients to equal 1.0 because this would indicate that 100% of the test score variance is due to true differences between individuals. However, due to measurement error, perfectly reliable measurement does not exist. There is not a single, simple answer to our question about what is an acceptable level of reliability. What constitutes an acceptable reliability coefficient depends on several factors, including the construct being measured, the amount of time available for testing, the way the scores will be used, and the method of estimating reliability. We will now briefly address each of these factors.

What constitutes an acceptable reliability coefficient depends on several factors, including the construct being measured, the amount of time available for testing, the way the scores will be used, and the method of estimating reliability.

Construct. Some constructs are more difficult to measure than others simply because the item domain is more difficult to sample adequately. As a general rule, personality variables are more difficult to measure than academic knowledge. As a result, what might be an acceptable level of reliability for a measure of "dependency" might be regarded as unacceptable for a measure of reading comprehension. In evaluating the acceptability of a reliability coefficient one should consider the nature of the variable under investigation and how difficult it is to measure. By carefully reviewing and comparing the reliability estimates of different instruments available for measuring a construct, one can determine which is the most reliable measure of the construct.

Time Available for Testing. If the amount of time available for testing is limited, only a limited number of items can be administered and the sampling of the test domain is open to greater error. This could occur in a research project in which the school principal allows you to conduct a study in his or her school but allows only 20 minutes to measure all the variables in your study. As another example, consider a districtwide screening for reading problems wherein the budget allows only 15 minutes of testing per student. In contrast, a psychologist may have two hours to administer a standardized intelligence test individually. It would be unreasonable to expect the same level of reliability from these significantly different measurement processes. However, comparing the reliability coefficients associated with instruments that can be administered within the parameters of the testing situation can help one select the best instrument for the situation.

Test Score Use. The way the test scores will be used is another major consideration when evaluating the adequacy of reliability coefficients. Diagnostic tests that form the basis for major decisions about individuals should be held to a higher standard than tests used with group research or for screening large numbers of individuals. For example, an individually administered test of intelligence that is used in the diagnosis of mental retardation would be expected to produce scores with a very high level of reliability. In

this context, performance on the intelligence test provides critical information used to determine whether the individual meets the diagnostic criteria. In contrast, a brief test used to screen all students in a school district for reading problems would be held to less rigorous standards. In this situation, the instrument is used simply for screening purposes and no decisions are being made that cannot easily be reversed. It helps to remember that although high reliability is desirable with all assessments, standards of acceptability vary according to the way test scores will be used. High-stakes decisions demand highly reliable information!

Method of Estimating Reliability. The size of reliability coefficients is also related to the method selected to estimate reliability. Some methods tend to produce higher estimates than other methods. As a result, it is important to take into consideration the method used to produce correlation coefficients when evaluating and comparing the reliability of different tests. For example, KR-20 and coefficient alpha typically produce reliability estimates that are smaller than ones obtained using the split-half method. As indicated in Table 4.5, alternate-form reliability with delayed administration takes into account more sources of error than other methods do and generally produces lower reliability coefficients. In summary, some methods of estimating reliability are more rigorous and tend to produce smaller coefficients, and this variability should be considered when evaluating reliability coefficients.

General Guidelines. Although it is apparent that many factors deserve consideration when evaluating reliability coefficients, we will provide some general guidelines that can provide some guidance.

If a test is being used to make important decisions that are likely to impact individuals significantly and are not easily reversed, it is reasonable to expect reliability coefficients of 0.90 or even 0.95.

- If a test is being used to make important decisions that are likely to significantly impact individuals and are not easily reversed, it is reasonable to expect reliability coefficients of 0.90 or even 0.95. This level of reliability is regularly obtained with individually administered tests of intelligence. For example, the reliability of the Wechsler Adult Intelligence Scale—Third Edition (Wechsler, 1997), an individually administered intelligence test, is 0.98.

- Reliability estimates of 0.80 or more are considered acceptable in many testing situations and are commonly reported for group and individually administered achievement and personality tests. For example, the *California Achievement Test/5* (CAT/5)(CTB/Macmillan/McGraw-Hill, 1993), a set of group-administered achievement tests frequently used in public schools, has reliability coefficients that exceed 0.80 for most of its subtests.

- For teacher-made classroom tests and tests used for screening, reliability estimates of at least 0.70 are expected. Classroom tests are frequently combined to form linear composites that determine a final grade, and the reliability of these composites is expected to be greater than the reliabilities of the individual tests. Marginal coefficients in the 0.70s might also be acceptable when more thorough assessment procedures are available to address concerns about individual cases.

Some writers suggest that reliability coefficients as low as 0.60 are acceptable for group research, performance assessments, and projective measures, but we are reluctant to endorse the use of any assessment that produces scores with reliability estimates below 0.70. As you recall, a reliability coefficient of 0.60 indicates that 40% of the observed variance can be attributed to random error. How much confidence can you place in assessment results when you know that 40% of the variance is attributed to random error?

The preceding guidelines on reliability coefficients and qualitative judgments of their magnitude must also be considered in context. Some constructs are just a great deal more difficult to measure reliably than others. From a developmental perspective, we know that emerging skills or behavioral attributes in children are more difficult to measure than mature or developed skills. When a construct is very difficult to measure, any reliability coefficient greater than 0.50 may well be acceptable just because there is still more true score variance present in such values relative to error variance. However, before choosing measures with reliability coefficients below 0.70, be sure no better measuring instruments are available that are also practical and whose interpretations have validity evidence associated with the intended purposes of the test.

How to Improve Reliability

Possibly the most obvious way to improve the reliability of measurement is simply to increase the number of items on a test. If we increase the number of items while maintaining the same quality as the original items, we will increase the reliability of the test.

A natural question at this point is "What can we do to improve the reliability of our assessment results?" In essence we are asking what steps can be taken to maximize true score variance and minimize error variance. Probably the most obvious approach is simply to increase the number of items on a test. In the context of an individual test, if we increase the number of items while maintaining the same quality as the original items, we will increase the reliability of the test. This concept was introduced when we discussed split-half reliability and presented the Spearman-Brown formula. In fact, a variation of the Spearman-Brown formula can be used to predict the effects on reliability achieved by adding items:

$$r = \frac{n \times r_{xx}}{1 + (n - 1)r_{xx}}$$

where r = estimated reliability on test with new items
n = factor by which the test length is increased
r_{xx} = reliability of the original test

For instance, consider the example of our 25-item math test. If the reliability of the test were 0.80 and we wanted to estimate the increase in reliability we would achieve by increasing the test to 30 items (a factor of 1.2), the formula would be:

$$r = \frac{1.2 \times 0.80}{1 + [(1.2 - 1)0.80]}$$

$$r = \frac{0.96}{1.16}$$

$$r = 0.83$$

Table 4.6 provides other examples illustrating the effects of increasing the length of our hypothetical test on reliability. By looking in the first row of this table you see that increasing the number of items on a test with a reliability of 0.50 by a factor of 1.25 results in a predicted reliability of 0.56. Increasing the number of items by a factor of 2.0 (i.e., doubling the length of the test) increases the reliability to 0.67.

In some situations various factors will limit the number of items we can include in a test. For example, teachers generally develop tests that can be administered in a specific time interval, usually the time allocated for a class period. In these situations, one can enhance reliability by using multiple measurements that are combined for an average or composite score. As noted earlier, combining multiple tests in a linear composite will increase the reliability of measurement over that of the component tests. In summary, anything we do to get a more adequate sampling of the content domain will increase the reliability of our measurement.

In Chapter 6 we will discuss a set of procedures collectively referred to as "item analyses." These procedures help us select, develop, and retain test items with good measurement characteristics. While it is premature to discuss these procedures in detail, it should be noted that selecting or developing good items is an important step in developing a good test. Selecting and developing good items will enhance the measurement characteristics of the assessments you use.

Another way to reduce the effects of measurement error is what Ghiselli, Campbell, and Zedeck (1981) refer to as "good housekeeping procedures." By this they mean test developers should provide precise and clearly stated procedures regarding the administra-

TABLE 4.6 Reliability Expected when Increasing the Number of Items

Current Reliability	The Reliability Expected when the Number of Items Is Increased By:			
	× *1.25*	× *1.50*	× *2.0*	× *2.5*
0.50	0.56	0.60	0.67	0.71
0.55	0.60	0.65	0.71	0.75
0.60	0.65	0.69	0.75	0.79
0.65	0.70	0.74	0.79	0.82
0.70	0.74	0.78	0.82	0.85
0.75	0.79	0.82	0.86	0.88
0.80	0.83	0.86	0.89	0.91
0.85	0.88	0.89	0.92	0.93
0.90	0.92	0.93	0.95	0.96

When est
of the res
estimates
administi
appropri

tion and scoring of tests. Examples include providing explicit instructions for standardized administration, developing high-quality rubrics to facilitate reliable scoring, and requiring extensive training before individuals can administer, grade, or interpret a test.

Special Problems in Estimating Reliability

Reliability of Speed Tests. A speed test generally contains items that are relatively easy but has a time limit that prevents any test takers from correctly answering all questions. As a result, the test taker's score on a speed test primarily reflects the speed of performance. When estimating the reliability of the results of speed tests, estimates derived from a single administration of a test are not appropriate. Therefore, with speed tests, test-retest or alternate-form reliability is appropriate, but split-half, coefficient alpha, and KR-20 should be avoided.

When estimating the reliability of the results of speed tests, estimates derived from a single administration of a test are not appropriate.

Reliability as a Function of Score Level. Though it is desirable, tests do not always measure with the same degree of precision throughout the full range of scores. If a group of individuals is tested for whom the test is either too easy or too difficult, we are likely to have additional error introduced into the scores. At the extremes of the distribution, at which scores reflect either all correct or all wrong responses, little accurate measurement has occurred. It would be inaccurate to infer that a child who missed every question on an intelligence test has "no" intelligence. Rather, the test did not adequately assess the low-level skills necessary to measure the child's intelligence. This is referred to as the test having an insufficient "floor." At the other end, it would be inaccurate to report that a child who answers all of the questions on an intelligence test correctly has an "infinite level of intelligence." The test is simply too easy to provide an adequate measurement, a situation referred to as a test having an insufficient "ceiling." In both cases we need a more appropriate test. Generally, aptitude and achievement tests are designed for use with individuals of certain ability levels. When a test is used with individuals who fall either at the extremes or outside this range, the scores might not be as accurate as the reliability estimates suggest. In these situations, further study of the reliability of scores at this level is indicated.

Range Restriction. The values we obtain when calculating reliability coefficients are dependent on characteristics of the sample or group of individuals on which the analyses are based. One characteristic of the sample that significantly impacts the coefficients is the degree of variability in performance (i.e., variance). More precisely, reliability coefficients based on samples with large variances (referred to as heterogeneous samples) will generally produce higher estimates of reliability than those based on samples with small variances (referred to as homogeneous samples). When reliability coefficients are based on a sample with a restricted range of variability, the coefficients may actually underestimate the reliability of measurement. For example, if you base a reliability analysis on students in a gifted and talented class in which practically all of the scores reflect exemplary performance (e.g., >90% correct), you will receive lower estimates of reliability than if the analyses are based on a class with a broader and more nearly normal distribution of scores.

The reliability estimates discussed in this chapter are usually not applicable to scores of mastery tests. Because mastery tests emphasize classification, a recommended approach is to use an index that reflects the consistency of classification.

Mastery Testing. Criterion-referenced tests are used to make interpretations relative to a specific level of performance. Mastery testing is an example of a criterion-referenced test by which a test taker's performance is evaluated in terms of achieving a cut score instead of the degree of achievement. The emphasis in this testing situation is on classification. Either test takers score at or above the cut score and are classified as having mastered the skill or domain, or they score below the cut score and are classified as having not mastered the skill or domain. Mastery testing often results in limited variability among test takers, and, as we just described, limited variability in performance results in small reliability coefficients. As a result, the reliability estimates discussed in this chapter are typically inadequate for assessing the reliability of mastery test scores. Given the emphasis on classification, a recommended approach is to use an index that reflects the consistency of classification (AERA et al., 1999). Special Interest Topic 4.3 illustrates a useful procedure for evaluating the consistency of classification when using mastery tests.

Reliability coefficients are useful when comparing the reliability of the scores produced by different tests, but when the focus is on interpreting the test scores of individuals, the *standard error of measurement* is a more practical statistic.

The Standard Error of Measurement

Reliability coefficients are interpreted in terms of the proportion of observed variance attributable to true variance and are a useful way of comparing the reliability of scores produced by different assessment procedures. Other things being equal, you will want to select the test that produces scores with the best reliability. However, once a test has been selected and the focus is on interpreting scores, the **standard error of measurement (SEM)** is a more practical statistic. The SEM is the standard deviation of the distribution of scores that would be obtained by one person if he or she were tested on an infinite number of parallel forms of a test comprised of items randomly sampled from the same content domain. In other words, if we created an infinite number of parallel forms of a test and had the same person take them with no carryover effects, the presence of measurement error would prevent the person from earning the same score every time. Although each test might represent the content domain equally well, the test taker would perform better on some tests and worse on others simply due to random error. By taking the scores obtained on all of these tests, a distribution of scores would result. The mean of this distribution is the individual's true score (T) and the SEM is the standard deviation of this distribution of error score. Obviously, we are never actually able to follow these procedures and must estimate the SEM using information available to us.

Evaluating the Standard Error of Measurement

The SEM is a function of the reliability (r_{xx}) and standard deviation (SD) of a test. When calculating the SEM, the reliability coefficient takes into consideration measurement errors

SPECIAL INTEREST TOPIC 4.3

Consistency of Classification with Mastery Tests

As noted in the text, the size of reliability coefficients is substantially affected by the variance of the test scores. Limited test score variance results in lower reliability coefficients. Because mastery tests often do not produce test scores with much variability, the methods of estimating reliability described in this chapter will often underestimate the reliability of these tests. To address this, reliability analyses of mastery tests typically focus on the consistency of classification. That is, because the objective of mastery tests is to determine if a student has mastered the skill or knowledge domain, the question of reliability can be framed as one of how consistent mastery–nonmastery classifications are. For example, if two parallel or equivalent mastery tests covering the same skill or content domain consistently produce the same classifications for the same test takers (i.e., mastery versus nonmastery), we would have evidence of consistency of classification. If two parallel mastery tests produced divergent classifications, we would have cause for concern. In this case the test results are not consistent or reliable.

The procedure for examining the consistency of classification on parallel mastery tests is fairly straightforward. Simply administer both tests to a group of students and complete a table like the one that follows. For example, consider two mathematics mastery tests designed to assess students' ability to multiply fractions. The cut score is set at 80%, so all students scoring 80% or higher are classified as having mastered the skill while those scoring less than 80% are classified as not having mastered the skill. In the following example, data are provided for 50 students:

	Form B: Nonmastery (score <80%)	**Form B: Mastery (score of 80% or better)**
Form A: Mastery (score of 80% or better)	4	32
Form A: Nonmastery (score <80%)	11	3

Students classified as achieving mastery on both tests are denoted in the upper right-hand cell while students classified as not having mastered the skill are denoted in the lower left-hand cell. There were four students who were classified as having mastered the skills on Form A but not on Form B (denoted in the upper left-hand cell). There were three students who were classified as having mastered the skills on Form B but not on Form A (denoted in the lower right-hand cell). The next step is to calculate the percentage of consistency by using the following formula:

$$\text{Percent Consistency} = \frac{\text{Mastery on Both Forms} + \text{Nonmastery on Both Forms}}{\text{Total Number of Students}} \times 100$$

$$\text{Percent Consistency} = \frac{32 + 11}{50} \times 100$$

$$\text{Percent Consistency} = 0.86 \times 100$$

$$\text{Percent Consistency} = 86\%$$

(continued)

SPECIAL INTEREST TOPIC 4.3 Continued

This approach is limited to situations in which you have parallel mastery tests. Another limitation is that there are no clear standards regarding what constitutes "acceptable" consistency of classification. As with the evaluation of all reliability information, the evaluation of classification consistency should take into consideration the consequences of any decisions that are based on the test results (e.g., Gronlund, 2003). If the test results are used to make high-stakes decisions (e.g., awarding a diploma), a very high level of consistency is required. If the test is used only for low-stake decisions (e.g., failure results in further instruction and retesting), a lower level of consistency may be acceptable. Subkoviak (1984) provides a good discussion of several techniques for estimating the classification consistency of mastery tests, including some rather sophisticated approaches that require only a single administration of the test.

The greater the reliability of a test score, the smaller the SEM and the more confidence we have in the precision of test scores.

present in test scores, and the SD reflects the variability of the scores in the distribution. The SEM is estimated using the following formula:

$$SEM = SD\sqrt{1 - r_{xx}}$$

where SD = the standard deviation of the obtained scores
r_{xx} = the reliability of the test

Let's work through two quick examples. First, let's assume a test with a standard deviation of 10 and reliability of 0.90.

Example 1: $SEM = \sqrt{1 - 0.90}$

$SEM = \sqrt{0.10}$

$SEM = 3.2$

Now let's assume a test with a standard deviation of 10 and reliability of 0.80. The SD is the same as in the previous example, but the reliability is lower.

Example 2: $SEM = \sqrt{1 - 0.80}$

$SEM = \sqrt{0.20}$

$SEM = 4.5$

Notice that as the reliability of the test scores decreases, the SEM increases. Because the reliability coefficient reflects the proportion of observed score variance due to true score variance and the SEM is an estimate of the amount of error in test scores, this inverse relationship is what one would expect. The greater the reliability of test scores, the smaller the SEM and the more confidence we have in the precision of test scores. The lower the reliability of a test, the larger the SEM and the less confidence we have in the precision of test scores. Table 4.7 shows the SEM as a function of SD and reliability. Examining the

A *confidence* range of scor the individu a prescribed et al., 1999).

TABLE 4.7 Standard Errors of Measurement for Values of Reliability and Standard Deviation

Standard Deviation	Reliability Coefficients					
	0.95	*0.90*	*0.85*	*0.80*	*0.75*	*.70*
30	6.7	9.5	11.6	13.4	15.0	16.4
28	6.3	8.9	10.8	12.5	14.0	15.3
26	5.8	8.2	10.1	11.6	13.0	14.2
24	5.4	7.6	9.3	10.7	12.0	13.1
22	4.9	7.0	8.5	9.8	11.0	12.0
20	4.5	6.3	7.7	8.9	10.0	11.0
18	4.0	5.7	7.0	8.0	9.0	9.9
16	3.6	5.1	6.2	7.2	8.0	8.8
14	3.1	4.4	5.4	6.3	7.0	7.7
12	2.7	3.8	4.6	5.4	6.0	6.6
10	2.2	3.2	3.9	4.5	5.0	5.5
8	1.8	2.5	3.1	3.6	4.0	4.4
6	1.3	1.9	2.3	2.7	3.0	3.3
4	0.9	1.3	1.5	1.8	2.0	2.2
2	0.4	0.6	0.8	0.9	1.0	1.1

first row in the table shows that on a test with a standard deviation of 30 and a reliability coefficient of 0.95 the SEM is 6.7. In comparison, if the reliability of the test score is 0.90 the SEM is 9.5; if the reliability of the test is 0.85 the SEM is 11.6; and so forth. The SEM is used in calculating intervals or bands around observed scores in which the true score is expected to fall. We will now turn to this application of the SEM.

A *confidence interval* reflects a range of scores that will contain the individual's true score with a prescribed probability (AERA et al., 1999).

Calculating Confidence Intervals. A **confidence interval** reflects a range of scores that will contain the individual's true score with a prescribed probability (AERA et al., 1999). We use the SEM to calculate confidence intervals. When introducing the SEM, we said it provides information about the distribution of observed scores around true scores. More precisely, we defined the SEM as the standard deviation of the distribution of error scores. Like any standard deviation, the SEM can be interpreted in terms of frequencies represented in a normal distribution. In the previous chapter we showed that approximately 68% of the scores in a normal distribution are located between one SD below the mean and one SD above the mean. As a result, approximately 68% of the time an individual's observed score would be expected to be within ±1 SEM of the true score. For example, if an individual had a true score of 70 on a test with a SEM of 3,

then we would expect him or her to obtain scores between 67 and 73 two-thirds of the time. To obtain a 95% confidence interval we simply determine the number of standard deviations encompassing 95% of the scores in a distribution. By referring to a table representing areas under the normal curve (see Appendix F), you can determine that 95% of the scores in a normal distribution fall within ±1.96 of the mean. Given a true score of 70 and SEM of 3, the 95% confidence interval would be 70 ± 3(1.96) or 70 ± 5.88. Therefore, in this situation an individual's observed score would be expected to be between 64.12 and 75.88 95% of the time.

You might have noticed a potential problem with this approach to calculating confidence intervals. So far we have described how the SEM allows us to form confidence intervals around the test taker's true score. The problem is that we don't know a test taker's true score, only the observed score. Although it is possible for us to estimate true scores (see Nunnally & Bernstein, 1994), it is common practice to use the SEM to establish confidence intervals around obtained scores (see Gulliksen, 1950). These confidence intervals are calculated in the same manner as just described, but the interpretation is slightly different. In this context the confidence interval is used to define the range of scores that will contain the individual's true score. For example, if an individual obtains a score of 70 on a test with a SEM of 3.0, we would expect his or her true score to be between 67 and 73 (obtained score ±1 SEM) 68% of the time. Accordingly, we would expect his or her true score to be between 64.12 and 75.88 95% of the time (obtained score ±1.96 SEM).

It may help to make note of the relationship between the reliability of the test score, the SEM, and confidence intervals. Remember that we noted that as the reliability of scores increases the SEM decreases. The same relationship exists between test reliability and confidence intervals. As the reliability of test scores increases (denoting less measurement error), the confidence intervals become smaller (denoting more precision in measurement).

A major advantage of the SEM and the use of confidence intervals is that they serve to remind us that measurement error is present in all scores and that we should interpret scores cautiously. A single numerical score is often interpreted as if it is precise and involves no error. For example, if you report that Susie has a Full Scale IQ of 113, her parents might interpret this as implying that Susie's IQ is exactly 113. If you are using a high-quality IQ test such as the *Wechsler Intelligence Scale for Children—4th Edition* or the *Reynolds Intellectual Assessment Scales,* the obtained IQ is very likely a good estimate of her true IQ. However, even with the best assessment instruments the obtained scores contain some degree of error and the SEM and confidence intervals help us illustrate this. This information can be reported in different ways in written reports. For example, Kaufman and Lichtenberger (1999) recommend the following format:

A major advantage of the SEM and the use of confidence intervals is that they serve to remind us that measurement error is present in all scores and that we should interpret scores cautiously.

> Susie obtained a Full Scale IQ of 113 (between 108 and 118 with 95% confidence).

Kamphaus (2001) recommends a slightly different format:

> Susie obtained a Full Scale IQ in the High Average range, with a 95% probability that her true IQ falls between 108 and 118.

Regardless of the exact format used, the inclusion of confidence intervals highlights the fact that test scores contain some degree of measurement error and should be interpreted with caution. Most professional test publishers either report scores as bands within which the test taker's true score is likely to fall or provide information on calculating these confidence intervals.

Reliability: Practical Strategies for Teachers

Now that you are aware of the importance of the reliability of measurement, a natural question is "How can I estimate the reliability of scores on my classroom tests?" Most teachers have a number of options. First, if you use multiple-choice or other tests that can be scored by a computer scoring program, the score printout will typically report some reliability estimate (e.g., coefficient alpha or KR-20). If you do not have access to computer scoring, but the items on a test are of approximately equal difficulty and scored dichotomously (i.e., correct/incorrect), you can use an internal consistency reliability estimate known as the Kuder-Richardson formula 21 (KR-21). This formula is actually an estimate of the KR-20 discussed earlier and is usually adequate for classroom tests. To calculate KR-21 you need to know only the mean, variance, and number of items on the test:

Most teachers have multiple options for estimating the reliability of scores produced by their classroom tests.

$$\text{KR-21} = 1 - \frac{X(n - X)}{n\sigma^2}$$

where X = mean
σ^2 = variance
n = number of items

Consider the following set of 20 scores: 50, 48, 47, 46, 42, 42, 41, 40, 40, 38, 37, 36, 36, 35, 34, 32, 32, 31, 30, and 28. Here the $X = 38.25$, $\sigma^2 = 39.8$, and $n = 50$. Therefore,

$$\text{KR-21} = 1 - \frac{38.25(50 - 38.25)}{50(39.8)}$$

$$= 1 - \frac{449.4375}{1990}$$

$$= 1 - 0.23 = 0.77$$

As you see, this is a fairly simple procedure. If you have access to a computer with a spreadsheet program or a calculator with mean and variance functions, you can estimate the reliability of a classroom test easily in a matter of minutes with this formula.

Special Interest Topic 4.4 presents a shortcut approach for calculating the Kuder-Richardson formula 21 (KR-21). If you want to avoid even these limited computations, we prepared Table 4.8, which allows you to estimate the KR-21 reliability for dichotomously

SPECIAL INTEREST TOPIC 4.4

A Quick Way to Estimate Reliability for Classroom Exams

Saupe (1961) provided a quick method for teachers to calculate reliability for a classroom exam in the era prior to easy access to calculators or computers. It is appropriate for a test in which each item is given equal weight and each item is scored either right or wrong. First, the standard deviation of the exam must be estimated from a simple approximation:

SD = [sum of top 1/6th of scores – sum of bottom 1/6th of scores] / [total # of scores – 1] / 2

Then reliability can be estimated from:

Reliability = 1 – [0.19 × number of items] / SD^2

Thus, for example, in a class with 24 student test scores, the top one-sixth of the scores are 98, 92, 87, and 86, while the bottom sixth of the scores are 48, 72, 74, and 75. With 25 test items, the calculations are:

SD = [98 + 92 + 87 + 86 – 48 – 72 – 74 – 75] / 23/2
= [363 – 269] / 11.5
= 94 / 11.5 = 8.17

So,

Reliability = 1 – [0.19 × 25] / 8.17^2
= 1 – 0.07
= 0.93

A reliability coefficient of 0.93 for a classroom test is excellent! Don't be dismayed if your classroom tests do not achieve this high a level of reliability.

Source: Saupe, J. L. (1961). Some useful estimates of the Kuder-Richardson formula number 20 reliability coefficient. *Educational and Psychological Measurement, 2,* 63–72.

TABLE 4.8 KR-21 Reliability Estimates for Tests with a Mean of 80%

	Standard Deviation of Test		
Number of Items (*n*)	*0.10*(n)	*0.15*(n)	*0.20*(n)
10	—	0.29	0.60
20	0.20	0.64	0.80
30	0.47	0.76	0.87
40	0.60	0.82	0.90
50	0.68	0.86	0.92
75	0.79	0.91	0.95
100	0.84	0.93	0.96

Sum

scored classroom tests if you know the standard deviation and number of items (this table was modeled after tables originally presented by Deiderich, 1973). This table is appropriate for tests with a mean of approximately 80% correct (we are using a mean of 80% correct because it is fairly representative of many classroom tests). To illustrate its application, consider the following example. If your test has 50 items and an SD of 8, select the "Number of Items" row for 50 items and the "Standard Deviation" column for $0.15n$, because $0.15(50) = 7.5$, which is close to your actual SD of 8. The number at the intersection is 0.86, which is a very respectable reliability for a classroom test (or a professionally developed test for that matter).

If you examine Table 4.8, you will likely detect a few fairly obvious trends. First, the more items on the test the higher the estimated reliability coefficients. We alluded to the beneficial impact of increasing test length previously in this chapter and the increase in reliability is due to enhanced sampling of the content domain. Second, tests with larger standard deviations (i.e., variance) produce more reliable results. For example, a 30-item test with an SD of 3—i.e., $0.10(n)$—results in an estimated reliability of 0.47, while one with an SD of 4.5—i.e., $0.15(n)$—results in an estimated reliability of 0.76. This reflects the tendency we described earlier that restricted score variance results in smaller reliability coefficients. We should note that while we include a column for standard deviations of $0.20(n)$, standard deviations this large are rare with classroom tests (Deiderich, 1973). In fact, from our experience it is more common for classroom tests to have standard deviations closer to $0.10(n)$. Before leaving our discussion of KR-21 and its application to classroom tests, we do want to caution you that KR-21 is only an approximation of KR-20 or coefficient alpha. KR-21 assumes the test items are of equal difficulty and it is usually slightly lower than KR-20 or coefficient alpha (Hopkins, 1998). Nevertheless, if the assumptions are not grossly violated, it is probably a reasonably good estimate of reliability for many classroom applications.

Our discussion of shortcut reliability estimates to this point has been limited to tests that are dichotomously scored. Obviously, many of the assessments teachers use are not dichotomously scored and this makes the situation a little more complicated. If your items are not scored dichotomously, you can calculate coefficient alpha with relative ease using a commonly available spreadsheet such as Microsoft Excel. With a little effort you should be able to use a spreadsheet to perform the computations illustrated previously in Tables 4.3 and 4.4.

Summary

Reliability refers to consistency in test scores. If a test or other assessment procedure produces consistent measurements, its scores are reliable. Why is reliability so important? As we have emphasized, assessments are useful because they provide information that helps educators make better decisions. However, the reliability (and validity) of that information is of paramount importance. For us to make good decisions, we need reliable information. By estimating the reliability of our assessment results, we get an indication of how much confidence we can place in them. If we have highly reliable and valid information, it is probable that we can use that information to make better decisions. If the results are unreliable, they are of little value to us.

Errors of measurement undermine the reliability of measurement and therefore reduce the utility of the measurement. Although there are multiple sources of measurement error, the major ones are content sampling and time sampling errors. Content sampling errors are the result of less than perfect sampling of the content domain. The more representative tests are of the content domain, the less content sampling errors threaten the reliability of the test. Time sampling errors are the result of random changes in the test taker or environment over time. Experts in testing and measurement have developed methods of estimating errors due to these and other sources, including the following major approaches to estimating reliability:

- *Test-retest reliability* involves the administration of the same test to a group of individuals on two different occasions. The correlation between the two sets of scores is the test-retest reliability coefficient and reflects errors due to time sampling.
- *Alternate-form reliability* involves the administration of parallel forms of a test to a group of individuals. The correlation between the scores on the two forms is the reliability coefficient. If the two forms are administered at the same time, the reliability coefficient reflects only content sampling error. If the two forms of the test are administered at different times, the reliability coefficient reflects both content and time sampling errors.
- *Internal-consistency reliability* estimates are derived from a single administration of a test. Split-half reliability involves dividing the test into two equivalent halves and calculating the correlation between the two halves. Instead of comparing performance on two halves of the test, coefficient alpha and the Kuder-Richardson approaches examine the consistency of responding among all of the individual items of the test. Split-half reliability reflects errors due to content sampling whereas coefficient alpha and the Kuder-Richardson approaches reflect both item heterogeneity and errors due to content sampling.
- *Inter-rater reliability* is estimated by administering the test once but having the responses scored by different examiners. By comparing the scores assigned by different examiners, one can determine the influence of different raters or scorers. Inter-rater reliability is important to examine when scoring involves considerable subjective judgment.

We also discussed a number of issues important for understanding and interpreting reliability estimates. We provided some guidelines for selecting the type of reliability estimate most appropriate for specific assessment procedures, some guidelines for evaluating reliability coefficients, and some suggestions on improving the reliability of measurement.

Although reliability coefficients are useful when comparing the reliability of different tests, the standard error of measurement (SEM) is more useful when interpreting scores. The SEM is an index of the amount of error in test scores and is used in calculating confidence intervals within which we expect the true score to fall. An advantage of the SEM and the use of confidence intervals is that they serve to remind us that measurement error is present in all scores and that we should use caution when interpreting scores. We closed the chapter by illustrating some shortcut procedures that teachers can use to estimate the reliability of their classroom tests.

KEY TERMS AND CONCEPTS

Alternate-form reliability, p. 98
Coefficient alpha, p. 101
Composite score, p. 103
Confidence interval, p. 115
Content heterogeneity, p. 100
Content sampling error, p. 94
Error score, p. 92
Error variance, p. 95
Internal-consistency reliability, p. 98
Inter-rater differences, p. 95
Inter-rater reliability, p. 101
Kuder-Richardson formula 20, p. 101
Measurement error, p. 91
Obtained score, p. 92
Reliability, p. 91
Reliability coefficient, p. 95
Spearman-Brown formula, p. 99
Split-half reliability, p. 99
Standard error of measurement (SEM), p. 112
Test-retest reliability, p. 97
Time sampling error, p. 95
True score, p. 92
True score variance, p. 95

RECOMMENDED READINGS

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: AERA. Chapter 5, Reliability and Errors of Measurement, is a great resource!

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). Upper Saddle River, NJ: Merrill/Prentice Hall. A little technical at times, but a great resource for students wanting to learn more about reliability.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: W. H. Freeman. Chapters 8 and 9 provide outstanding discussions of reliability. A classic!

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill. Chapter 6, The Theory of Measurement Error, and Chapter 7, The Assessment of Reliability are outstanding chapters. Another classic!

Subkoviak, M. J. (1984). Estimating the reliability of mastery–nonmastery classifications. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 267–291). Baltimore: Johns Hopkins University Press. An excellent discussion of techniques for estimating the consistency of classification with mastery tests.

PRACTICE ITEMS

1. Consider these data for a five-item test that was administered to six students. Each item could receive a score of either 1 or 0. Calculate KR-20 using the following formula:

$$\text{KR-20} = \frac{k}{k-1}\left(\frac{\text{SD}^2 - \sum p_i \times q_i}{\text{SD}^2}\right)$$

where k = number of items
SD^2 = variance of total test scores
p_i = proportion of correct responses on item
q_i = proportion of incorrect responses on item

	Item 1	Item 2	Item 3	Item 4	Item 5	Total Score
Student 1	0	1	1	0	1	
Student 2	1	1	1	1	1	
Student 3	1	0	1	0	0	
Student 4	0	0	0	1	0	
Student 5	1	1	1	1	1	
Student 6	1	1	0	1	0	
p_i						
q_i						SD^2
$p_i \times q_i$						

Note: When calculating SD^2, use n in the denominator.

2. Consider these data for a five-item test that was administered to six students. Each item could receive a score ranging from 1 to 5. Calculate coefficient alpha using the following formula:

$$\text{Coefficient alpha} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma SD_i^2}{SD^2}\right)$$

where k = number of items
SD_i^2 = variance of individual items
SD^2 = variance of total test scores

	Item 1	Item 2	Item 3	Item 4	Item 5	Total Score
Student 1	4	5	4	5	5	
Student 2	3	3	2	3	2	
Student 3	2	3	1	2	1	
Student 4	4	4	5	5	4	
Student 5	2	3	2	2	3	
Student 6	1	2	2	1	3	
SD_i^2						SD^2 =

Note: When calculating SD_i^2 and SD^2, use n in the denominator.

Go to **www.pearsonhighered.com/reynolds2e** to view a PowerPoint™ presentation and to listen to an audio lecture about this chapter.