7

Reliability

This chapter introduces reliability—a topic that is broad and has important implications for any research endeavor. In this chapter, the classical true score model is introduced providing the foundation for the conceptual and mathematical underpinnings of reliability. After the foundations of reliability are presented, several approaches to the estimation of reliability are provided. Throughout the chapter, theory is linked to practical application.

7.1 INTRODUCTION

Broadly speaking, the term *reliability* refers to the degree to which scores on tests or other instruments are free of errors of measurement. The degree to which scores are free from errors of measurement dictates their level of consistency or **reliability**. Reliability of measurement is a fundamental issue in any research endeavor because some form of measurement is used to acquire data. The process of data acquisition involves the issues of measurement precision (or imprecision) and the manner by which it is reported in relation to test scores. As you will see, reliability estimation is directly related to measurement precision or imprecision (i.e., error of measurement). Estimating the reliability of scores according to the classical true score model involves certain assumptions about a person's observed, true, and error scores. This chapter introduces the topic of reliability in light of the assumptions of the true score model, how it is conceptualized, requisite assumptions about true and error scores, and how various coefficients of reliability are derived.

Two issues central to reliability are (1) the consistency or degree of similarity of at least two scores on a set of test items and (2) the stability of at least two scores on a set of test items over time. Different methods of estimating reliability are based on specific assumptions about true and error scores and, therefore, address different sources of error. The assumptions explicitly made regarding true and error scores are integral to correctly reporting and interpreting score reliability. Although the term *reliability* is used in a general sense in many instances, reliability is clearly a property of scores rather than measurement instruments or tests. *It is the consistency or stability of scores that provides evidence of reliability when using a test or instrument in a particular context or setting.*

This chapter is organized as follows. First, a conceptual overview of reliability is presented followed by an introduction to the **classical true score model**—a model that serves as the foundation for **classical test theory**. Next, several methods commonly used to estimate reliability are presented using the classical test theory approach. Specifically, we present three approaches to estimating reliability: (1) the test–retest method for estimating the stability of scores over time, (2) the internal consistency method based on the model of randomly parallel tests, and (3) the splithalf method—also related to the model of parallel tests. A subset of the dataset introduced in Chapter 2 that includes three components of the theory of generalized intelligence—fluid (*Gf*), crystallized (*Gc*), and short-term memory (*Gsm*)—is used throughout the chapter in most examples. As a reminder, the dataset used throughout this chapter includes a randomly generated set of item responses based on a sample size N = 1,000 persons. For convenience, the data file is available in SPSS (GfGc.sav), SAS (GfGc.sd7), or delimited file (GfGc.dat) formats and is downloadable from the companion website (*www.guilford.com/price2-materials*).

7.2 CONCEPTUAL OVERVIEW

As noted earlier, measurement precision is a critical component of reliability. For example, a useful way to envision the concept of reliability is to determine how free a set of scores is from measurement error. How one evaluates (or estimates) the degree of measurement error in a set of scores is a primary focus of this chapter and is foundational to understanding the various approaches to the estimation of reliability. Reliability is perhaps most concretely illustrated in fields such as chemistry, physics, or engineering. For example, measurements acquired in traditional laboratory settings are often acquired within the context of well-defined conditions, with precisely calibrated instrumentation, where the object of the measurement physically exists (i.e., directly observable and measureable physical properties). Consider two examples from chemistry: (1) measurement such as the volume of a gas in a rigid container at an exact temperature and (2) the precise amount of heat required to produce a chemical reaction. In the first example, say that a researcher measures the volume of gas in a rigid container on 10 different occasions. In summarizing the 10 measurements, one would expect a high degree of consistency, although there will be some **random error** variability in the numerical values acquired from the measurement due to fluctuations in instrumentation (e.g., calibration issues or noise introduced through the instruments used for the data collection). When research is conducted with human subjects, random error may occur due to distractions, guessing, content sampling, or intermittent changes in a person's mental state (see Table 7.1).

Another type of error is called systematic or **constant error** of measurement (Gulliksen, 1950b; 1987, p. 6). For example, systematic error occurs when all test scores are

TABLE 7.1. General and Specific Origins of Test Score Variance Attributable to Persons

General: Enduring traits or attributes

1. Skill in an area tested such as reading, mathematics, science
2. Test-taking ability such as careful attention to and comprehension of instructions
3. Ability to respond to topics or tasks presented in the items on the test
4. Self-confidence manifested as positive attitude toward testing as a way to measure ability, achievement, or performance

Specific: Enduring traits or attributes

1. Requisite knowledge and skill specific to the area or content being measured or tested
2. Emotional reactivity to a certain type of test item or question (e.g., the content of the item includes a topic that elicits an emotional reaction)
3. Attitude toward the content or information included on the test
4. Self-confidence manifested as positive attitude toward testing as a way to measure ability, achievement, or perfomance

General: Limited or fluctuating

1. Test-taking anxiety
2. Test preparation (e.g., amount and quality of practice specific to the content of items on the test)
3. Impact of test-taking environment (e.g., comfort, temperature, noise)
4. Current attitude toward the test and testing enterprise
5. Current state of physical health and level of mental/physical fatigue
6. Motivation to participate in the testing occasion
7. Relationship with person(s) administering the test

Specific: Limited or fluctuating

1. Momentary changes in memory specific to factual information
2. Test preparation (e.g., amount and quality of practice specific to the content of items on the test)
3. Guessing correct answers to items on the test
4. Momentary shift in emotion triggered by information included on test item
5. Momentary shifts in attention or judgment

Note. Based on Cronbach (1970).

excessively high or low. In the physical sciences, consider the process of measuring the precise amount of heat required to produce a chemical reaction. Such a reaction may be affected systematically by an improperly calibrated thermometer being used to measure the temperature—resulting in a systematic shift in temperature by the amount or degree of calibration error. In the case of research conducted with human subjects, systematic error may occur owing to characteristics of the person, the test, or both. For example, in some situations persons' test scores may vary in a systematic way that yields a consistently lower or higher score over repeated test administrations. With regard to the crystallized intelligence dataset used in the examples throughout this book, suppose that all of the subtests on the total test were developed for a native English-speaking population.

Further suppose that a non-native English-speaking person responds to all questions on the subtests. The person's scores over repeated testing occasions will likely be consistently lower (due to the language component) than their true or actual level of intellectual ability because English is not the respondents' first or primary language. However, *systematic error is not part of the theoretical assumptions of the true score model*—only random error is. Therefore, systematic errors are not regarded as affecting the reliability of scores; rather, they are a source of construct-related variance (an issue related to validity).

The example with non-native English speaking persons introduces one aspect of an important topic in psychometrics and/or test theory known as **validity** (i.e., the test not being used with the population for which it was developed). *Evidence of test validity is related to reliability such that reliability is a necessary but not sufficient condition to establish the validity of scores on a test.* The validity example is important because errors of measurement place limitations on the validity of a test. Furthermore, even if no measurement error existed, complete absence of measurement error in no way guarantees the validity of test scores. Validity, a comprehensive topic, is covered in Chapters 3 and 4 of this text. Table 7.1 provides examples of sources of error variability that may affect the reliability of scores (either randomly or systematically) when conducting research in social and/or behavioral science.

7.3 THE TRUE SCORE MODEL

In 1904, Charles Spearman proposed a model-based framework of test theory known as the **true score model**. For approximately a century, Spearman's true score model has largely dominated approaches to the estimation of reliability. This model rests on the assumption that test scores represent fallible (i.e., less than perfectly objective or accurate) measurements of human traits or **attributes**. Because perfect measurement can never occur, observed scores always contain some error. Based on the idea that measurements are fallible, Spearman (1904, 1907) posited that the observed correlation between such fallible scores is lower than would be observed if one were able to use true objective values. Over the past century, the true score model has been revised and/or expanded with formal, comprehensive treatments published by Harold Gulliksen (1950b, 1987) in *The Theory of Mental Tests* and Fredrick Lord and Melvin Novick (1968) in their seminal text *Statistical Theories of Mental Test Scores*. The true score model for a person is provided in Equation 7.1 (Lord & Novick, 1968, p. 56).

Equation 7.1. True score model

$$X_i = T_i + E_i$$

- X_i = observed fallible score for person *i*.
- T_i = true score for person *i*.
- E_i = error score for person *i*.

Although Equation 7.1 makes intuitive sense and has proven remarkably useful historically, six assumptions are necessary in order for the equation to become practical for use. Before introducing the assumptions of the true score model, some connections between probability theory, true scores, and random variables are reviewed in the next section (see the Appendix for comprehensive information on probability theory and random variables).

7.4 PROBABILITY THEORY, TRUE SCORE MODEL, AND RANDOM VARIABLES

Random variables are associated with a set of probabilities (see the Appendix). In the true score model, test scores are random variables and, therefore, can take on a hypothetical set of outcomes. The set of outcomes is expressed as a probability (i.e., expressed as a frequency) distribution as illustrated in Table 7.2. For example, when a person takes a test, the score he or she receives is considered a random variable (expressed in uppercase letter X in Equation 7.1). The one time or single occasion a person takes the test, he or she receives a score, and this score is *one sample from a hypothetical distribution of possible outcomes*. Table 7.2 illustrates probability distributions based on a hypothetical set of scores for three people. In the distribution of scores in Table 7.2, we assume that the same person has taken the same test repeatedly and that each testing occasion is an independent

TABLE 7.2. Probability of Obtaining a Particular Score on a 25-Item Test of Crystallized Intelligence on a Single Testing Occasion

	Person		
	A	B	C
Raw score (X)	$p(X)$	$p(X)$	$p(X)$
4	0.01	0.04	0.00
5	0.01	0.05	0.00
6	0.02	0.10	0.00
7	0.05	0.28	0.02
8	0.06	0.45	0.03
11	0.08	0.08	0.12
13	0.40	0.00	0.13
14	0.23	0.00	0.18
15	0.10	0.00	0.40
17	0.02	0.00	0.07
18	0.02	0.00	0.04
20	0.00	0.00	0.01
$\Sigma(X)p =$	12.54	7.45	14.02

Note. Each person has a unique score distribution independently determined for a single person. The frequency distribution of scores in the table is not based on any actual dataset used throughout this text; rather, it is only provided as an example.

event. The result is a distribution of scores for each person with an associated probability. The probabilities expressed in Table 7.2 are synonymous with the relative frequency for a score based on the repeated testing occasions. The implication of Table 7.2 for the true score model or classical test theory is that *the mean (or expectation) of the hypothetical observed score distribution for a person based on an infinitely repeated number of independent trials represents his or her true score within the classical true score model.*

To clarify the role of the person-specific probability distribution, consider the following example in Table 7.2. Tabulation of the probability of a person's raw score (expressed as a random variable) multiplied by the probability of obtaining a certain score (due to probability theory) demonstrates that person C appears to possess the highest level of crystallized intelligence for the 25-item test. Furthermore, by Equation 7.6, person C's true score is 14.02. Notice that for person C the probability (i.e., expressed as the relative frequency) of scoring a 15 is .40—higher than the other two persons. Person A has a probability of .40 scoring a 13. Person B has a probability of .45 scoring an 8. Clearly, person C's probability distribution is weighted more heavily toward the high end of the score scale than person A or B.

Although a person's **true score** is an essential component of the true score model, true score is only a *hypothetical entity* owing to the implausibility of conducting an infinite number of independent testing occasions. True score is expressed as the expectation of a person's observed score over repeated independent testing occasions. Therefore, *the score for each person taking the test represents a different random variable regarding his or her person-specific probability distribution* (e.g., Table 7.2). The result is that such persons have their own probability distribution—one that is specific to their hypothetical distribution of observed scores (i.e., each person has an associated score frequency or probability given their score on a test). In actual testing situations, the interest is usually in studying individual differences among people (i.e., measurements over people rather than on a single person). The true score model can be extended to accommodate the study of individual differences by administering a test to a random sample of persons from a population. Ideally, this process could be repeated an infinite number of times (under standardized testing conditions), resulting in an observed score random variable taking on specific values of score *X*. In the context described here, the error variance over persons can be shown to be equal to the average, over persons (group-level), of the error variance within persons (hypothetical repeated testing occasions for a single person; Lord & Novick, 1968, p. 35). Formally, this is illustrated in Equation 7.5 in the next section.

In the Appendix, equations for the expectation (i.e., the mean) of continuous and discrete random variables are introduced along with examples. In the true score model, total test scores for persons are called **composite scores**. Formally, such composite scores are defined as the sum of responses (response to an item as a discrete number) to individual items. At this point, readers are encouraged to review the relevant parts of Chapter 2 and the Appendix before proceeding through this chapter; this will reinforce key foundational information essential to understanding the true score model and reliability estimation. Next, we turn to a presentation of the assumptions of the true score model.

7.5 PROPERTIES AND ASSUMPTIONS OF THE TRUE SCORE MODEL

In the true score model, the human traits or attributes being measured are assumed to remain constant regardless of the number of times they are measured. Imagine for a moment that a single person is tested an infinite number of times repeatedly. For example, say Equation 7.1 is repeated infinitely for one person and the person's true state of knowledge about the construct remains unchanged (i.e., is constant). This scenario is illustrated in Figure 7.1.

Table 7.3 illustrates observed, true, and error scores for 10 individuals. Given this scenario, the person's observed score would fluctuate owing to random measurement error. The hypothetical trait or attribute that remains constant and that observed score fluctuates about is represented as a person's true score or *T*. Because of random error during the measurement process, a person's observed score *X* fluctuates over repeated trials or measurement occasions. The result of random error is that differences between a person's observed score and true score will fluctuate in a way that some are positive

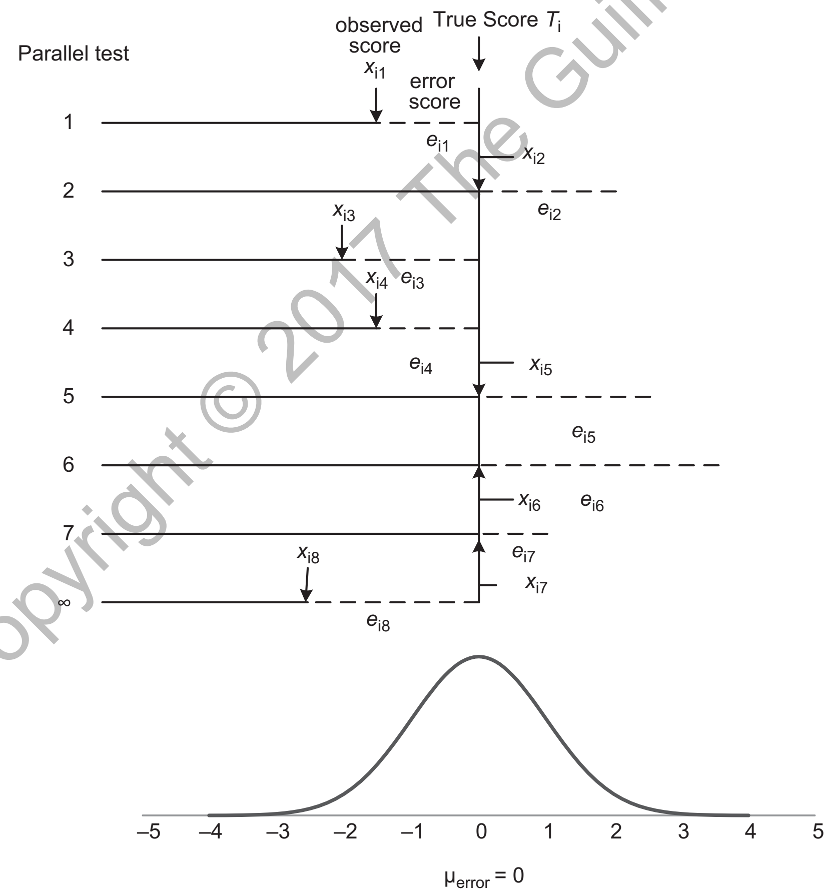


FIGURE 7.1. True score for a person. Adapted from Magnusson (1967, p. 63). Copyright 1967. Reprinted by permission of Pearson Education, Inc. New York, New York.

TABLE 7.3. Crystallized Intelligence Test Observed, True, and Error Scores for 10 Persons

Person (i)	Observed score (X)		True score (T)		Error score (E)
A	12.00	=	13.00	+	–1.00
B	14.50	=	12.00	+	2.50
C	9.50	=	11.00	+	–1.50
D	8.50	=	10.00	+	–1.50
E	11.50	=	9.00	+	2.50
F	7.00	=	8.00	+	–1.00
G	17.00	=	17.25	+	–0.25
H	17.00	=	16.75	+	0.25
I	10.00	=	9.00	+	1.00
J	8.00	=	9.00	+	–1.00
Mean	11.50		11.50		0.00
Standard deviation	3.43		3.11		1.45
Variance	11.75		9.66		2.11
Sum of cross products	96.50				
Covariance	9.65				

Note. Correlation of observed scores with true scores = .91. Correlation of observed scores with error scores = .42. Correlation of true scores with error scores = 0. True score values are arbitrarily assigned for purposes of illustration. Variance is population formula and is calculated using *N*. Partial credit is possible on test items. Covariance is the average of the cross products of observed and true deviation scores.

and some are negative. *Over an infinite number of testing occasions, the positive and negative errors cancel in a symmetric fashion, yielding an observed score equaling true score for a person* (see Equations 7.5 and 7.6).

Notice that in Table 7.4, all of the components are in place to evaluate the reliability of scores based on errors of measurement.

In the situation where score changes or shifts occur systematically, the difference between observed and true scores will be either systematically higher or lower by the factor of some constant value. For example, all test takers may score consistently lower on a test because the examinees are non-English speakers, yet the test items were written and/or developed for native English-speaking persons. Technically, *such systematic influences on test scores are not classified as error in the true score model* (only random error is assumed by the model). The error of measurement for a person in the true score model is illustrated in Equation 7.2. Alternatively, in Figure 7.2, the relationship between observed and true

TABLE 7.4. Correlations among Observed, True, and Error Scores for 10 Persons

	1	2	3
1. Observed	1	0.91	0.42
2. True		1	0.00
3. Error			1

Note. $\rho_{TE} = 0.0$; $\rho_{OE} = .42$; $\rho_{OT} = .91$; $\rho_{XX} = .82$ (which is the reliability coefficient expressed as the square of $\rho_{OT} = .91$); $\rho^2_{OE} = .42$; $\rho_{OT} = .91$. The correlation between true and error scores is actually .003 in the above example.

Equation 7.2. Error of measurement in the true score model for person *i*

$$E_i = X_i - T_i$$

- E_i = error score for person *i*.
- X_i = observed score for person *i*.
- T_i = true score for person *i*.

scores is expressed as the regression of true score on observed score (e.g., the correlation between true and observed score is .91 and $.91^2 = .82$ or the reliability coefficient).

Next, in Equation 7.3, the mean of the distribution of error is expressed as the expected difference between the observed score and true score for a person over infinitely repeated testing occasions (e.g., as in Table 7.3).

Because *X* and *T* are equal in the true score model (inasmuch as the mean observed score distribution over infinite occasions equals a person's true score distribution), the mean error over repeated testing occasions is also zero (Table 7.3; Figure 7.1; Equation 7.4;

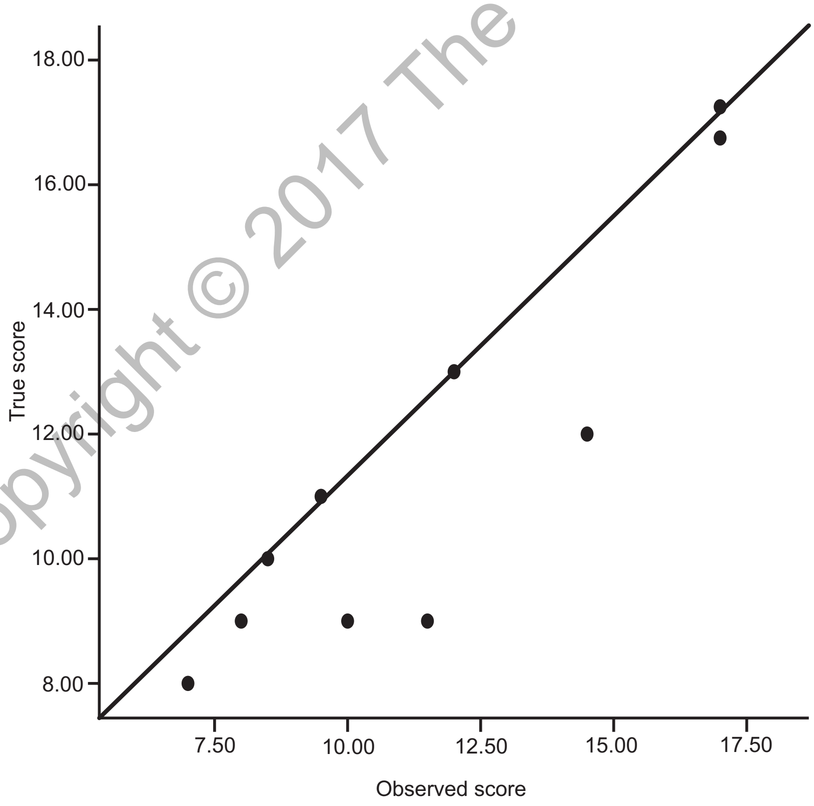


FIGURE 7.2. Regression line and scatterplot of true and observed scores for data in Table 7.3.

Equation 7.3. Mean error score for person *i* as expectation of the difference between observed score and true score

$$\mu_{E_i} = \varepsilon(E_i) = \varepsilon(X_i - T_i)$$

- T_i = true score for person *i*.
- μ_{E_i} = mean error score for person *i*.
- ε = expectation operator.
- (E_i) = observed error score for person *i*.
- X_i = mean of observed score *X* for subject *i*.

Equation 7.4. The expectation of random variable *E* for person *i*

$$\varepsilon = (E_i) = 0$$

- ε = expectation operator.
- (E_i) = expected value of random variable E_i over an indefinite number of repeated trials.

Lord & Novick, 1968, p. 36; Crocker & Algina, 1986, p. 111). Also, since the error component is random, then from classical probability theory (e.g., Rudas, 2008), the mean error over repeated trials equals zero (Figure 7.1). Accordingly, the first assumption in the true score model is that the mean error of measurement over repeated trials or testing occasions equals zero (Equation 7.4). *The preceding statement is true for (a) an infinite number of persons taking the same test—regardless of their true score and (b) for a single person's error scores on an infinite number of parallel repeated testing occasions.*

Assumption 1: The expectation (population mean) error for person *i* over an infinite number of trials or testing occasions on the same test is zero.

Extension to the Group Level

The expectation (mean) error for a *population* of persons (i.e., represented at the group level) over an infinite number of trials or testing occasions is zero. Equation 7.5 includes the double expectation operator to illustrate that the error variance over persons can be shown to be equal to the average over persons in a group of the error variance within persons (Lord & Novick, 1968, pp. 34–37). Here, the group notation is denoted by subscript *j* as presented in Crocker and Algina (1986, p. 111).

Equation 7.5. Mean error score for a population of persons

$$\mu_E = \varepsilon_j \varepsilon X_j$$

and

$$\mu_E = \varepsilon_j(0)$$

- μ_E = mean error for a population or group of persons.
- $\varepsilon_j \varepsilon$ = double expectation operator reflecting that the error variance over persons is equal to the average error variance within persons.
- ε_j = expectation for population or group j.
- εX_j = expectation taken over all persons in group j.

A main caveat regarding Equation 7.5 is that for a random sample of persons from a population, the average error may not actually be zero. The discrepancy between true score theory and applied testing settings may be due to sampling error or other sources of error. Also, in the true score model, one is *hypothetically* drawing a random sample of error scores from each person in the sample of examinees. The expected value or population mean of these errors may or may not be realized as zero.

Assumption 2: True score for person *i* is equal to the expectation (mean) of their observed scores over infinite repeated trials or testing occasions (Equation 7.6; Table 7.2).

Equation 7.6. True score for person *i* as expectation of mean observed score

$$T_i = \varepsilon(X_i) = \mu_{X_i}$$

- T_i = true score for person i.
- ε = expectation operator.
- (X_i) = observed score for person i.
- μ_{X_i} = mean of observed score X for subject i over independent trials.

The fact that a person's true score remains constant, yet unknown, over repeated testing occasions makes using Equation 7.1 for the estimation of reliability with empirical data intractable because without knowing a person's true score, deriving errors of measurement is impossible. To overcome the inability of knowing a person's true score, *items comprising a test are viewed as different parallel parts of a test, enabling estimation of the* ***reliability coefficient***. Given that items serve as parallel components on a test, reliability estimation proceeds in one of two ways. First, the estimation of reliability can proceed by evaluating the **internal consistency** of scores by using a sample of persons tested once, with test items serving as component pieces (each item being a "micro test") within the overall composite or total test score. Second, the estimation of reliability can proceed by deriving the stability of scores as the correlation coefficient for a sample of persons tested twice with the same instrument or on a parallel form of a test. Later in this chapter, several methods for estimating the reliability of scores are presented based on the true score model—all of which are based on the assumption of **parallel tests**.

Extension to the Group Level

True score for a group of persons is equal to the expectation (mean) of their observed scores over infinite repeated trials or testing occasions (Equation 7.7; Lord & Novick, 1968, p. 37; Gulliksen, 1950b, p. 29; Crocker & Algina, 1986, p. 111).

At this point, the properties of true and error scores within the true score model can be summarized as follows: (1) the mean of the error scores in a population or group of persons equals zero and (2) the expected population or group mean of observed scores equals the mean of true scores. We now turn to Assumption 3.

Assumption 3: In the true score model, the correlation between true and error scores on a test in a population of persons equals zero (Equation 7.8; Table 7.4; Figure 7.3).

Equation 7.7. True score as expectation of mean observed score for group *j*

$$T_j = \varepsilon(X_j) = \mu_{X_j}$$

- T_j = true score for a group *j*.
- ε = expectation operator.
- (X_j) = observed score for a group *j*.
- μ_{X_j} = mean of observed score *X* for group *j* over independent trials.

Equation 7.8. Correlation between true and error scores in the true score model

$$\rho_{TE} = 0$$

- ρ_{TE} = correlation between true and error scores in a population.

A consequence of the absence of correlation between true and error scores (Assumption 3, Equation 7.8) is that deriving the observed score variance is accomplished by summing true score variance and error variance (as linear components in Equation 7.9). This assumption implies that persons with low or high true scores do not exhibit systematically high or low errors of measurement because errors are randomly distributed (as in Figure 7.3). To illustrate the relationships between true and error scores, we return to the data in Table 7.3. In Table 7.4, we see that the correlation between true and error scores is zero (readers should calculate this for themselves by entering the data into SPSS or Excel

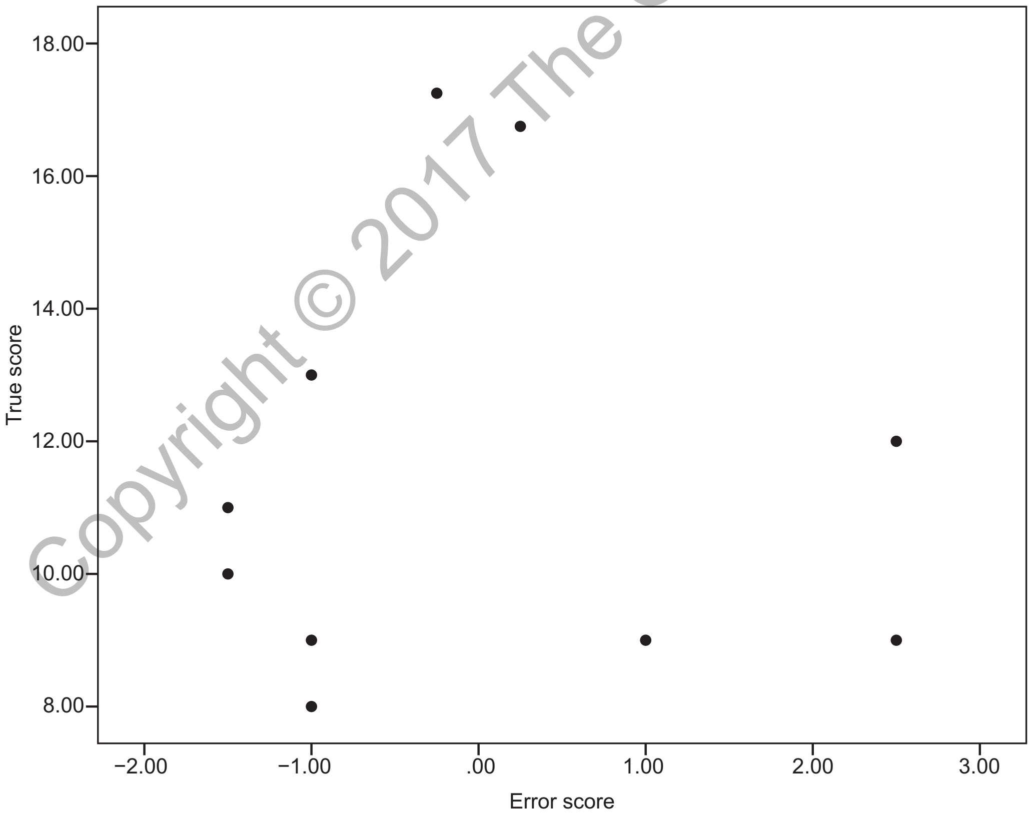


FIGURE 7.3. Correlation of true score with error score from data in Table 7.3.

Equation 7.9. Observed score variance as the sum of true score and error score

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

- σ_X^2 = observed score variance.
- σ_T^2 = true score variance.
- σ_E^2 = error score variance.

and conducting a correlation analysis). Next, because true score and error scores are uncorrelated, observed score variance is simply the sum of true and error score variance. To verify this statement, return to Table 7.3 and add the variance of true scores (9.66) to the variance of error scores (2.11) and you will see that the result is 11.75—the observed score variance. Formally, the additive, linear nature of observed score variance in the true score model is illustrated in Equation 7.9.

Assumption 4: When an independent random sample of persons from a population takes two separate tests that are parallel in structure and content, the correlation between the error scores on the two tests is zero (Equation 7.10; Lord & Novick, 1968, pp. 47–49; Crocker & Algina, 1986, p. 111).

Equation 7.10. Correlation between two sets of random error scores from two tests in the true score model

$$\rho_{E_1E_2} = 0$$

- $\rho_{E_1E_2}$ = population correlation between random errors of measurement for test 1 and parallel test 2.

Intuitively, Assumption 4 should be clear to readers at this point based on the presentation thus far regarding the nature of random variables as having no relationship (in this case zero correlation between errors of measurement on two parallel tests).

Assumption 5: Error scores on one test are uncorrelated with true scores on another test (Equation 7.11). For example, the error component on one intelligence test is not correlated with true score on a second, different test of intelligence.

Equation 7.11. Correlation between the error on test 1 and true score on test 2 are uncorrelated

$$\rho_{E_1T_2} = 0$$

- $\rho_{E_1T_2}$ = population correlation between the error on test 1 and true score on test 2 are uncorrelated.

Assumption 6: Two tests are exactly parallel if, for every population, their *true scores* and *error scores* are equal (Lord & Novick, 1968; Equation 7.12). Further, all items on a test are assumed to measure a single construct. This assumption of measuring a single construct is called unidimensionality and is covered in greater detail in Chapters 8 and 9 on factor analysis and item response theory. If two tests meet the assumptions of parallelism, they should be correlated with other external or criterion-related test scores that are parallel based on the content of the test. The parallel tests assumption is difficult to meet in practical testing situations because in order for the assumption to be tenable, the testing conditions that contribute to error variability presented in Table 7.1 (e.g., fatigue, environment, etc.) must vary in the same manner in each of the testing scenarios. Also, part of Assumption 6 is that every population of persons will exhibit equal observed score means (i.e., mean expressed the degree of **measurement precision** expressed as how close scores are to one another) and variances (i.e., as a measure of error) on parallel tests.

Equation 7.12. Definition of parallel tests

$$X_1 = T + E_1$$

$$X_2 = T + E_2$$

$$\sigma^2_{E_1} = \sigma^2_{E_2}$$

- X_1 = observed score on test 1.
- X_2 = observed score on test 2.
- T = true score (assumed as equal on both tests).
- $\sigma^2_{E_1}$ = variance of test 1.
- $\sigma^2_{E_2}$ = variance of test 2.

As previously stated, the model of parallel tests is important because it allows the true score model to become functional with empirical data. In fact, without the model of parallel tests, the true score model would be only theoretical because true scores are not actually measureable. Also, without knowing true scores, calculation of error scores would not be possible, making the model ineffective in empirical settings. To illustrate the importance of the model of parallel tests relative to its role in estimating the coefficient of reliability, consider Equations 7.13 and 7.14 (Crocker & Algina, 1986, pp. 115–116).

Equation 7.13. Deviation score formula as the correlation on parallel tests 1 and 2

$$\rho_{x_1 x_2} = \frac{\sum x_1\, x_2}{N\sigma_{x_1}\sigma_{x_2}}$$

- $\rho_{x_1x_2}$ = correlation between scores on two parallel tests.
- x_1 = observed deviation score on test 1.
- x_2 = observed deviation score on test 2.
- σ_{x_1} = observed standard deviation on test 1.
- σ_{x2} = observed standard deviation on test 2.
- N = sample size.

Equation 7.14. Deviation score formula as the correlation between parallel tests 1 and 2 with substitution of portions of Equation 7.12 in numerator

$$\rho_{x_1x_2} = \frac{\sum (t_1 + e_1)(t_2 + e_2)}{N\sigma_{x_1}\sigma_{x_2}} = \frac{\sigma_t^2}{\sigma_x^2}$$

- $\rho_{x_1x_2}$ = coefficient of reliability expressed as the correlation between parallel tests.
- t_1 = true score on test 1 in deviation score form.
- t_2 = true score on test 2 in deviation score form.
- x_1 = observed score on test 1 in deviation score form.
- x_2 = observed score on test 2 in deviation score form.
- σ_{x_1} = observed score on test 1.
- σ_{x_2} = observed score on test 2.
- N = sample size.
- $\frac{\sigma_t^2}{\sigma_x^2}$ = the coefficient of reliability expressed as the ratio of true score variance to observed score variance.

The first two lines in Equation 7.12 can be substituted into the numerator of Equation 7.13 yielding an expanded numerator in Equation 7.14. Notice in Equation 7.14 that *x*, *t*, and *e* are now lowercase letters in the numerator. The lowercase letters represent deviation scores (as opposed to raw test scores). A **deviation score** is defined as follows: $x-\bar{x}$; $t-\bar{t}$; $e-\bar{e}$; where raw scores are subtracted from their respective means.

The final bullet point in Equation 7.14, the *coefficient of reliability expressed as the ratio of true score variance to observed score variance*, is the most common definition of reliability in the true score model.

7.6 TRUE SCORE EQUIVALENCE, ESSENTIAL TRUE SCORE EQUIVALENCE, AND CONGENERIC TESTS

Returning to the example data in Table 7.3, notice that the assumption of exactly parallel tests is not met because, although the true and observed score means are equivalent, their standard deviations (and therefore variances) are different. This variation on the model of parallel tests is called **tau-equivalence**, meaning that *only the true (i.e., tau) scores are equal* (Lord & Novick, 1968, pp. 47–50). **Essential tau-equivalence** (Lord & Novick, 1968, pp. 47–50) is expressed by further relaxing the assumptions of tau-equivalence, thereby allowing true scores to differ by an additive constant (Lord & Novick, 1968; Miller, 1995). Including an additive constant in no way affects score reliability since the reliability coefficient is estimated using the covariance components of scores and is expressed in terms of the ratio of true to observed score variance (or as the amount of variance explained as depicted in Figure 7.1).

Finally, the assumption of **congeneric tests** (Lord & Novick, 1968, pp. 47–50; Raykov, 1997, 1998) is the least restrictive variation on the model of parallel tests because the only requirement is that true scores be perfectly correlated on tests that are designed to measure the same construct. The congeneric model also allows for either an additive and/or a multiplicative constant between each pair of item-level true scores so that the model is appropriate for estimating reliability in datasets with unequal means and variances. Table 7.5 summarizes variations on the assumptions of parallel tests within the classical true score model.

7.7 RELATIONSHIP BETWEEN OBSERVED AND TRUE SCORES

To illustrate the relationship among observed, true, and error scores, we return to using deviation scores based on a group of persons—a metric that is convenient for deriving the covariance (i.e., the unstandardized correlation presented in Chapter 2) among these score components. Recall that in Equation 7.1 the definition of observed score is the sum of the true score and error score. Alternatively, Equation 7.15 illustrates the same

TABLE 7.5. Four Measurement Models of Reliability Theory

Model assumption	Parallel tests	Tau-equivalent tests	Essentially tau-equivalent tests	Congeneric tests[a]
1. Equal *expected* observed scores	X	X	—	—
2. Equal standard deviations (variances) of *expected* observed scores	X	—	—	—
3. Equal covariance components for *expected* observed scores for any set of parallel tests or for any single parallel test and another test of a different construct	X	X	X	—
4. Equal coefficients of covariance or correlation	X	—	—	—
5. Equal coefficients of reliability	X	—	—	—

Note. Due to the axioms of classical test theory, expected observed scores equal true scores.
[a] In congeneric tests, there is no mathematically unique solution to the estimation of a reliability coefficient; thus only a lower bound should be reported.

Equation 7.15. Observed score, true score, and error score in deviation score units

$$x = t + e$$

- x = observed score on a test derived as a raw score minus the mean of the group scores.
- t = true score on a test derived as a true score minus the mean of the group of true scores.
- e = error score derived as an error score minus the mean of the group error scores.

elements in Equation 7.1 as deviation scores. In the previous section, a deviation score was defined as $x-\bar{x}$; $t-\bar{t}$; $e-\bar{e}$; where raw scores are subtracted from their respective means. An advantage of working through calculations in deviation score units is that the derivation includes the standard deviations of observed, true, and error scores—elements required for deriving the covariance among the score components. The covariance is expressed as the product of observed and true deviation scores divided by the sample size (N). For the data in Table 7.3, the covariance is 9.65: $\text{cov}_{ot} = \left[\sum (X_o - \overline{X_o})(X_t - \overline{X_t})/N\right]$ (as an exercise, you should use the data in Table 7.3 and apply it to the equation in this sentence to derive the covariance between true and observed scores). *Notice that in*

Equation 7.14 the covariance is incorporated into the derivation of the reliability index by including the standard deviations of observed and true scores in the denominator.

Next, recall that the true score model is based on a linear equation that yields a composite score for a person. By extension and analogy, a composite score is also expressed as the sum of the responses to individual test items (e.g., each test item is a micro-level test). Working with the covariance components of total or composite scores (e.g., observed, true, and error components) provides a unified or connecting framework for illustrating how the true score model works regarding the *estimation of reliability with individual and group-level scores* in the true score model and classical test theory.

7.8 THE RELIABILITY INDEX AND ITS RELATIONSHIP TO THE RELIABILITY COEFFICIENT

The **reliability index** (Equation 5.16; Crocker & Algina, 1986, pp. 114–115; Kelley, 1927; Lord & Novick, 1968) is defined as the correlation between observed scores and true scores. From the example data in Table 7.4 we see that this value is .91. The square of the reliability index (.91) is .82—the **coefficient of reliability** (see Table 7.4). Equation 7.16 illustrates the calculation of the reliability index working with deviation scores. Readers can insert the score data from Table 7.3 into Equation 7.16, then work through the steps and compare the results reported in Table 7.4 presented earlier.

7.9 SUMMARIZING THE WAYS TO CONCEPTUALIZE RELIABILITY

The observed score variance variable σ_x^2 can be expressed as the sum of the random true score variance σ_T^2 plus the random observed score error variance σ_E^2. Computing the observed score variance as a linear sum using separate, independent components is possible because true score errors are uncorrelated with observed score errors. Next, using the component pieces of true score error and observed score error, the coefficient of reliability can be conceptually expressed in Equation 7.17 as the ratio of true score variance to observed score variance.

Returning to the data in Table 7.3, we can insert the variance components from the table in Equation 7.17 to calculate the reliability coefficient. For example, the true score variance (9.66) divided by the observed score variance (11.75) equals .82, the coefficient of reliability (Table 7.4). The type of reliability estimation just mentioned uses the variance to express the proportion of variability in observed scores explained by true scores. To illustrate, notice that the correlation between true scores and error scores in Table 7.4 is .91. Next, if we square .91, a value of .82 results, or the reliability coefficient. In linear regression terms, the reliability (.82) is expressed as the proportion of variance in true scores explained by variance in observed scores (see Figure 7.2).

Equation 7.16. The reliability index or the correlation between observed scores and true scores expressed as the ratio of standard deviation of true scores to the standard deviation of observed scores

$$\begin{aligned}
\rho_{xt} &= \frac{\sum (t+e)\,t}{N\sigma_X\sigma_t} \\
&= \frac{\sum t^2 + \sum te}{N\sigma_X\sigma_t} \\
\rho_{xt} &= \frac{\sum t^2}{N\sigma_X\sigma_t} + \frac{\sum te}{N\sigma_X\sigma_t}
\end{aligned}$$

The last term above cancels because, by tautology, the correlation between true and error scores is zero, and since

$$\sigma_T^2 = \frac{\sum t^2}{N}\text{, then}$$

$$\rho_{XT} = \frac{\sigma_T^2}{\sigma_X\sigma_T}\text{, simplifying to}$$

$$\rho_{XT} = \frac{\sigma_T}{\sigma_X}$$

- ρ_{XT} = reliability index.
- σ_T = standard deviation of true scores.
- σ_X = standard deviation of observed scores.
- t = true score in deviation score units.
- e = error score in deviation score units.
- $\sum$ = summation operator.
- N = population size.
- σ_T^2 = variance of true scores.
- $\sum t^2$ = sum of true scores squared.

Finally,

ρ_{XT}^2 = the index of reliability squared is the coefficient of reliability.

Equation 7.17. Coefficient of reliability expressed as a ratio of variances

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E}$$

- ρ^2_{XT} = coefficient of reliability.
- σ^2_T = true score variance.
- σ^2_X = observed score variance.
- σ^2_E = error score variance.

Equation 7.17 illustrates that the squared correlation between true and observed scores is the coefficient of reliability. Yet another way to think of reliability is in terms of the lack of error variance. For example, we may think of the lack of error variability expressed as $1-\left(\sigma^2_E/\sigma^2_O\right)$. Referring to the data in Table 7.3, this value would be 1 – .18 = .82, or the coefficient of reliability. Finally, reliability may be described as the lack of correlation between observed and error scores, or $1-\rho^2_{\mathrm{OE}}$, which, based on the data in Table 7.3, is .82 or the coefficient of reliability.

7.10 RELIABILITY OF A COMPOSITE

Earlier in this chapter it was stated that individual items on a test can be viewed as parallel components of a test. This idea is essential to understanding how reliability coefficients are estimated within the model of parallel tests in the true score model. Specifically, test items serve as individual, yet parallel, parts of a test providing a way to estimate the coefficient of reliability from a single test administration. Recall that a score on an individual item is defined by a point value assigned based on a person's response to an item (e.g., 0 for incorrect or 1 for correct). In this sense, an item is a "micro-level" testing unit, and an item score is analogous to a "micro-level test." The variance of each item can be summed to yield a total variance for all items comprising a test. Equations 7.18a and 7.18b illustrate how the variance and covariance of individual test items can be used to derive the total variance of a test.

Based on Equation 7.18a, we see that total test variance for a composite is determined by the variance and covariance of a set of items. In Table 7.6, the total variance is the sum of the variances for each item (1.53), plus 2 times the sum of the individual covariance values (1.08), equaling a total test variance of 2.61.

Equation 7.18a. Test variance based on the sum of individual items

$$\sigma^2_{\text{test}} = \sum \sigma_i^2 + 2\sum \rho_{ik}\sigma_i\sigma_k, \quad i > k$$

- σ^2_{test} = variance of total test.
- σ^2_i = variance of an individual item.
- ρ_{ik} = correlation between items i and k.
- σ_i = standard deviation of item i.
- σ_k = standard deviation of item k.
- $\rho_{ik}\sigma_i\sigma_k$ = covariance of items i through k resulting in $n(n-1)$ terms.
- $2\sum \rho_{ik}\sigma_i\sigma_k$ = two times (2×) the sum of all $n(n-1)$ covariance terms.

Equation 7.18b. Test variance based on the data in Table 7.6

$$\begin{aligned}\sigma^2_{\text{test}} &= 1.53 + 2(.54)\\ &= 1.53 + 1.08\\ &= 2.61\end{aligned}$$

TABLE 7.6. Variance–Covariance Matrix Based on 10 Crystallized Intelligence Test Items

Item	1	2	3	4	5	6	7	8	9	10
1	**0.10**	–0.01	–0.01	–0.02	–0.01	–0.02	–0.02	–0.01	–0.04	0.08
2		**0.10**	–0.01	–0.02	–0.01	–0.02	–0.02	–0.01	0.07	–0.03
3			**0.10**	0.09	–0.01	0.09	–0.02	–0.01	0.07	–0.03
4				**0.18**	–0.02	0.18	0.07	0.09	0.13	–0.07
5					**0.10**	–0.02	–0.02	–0.01	–0.04	0.08
6						**0.18**	0.07	0.09	0.13	–0.07
7							**0.18**	0.09	0.02	–0.07
8								**0.10**	0.07	–0.03
9									**0.27**	–0.13
10										**0.23**

Note. Variances are in **bold** on the diagonal and covariance elements are off-diagonal entries. Σ variances = **1.53**; Σ covariances = **0.54**.

If we replace the "items" in Table 7.6 with "total test scores" (i.e., the total score being based on the sum of items comprising a test), the same concept and statistical details will apply regarding how to derive the total variance for a set of total test scores. Next, we turn to the use of total test scores that are useful as individual components for deriving a composite score.

In the true score model, total test scores are created by *summing the item response values* (i.e., score values yielding points awarded) for each person. The total score for a test derived in this manner is one form of a composite score. Another form of composite score is derived by summing total test scores for two or more tests. In this case, a composite score is defined as the sum of *individual total test scores*. Returning to the data used throughout this book, suppose that you want to create a composite score for crystallized intelligence by summing the total scores obtained on each of the four subtests for crystallized intelligence. The summation of the four total test scores yields a composite score that represents crystallized intelligence. Equation 7.19 illustrates the derivation of a composite score for crystallized intelligence (labeled CIQ). The composite score, *CIQ*, represents the sum of four subtests, each representing a different measure of crystallized intelligence.

Given that composites are based on item total scores (for a single test) or total test scores (for a linear composite comprised of two or more tests), these composites formally serve as parallel components on a test. Applying the definition of parallel test components, reliability estimation proceeds according to the technique(s) appropriate for accurately representing the reliability of scores given the type of study. Specifically, the estimation of reliability may proceed by one or more of the following techniques. First, you may derive the stability of scores using the test–retest method. Second, you may derive the equivalence of scores based on parallel test forms. Third, you may derive the internal consistency of scores by using a sample of persons tested once with test items

Equation 7.19. Observed score composite based on the linear sum of four crystallized intelligence tests

$$CIQ = X1_{\text{crystallized1}} + X2_{\text{crystallized2}} + X3_{\text{crystallized3}} + X4_{\text{crystallized4}}$$

- CIQ = composite score expressed as the linear combination of crystallized intelligence tests 1–4.
- $X1_{\text{crystallized1}}$ = total score for crystallized intelligence test 1.
- $X2_{\text{crystallized2}}$ = total score for crystallized intelligence test 2.
- $X3_{\text{crystallized3}}$ = total score for crystallized intelligence test 3.
- $X4_{\text{crystallized4}}$ = total score for crystallized intelligence test 4.

serving as parallel pieces within the overall composite using the **split-half reliability** method or by deriving the internal consistency of scores using the Küder–Richardson formula 20 (KR20) or (21) or Cronbach's coefficient alpha. Each of the internal consistency techniques is based on there being as many parallel tests as there are items on the test. To derive the variance of the composite score, Equation 7.20a is required. Equation 7.20b illustrates the application of Equation 7.20a with data from Table 7.7.

Based on Equation 7.20b, the total variance of the composite using the data in Table 7.7 is 214.92.

To conclude this section, recall that earlier in this chapter individual test items comprising a test were viewed as parallel parts of a test. The requirements for parallel tests or measurements include (1) equal mean true scores, (2) equal (item or test) standard deviations, and (3) equal item (or test) variances. Specifically, test items (or total test scores)

Equation 7.20a. Observed score variance of a composite score derived from crystallized tests 1–4

$$\sigma^2_{CIQ} = \sigma^2_{\text{crystallized1}} + \sigma^2_{\text{crystallized2}} + \sigma^2_{\text{crystallized3}} + \sigma^2_{\text{crystallized4}} + \sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j$$

- σ^2_{CIQ} = variance of a composite score expressed as crystallized intelligence based on the sum of individual total test scores.
- $\sigma^2_{\text{crystallized1}}$ = variance of the crystallized intelligence test 1.
- $\sigma^2_{\text{crystallized2}}$ = variance of the crystallized intelligence test 2.
- $\sigma^2_{\text{crystallized3}}$ = variance of the crystallized intelligence test 3.
- $\sigma^2_{\text{crystallized4}}$ = variance of the crystallized intelligence test 4.
- $\sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j$ = sum of $k(k-1)$ covariance terms (i.e., k = intelligence tests 1–4), where i and j represent any pair of tests.

Equation 7.20b. Observed score variance of a composite score derived from crystallized tests 1–4 based on data in Table 7.7

$$\begin{aligned}\sigma^2_{CIQ} &= 47.12 + 24.93 + 12.40 + 21.66 + 108.81 \\ &= 214.92\end{aligned}$$

TABLE 7.7. Composite Scores for Crystallized Intelligence Tests 1–4

	Crystallized total score test 1	Crystallized total score test 2	Crystallized total score test 3	Crystallized total score test 4
	39	14	23	17
	47	17	24	24
	28	8	14	12
	29	6	19	11
	27	5	22	17
	35	11	18	11
	44	15	25	22
	36	5	17	15
	42	17	22	21
	36	6	18	19
Mean	36.3	10.4	20.2	16.9
SD	6.86	4.99	3.52	4.65
Variance	47.12	24.93	12.40	21.66
		Variance–covariance matrix		
	47.12	28.64	15.93	25.48
	—	24.93	11.69	14.71
	—	—	12.40	12.36
	—	—	—	21.66
			Total variance =	214.92

serve as individual, yet parallel, parts of a test, providing a way to estimate the coefficient of reliability from a single test administration. Equation 7.21 provides a general form for deriving true score variance of a composite. Equations 7.20a and 7.21 are general because they can be used to estimate the variance of a composite when test scores exhibit unequal standard deviations and variances (i.e., the equations allow for the covariation between all items whether equal or unequal).

Equation 7.21. General form for true score variance of a composite

$$\sigma^2_{TIQ} = \sigma^2_{\text{true\_score\_crystallized1}} + \sigma^2_{\text{true\_score\_crystallized2}} + \sigma^2_{\text{true\_score\_crystallized3}} + \sigma^2_{\text{true\_score\_crystallized4}} + \sum_{i \neq j} \rho_{ij}\sigma_i\sigma_j$$

Using the foundations of the CTT model, in the next section, we review several techniques for estimating the coefficient of reliability in specific research or applied situations.

7.11 COEFFICIENT OF RELIABILITY: METHODS OF ESTIMATION BASED ON TWO OCCASIONS

Coefficient of Stability: Test–Retest Method

Estimating the stability of test scores involves administering the same test to the same persons twice in as similar situations as possible. Once the data are collected, one correlates the scores of two test administrations. Reliability estimation under this approach yields a **coefficient of stability**. For example, a researcher may want to know how consistently persons respond to the same test at different times. In this context, the interest is in how stable a person's observed scores are in relation to his or her true score on a trait or attribute of interest (e.g., intelligence).

The test–retest method relies on two assumptions. The first assumption is that a person's true score is stable over time and, therefore, does not change. The second assumption is that a person's error scores are stable over time. These two assumptions provide the basis for establishing the degree to which a group of persons' scores exhibit equal reliability over time. *The main challenge regarding the assumptions of the test–retest method is that true scores for persons do not change over time*. There are three reasons for challenging this assumption. First, constructs that reflect "states" such as mood or anxiety are unlikely to remain stable over time (i.e., state-type attributes are highly variable over time such as days or weeks). For this reason, if a test is measuring mental "states," the test–retest method for estimating reliability is seldom useful. Conversely, the construct of adult intelligence is classified as a "trait" or attribute that is stable over time. For constructs that reflect traits, the test–retest method is often useful because it provides a basis for establishing the degree to which a group of persons' scores on a trait is equally reliable over time.

The second challenge to the assumption of the lack of change in a person's true score over time is attributed to the length of the interval between the first and second test administrations. The longer the interval between the first and second testing periods, the greater the likelihood of change in the psychological attribute. If the time between the first and second testing periods is too short (i.e., less than 14 days), the chances of a carryover (memory or practice) or contamination (additional information acquired by persons) effect are high. The ideal time between the first and second test administrations is between 14 and 28 days (Nunnally & Bernstein, 1994). Regarding the acceptable level of test–retest reliability coefficients for tests of ability or achievement where significant diagnostic or educational decisions often hinge, values of at least .90 are recommended.

For personality, attitude, or interest inventories, test–retest coefficients are usually lower, and the recommended range is between .80 and .90.

The final challenge to the test–retest method is related to chronological age. For example, although research has established that adult intelligence is stable over time (Wechsler, 1997b), this is not the case with the intelligence of children.

Coefficient of Equivalence: Parallel (Alternate) Forms Method

As previously stated, one way to define the reliability coefficient is the correlation between two strictly parallel tests. The parallel or alternate forms approach to reliability estimation directly incorporates this definition. The alternate forms approach to reliability estimation is useful when having parallel forms of a test is desirable. For example, parallel test forms may be useful (1) when persons are required to repeat an examination with a short time period between the two testing occasions or (2) to reduce the possibility of cheating when a single group of persons is taking a test in the same location.

To use the parallel forms technique, one creates two tests that, as nearly as possible, meet the requirement of strictly parallel tests. Recall that this requirement means that, for a group of persons, (1) the same set of true scores is being measured and the true scores are equal, and (2) error scores (or variances) are equal. If the requirements for strict parallelism are tenable, the two test forms are administered by using (1) the same persons in a retest situation or (2) a group of persons taking two forms of the test at the same time. Once the scores from the two tests are obtained, one proceeds by conducting a correlation analysis between the scores obtained.

Perhaps the strongest criticism of the alternate forms method is that one can argue that because two tests are composed of different items, the two forms can never be exactly parallel—at least theoretically speaking. A second criticism of the alternative forms method is related to carryover or memory effects. Earlier in this chapter, it was stated that in the true score model of parallel tests, error scores are required to be uncorrelated. However, if a carryover effect exists, as is sometimes the case, the errors of measurement for a group of persons will be correlated—sometimes substantially. For these reasons, if the parallel forms method involves retesting the same persons with an alternate form, the same concerns cited in the test–retest method apply (i.e., carryover effects due to memory or additional information gleaned by persons between testing occasions). In applied testing situations, if the researcher can demonstrate strong evidence that the assumptions of the true score model of parallel tests are tenable, then the alternate forms coefficient of reliability may be reported. Additionally, in order to provide comprehensive evidence, the parallel forms method is often accompanied by an estimate of internal consistency reliability—a subject covered in the next section.

7.12 METHODS BASED ON A SINGLE TESTING OCCASION

Split-Half Methods

Often it is not possible or desirable to compose and administer two forms of a test, as discussed earlier. Here we describe a method for deriving the reliability of total test scores based on parallel half tests. The split-half approach to reliability estimation involves dividing a test composed of a set of items into halves that, to the greatest degree possible, meet the assumptions of exact parallelism. The resulting scores on the respective half tests are then correlated to provide a **coefficient of equivalence**. *The coefficient of equivalence is actually the reliability based on one of the half tests*. However, remember that owing to the assumption of parallel test halves, we can apply a formula for deriving the reliability of scores on the total test using the **Spearman–Brown formula**. For tests composed of items with homogeneous content (a.k.a. **item homogeneity**; Coombs, 1950), the split-half method proceeds according to the following steps. First, after scores on the total test are obtained, items are assigned to each half test in either (a) a random fashion or (b) according to order of item difficulty. This process yields one parallel subtest that is composed of odd-numbered items, and a second half test is composed of even-numbered items. The split-half technique described allows one to create two parallel half tests that are of equal difficulty and have homogeneous item content.

Earlier it was stated that two parallel half tests can be created with the intent to target or measure the same true scores with a high degree of accuracy. One way to ascertain if two tests are parallel is to ensure that the half tests have equal means and standard deviations. Also, the test items in the two half tests should have the same content (i.e., exhibit item homogeneity). A high level of item homogeneity ensures that, as the correlation between the two half tests approaches 1.0, the approximation to equal true scores is as accurate as possible. If, however, the two half tests comprise items with partially heterogeneous content, then certain parts of the two half tests will measure different true scores. *In this case, the two half tests should be created based on matching test halves, where test items have been matched on difficulty and content.* Table 7.8 provides example

TABLE 7.8. Split-Half Data for 10 Persons from the 25-Item Crystallized Intelligence Test 2

	Half test 1	Half test 2
	Odd items (total score)	Even items (total score)
Mean	10.30	4.20
Variance	6.23	5.96
Variance of total test:		21.17
Odd/even correlation ($\rho_{ii'}$):		**0.69**
Split-half reliability:		**0.85**
Guttman split-half reliability:		**0.85**

data for illustrating the split-half and Guttman (1946) methods for estimating reliability based on half tests. **Rulon's formula** (1939) (equivalent to Guttman's formula) does not assume equal standard deviations (and variances) on the half test components. Finally, when the variances on the half tests are approximately equal, the Rulon formula and **Guttman's equation** yield the same result as the split-half method with the Spearman–Brown formula.

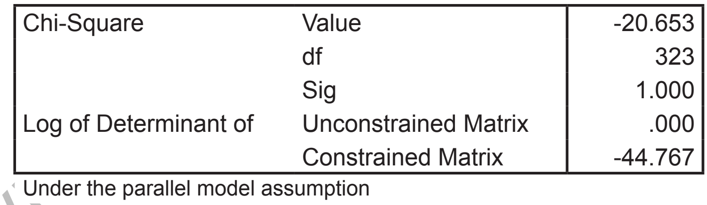
The SPSS syntax for computing the split-half reliability based on the model of parallel tests (not strictly parallel) is provided below.

```
RELIABILITY
/VARIABLES=cri2_01 cri2_02 cri2_03 cri2_04 cri2_05 cri2_06
cri2_07 cri2_08 cri2_09 cri2_10 cri2_11 cri2_12 cri2_13 cri2_14
cri2_15 cri2_16 cri2_17 cri2_18 cri2_19 cri2_20 cri2_21 cri2_22
cri2_23 cri2_24 cri2_25
/SCALE('ALL VARIABLES') ALL
/MODEL=PARALLEL.
```

The resulting output is provided in Tables 7.9a and 7.9b.

Equation 7.22 can be extended to deriving the reliability of any composite (e.g., the parallel components may be subtest total scores rather than individual items). Equation 7.23 illustrates Rulon's formula, as applied by Guttman, for total test score reliability. Rulon's formula is based on the error variances on half tests and the total test variance.

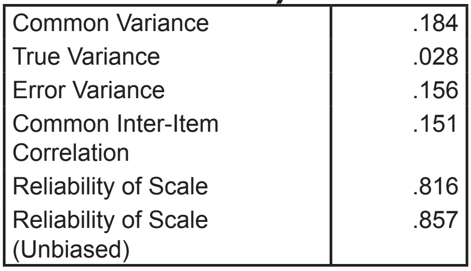
TABLE 7.9a. Test for Model Goodness of Fit



Chi-Square	Value	-20.653
	df	323
	Sig	1.000
Log of Determinant of	Unconstrained Matrix	.000
	Constrained Matrix	-44.767

Under the parallel model assumption

TABLE 7.9b. Reliability Statistics



Common Variance	.184
True Variance	.028
Error Variance	.156
Common Inter-Item Correlation	.151
Reliability of Scale	.816
Reliability of Scale (Unbiased)	.857

Equation 7.22. Spearman–Brown formula for total test score reliability based on the correlation between parallel split-halves

$$\rho_{xx'} = \frac{2(\rho_{ii'})}{1+\rho_{ii'}}$$

- $\rho_{ii'}$ = correlation between half tests.
- $\rho_{xx'}$ = split-half reliability based on the Spearman–Brown formula.

Equation 7.23. Rulon's formula for total test score reliability based on the correlation between parallel split-halves

$$\rho_{xx'} = 2\left[1-\left(\frac{\sigma^2_{\text{half test1}}+\sigma^2_{\text{half test2}}}{\sigma^2_{\text{total test}}}\right)\right]$$

The SPSS syntax for computing the Guttman model of reliability is as follows:

```
RELIABILITY
/VARIABLES=cri2_01 cri2_02 cri2_03 cri2_04 cri2_05 cri2_06
cri2_07 cri2_08 cri2_09 cri2_10 cri2_11 cri2_12 cri2_13 cri2_14
cri2_15 cri2_16 cri2_17 cri2_18 cri2_19 cri2_20 cri2_21 cri2_22
cri2_23 cri2_24 cri2_25
/SCALE('ALL VARIABLES') ALL
/MODEL=GUTTMAN.
```

The Guttman model provides six lower-bound coefficients (i.e., expressed as lambda coefficients). The output for the Guttman reliability model is provided in Table 7.10. The lambda 3 (L3) is based on estimates of the true variance of scores on each item and is also expressed as the average covariance between items and is analogous to coefficient alpha. Guttman's lambda 4 is interpreted as the greatest split-half reliability.

TABLE 7.10. Reliability Statistics

Lambda	1	.783
	2	.865
	3	.816
	4	.848
	5	.830
	6	.
N of Items		25

Internal Consistency: Methods Based on Covariation among Items

The final section of this chapter introduces approaches based on covariation among or between test items. The methods presented here were developed to provide a way to estimate the coefficient of reliability from a single test administration without splitting the single test into parallel halves. Specifically, the methods presented in this chapter include **coefficient alpha**, the **Küder–Richardson 20**, and the **Küder–Richardson 21** formulas.

Coefficient Alpha

The first and most general technique for the estimation of internal consistency reliability is known as coefficient alpha and is attributed to L. J. Cronbach (1916–2001). In his work (1951), Cronbach provided a general formula for deriving the internal consistency of scores. Coefficient alpha is a useful formula because of its generality. For example, alpha is effective for estimating score reliability for test items that are scored dichotomously (correct/incorrect), or for items scored on an ordinal level of measurement (e.g., Likert-type or rating scale items) and even for essay-type questions that often include differential scoring weights. For these reasons, coefficient alpha is reported in the research literature more often than any other coefficient. The general formula for coefficient alpha is provided in Equation 7.24. Table 7.11 includes summary data for 10 persons on the 25-item crystallized intelligence test 2 used in the previous section on split-half methods.

The total test variance for the crystallized intelligence test 2 is 19.05 (defined as the sum of the squared deviations from the mean) for 10 persons in this example data. Readers are encouraged to conduct the calculation of coefficient alpha using the required parts of Equation 7.24 by accessing the raw item-level Excel file: "Reliability_Calculation_Examples.xlsx" on the companion website (*www.guilford.com/price2-materials*). Knowing that the test is composed of 25 items, the total test variance is 19.05 and the sum of the

Equation 7.24. Coefficient alpha

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\Sigma\hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right)$$

- $\hat{\alpha}$ = coefficient alpha.
- k = number of items.
- $\hat{\sigma}_i^2$ = variance of item i.
- $\hat{\sigma}_X^2$ = total test variance.

TABLE 7.11. Item Summary Data for 10 Persons from Crystallized Intelligence Test 2

	Proportion correct	Proportion incorrect	Item variance
Item	*p*	*q*	*p*q*
1	0.9	0.1	0.09
2	0.9	0.1	0.09
3	0.8	0.2	0.16
4	0.8	0.2	0.16
5	0.9	0.1	0.09
6	0.8	0.2	0.16
7	0.9	0.1	0.09
8	0.9	0.1	0.09
9	0.6	0.4	0.24
10	0.7	0.3	0.21
11	0.7	0.3	0.21
12	0.6	0.4	0.24
13	0.8	0.2	0.16
14	0.8	0.2	0.16
15	0.6	0.4	0.24
16	0.7	0.3	0.21
17	0.4	0.6	0.24
18	0.3	0.7	0.21
19	0.3	0.7	0.21
20	0.8	0.2	0.16
21	0.3	0.7	0.21
22	0.2	0.8	0.16
23	0.2	0.8	0.16
24	0.1	0.9	0.09
25	0.1	0.9	0.09
Σp =	15.1	Σp^*q =	4.13

item-level variances is 4.13, we can insert these values into Equation 7.23 and derive the coefficient alpha as .82.

7.13 ESTIMATING COEFFICIENT ALPHA: COMPUTER PROGRAM AND EXAMPLE DATA

The SPSS syntax and SAS source code that produces output using the data file .sav is provided on the next page. The dataset may be downloaded from the companion website (*www.guilford.com/price2-materials*).

SPSS program syntax for coefficient alpha using data file Coefficient_Alpha_ Reliability_N_10_Data.SAV

```
RELIABILITY
/VARIABLES=cri2_01 cri2_02 cri2_03 cri2_04 cri2_05 cri2_06
cri2_07 cri2_08 cri2_09   cri2_10 cri2_11 cri2_12 cri2_13
cri2_14 cri2_15 cri2_16 cri2_17 cri2_18 cri2_19 cri2_20 cri2_21
cri2_22 cri2_23 cri2_24 cri2_25
/SCALE('ALL VARIABLES') ALL
/MODEL=ALPHA
/STATISTICS=DESCRIPTIVE SCALE
/SUMMARY=TOTAL.
```

Tables 7.12a–d are derived from the SPSS program.

TABLE 7.12a. Reliability Statistics

Cronbach's Alpha	N of Items
.816	25

TABLE 7.12b. Item Statistics

	Mean	Std. Deviation	N
cri2_01	.90	.316	10
cri2_02	.90	.316	10
cri2_03	.90	.316	10
cri2_04	.80	.422	10
cri2_05	.90	.316	10
cri2_06	.80	.422	10
cri2_07	.80	.422	10
cri2_08	.90	.316	10
cri2_09	.60	.516	10
cri2_10	.70	.483	10
cri2_11	.70	.483	10
cri2_12	.60	.516	10
cri2_13	.80	.422	10
cri2_14	.80	.422	10
cri2_15	.60	.516	10
cri2_16	.70	.483	10
cri2_17	.40	.516	10
cri2_18	.30	.483	10
cri2_19	.30	.483	10
cri2_20	.20	.422	10
cri2_21	.30	.483	10
cri2_22	.20	.422	10
cri2_23	.20	.422	10
cri2_24	.10	.316	10
cri2_25	.10	.316	10

TABLE 7.12c. Item–Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
cri2_01	13.60	21.378	-.106	.824
cri2_02	13.60	20.489	.202	.815
cri2_03	13.60	20.267	.281	.812
cri2_04	13.70	18.233	.765	.791
cri2_05	13.60	21.378	-.106	.824
cri2_06	13.70	18.233	.765	.791
cri2_07	13.70	19.344	.443	.806
cri2_08	13.60	19.156	.690	.799
cri2_09	13.90	18.544	.530	.800
cri2_10	13.80	22.178	-.274	.839
cri2_11	13.80	18.844	.498	.802
cri2_12	13.90	19.433	.322	.811
cri2_13	13.70	18.456	.699	.794
cri2_14	13.70	18.233	.765	.791
cri2_15	13.90	19.656	.272	.814
cri2_16	13.80	17.511	.847	.784
cri2_17	14.10	17.878	.692	.791
cri2_18	14.20	18.400	.611	.796
cri2_19	14.20	21.733	-.178	.834
cri2_20	14.30	20.233	.199	.816
cri2_21	14.20	20.844	.020	.825
cri2_22	14.30	19.344	.443	.806
cri2_23	14.30	20.233	.199	.816
cri2_24	14.40	20.267	.281	.812
cri2_25	14.40	21.156	-.031	.822

TABLE 7.12d. Scale Statistics

Mean	Variance	Std. Deviation	N of Items
14.50	21.167	4.601	25

SAS source code for coefficient alpha using SAS data file alpha_reliability_data

```
libname work 'LPrice_09';
data temp; set work.alpha_reliability_data;
proc corr data=temp nosimple alpha;
Title 'Coefficient Alpha using Crystallized Intelligence Example
Data N=10 ';
var cri2_01 - cri2_25;
run; quit;
```

Table 7.13 is produced from the SAS program.

TABLE 7.13. SAS Output for Coefficient Alpha

```
    Coefficient Alpha using Crystallized Intelligence Example Dat N=10              1

                                            10:45 Tuesday, November 15, 2011


                                The CORR Procedure
 25        CRI2_01   CRI2_02   CRI2_03   CRI2_04   CRI2_05   CRI2_06   CRI2_07   CRI2_08
 Vari-
 ables:

           CRI2_09   CRI2_10   CRI2_11   CRI2_12   CRI2_13   CRI2_14   CRI2_15   CRI2_16
           CRI2_17   CRI2_18   CRI2_19   CRI2_20   CRI2_21   CRI2_22   CRI2_23   CRI2_24
           CRI2_25

                             Cronbach Coefficient Alph
                              Variables        Alpha
                            ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
                                 Raw          0.815836
                            Standardized      0.808206

                       Cronbach Coefficient Alpha with Deleted Variabl
                      Raw Variables             Standardized Variables


                           Correla-                Correla-
                 Deleted  tion with               tion with
                 Variable   Total      Alpha        Total      Alpha      Label
                ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
                                         ƒƒƒƒƒƒƒƒƒƒƒ
                 CRI2_01   -.106391   0.824370   -.117827   0.821956   cri2_01
                 CRI2_02   0.201823   0.814864   0.176187   0.809214   cri2_02
                 CRI2_03   0.280976   0.812357   0.269409   0.805024   cri2_03
                 CRI2_04   0.765257   0.791034   0.766489   0.781409   cri2_04
                 CRI2_05   -.106391   0.824370   -.139250   0.822857   cri2_05
                 CRI2_06   0.765257   0.791034   0.766489   0.781409   cri2_06
                 CRI2_07   0.443376   0.805534   0.423210   0.797949   cri2_07
                 CRI2_08   0.690412   0.798951   0.664913   0.786412   cri2_08
                 CRI2_09   0.529629   0.800271   0.518662   0.793454   cri2_09
                 CRI2_10   -.273526   0.838547   -.252589   0.827561   cri2_10
                 CRI2_11   0.498087   0.802297   0.516984   0.793534   cri2_11
                 CRI2_12   0.322139   0.811395   0.307313   0.803299   cri2_12
                 CRI2_13   0.699294   0.794074   0.689362   0.785216   cri2_13
                 CRI2_14   0.765257   0.791034   0.766489   0.781409   cri2_14
                 CRI2_15   0.271781   0.814019   0.293075   0.803948   cri2_15
                 CRI2_16   0.846512   0.783933   0.851875   0.777130   cri2_16
                 CRI2_17   0.692078   0.791202   0.700026   0.784693   cri2_17
                 CRI2_18   0.611315   0.796471   0.625044   0.788351   cri2_18
                 CRI2_19   -.177627   0.834356   -.192228   0.825069   cri2_19
                 CRI2_20   0.199188   0.815987   0.203809   0.807980   cri2_20
                 CRI2_21   0.020153   0.825438   -.003166   0.817071   cri2_21
                 CRI2_22   0.443376   0.805534   0.470516   0.795731   cri2_22
                 CRI2_23   0.199188   0.815987   0.215163   0.807471   cri2_23
                 CRI2_24   0.280976   0.812357   0.297003   0.803769   cri2_24
                 CRI2_25   -.030557   0.822068   -.048887   0.819032   cri2_25
```

7.14 RELIABILITY OF COMPOSITE SCORES BASED ON COEFFICIENT ALPHA

In reality, tests rarely meet the assumptions required of strictly parallel forms. Therefore, a framework is needed for estimating composite reliability when the model of strictly parallel tests is untenable. Estimating the composite reliability of scores in the case of essentially tau-equivalent or congeneric tests is accomplished using the variance of the composite scores and all of the covariance components of the subtests (or individual items if one is working with a single test). An estimate is provided that is analogous to coefficient alpha and is simply an extension from the item-level data to subtest level data structures. *Importantly, alpha provides a lower bound to the estimation of reliability in the situation where tests are nonparallel*. The evidence that coefficient alpha provides a lower bound estimate of reliability is established as follows. First, there will be at least one subtest of those comprising a composite variable that exhibits a variance greater than or equal to its covariance with any other of the subtests. Second, for any two tests that are not strictly parallel, the sum of their true score variances is greater than or equal to twice their covariance. Finally, the sum of the true score variance for nonparallel tests (k) will be greater than or equal to the sum of their $k(k - 1)$ covariance components divided by $(k - 1)$. Application of the inequality yields Equation 7.25.

Equation 7.25. Reliability of a composite equivalent to coefficient alpha

$$\rho_{CC'} \geq \frac{k}{k-1}\left(1-\frac{\sum\hat{\sigma}_i^2}{\hat{\sigma}_C^2}\right)$$

- $\rho_{CC'}$ = reliability of the composite.
- $\sum\hat{\sigma}_i^2$ = variance for subtest i.
- $\hat{\sigma}_C^2$ = total composite test variance.

Küder–Richardson Formulas 20 (KR_{20}) and 21 (KR_{21})

In 1937, Küder and Richardson developed two formulas aimed at solving the problem of the lack of a unique solution provided by the split-half method of reliability estimation. Specifically, the Küder–Richardson approaches are based on item-level statistical properties rather than the creation of two parallel half tests. The two formulas developed, KR_{20} and KR_{21}, are numbered according to the steps involved in their derivation. Both KR_{20} and KR_{21} are closely related to coefficient alpha. In fact, the two formulas can be viewed as more restrictive versions of coefficient alpha. For example, the KR_{20} formula is only applicable to dichotomously (correct/incorrect) scored items (Equation 7.26).

To explain, notice that the numerator inside the brackets of Equation 7.26 is the sum of the product of the proportion of persons correctly responding to each item on the

Equation 7.26. Küder–Richardson formula 20

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{\hat{\sigma}_X^2} \div \right)$$

- KR_{20} = coefficient alpha.
- k = number of items.
- pq = variance of item i as the product of the proportion of correct and proportion incorrect responses over persons.
- $\hat{\sigma}_X^2$ = total test score variance.

test multiplied by the proportion of persons responding incorrectly to each item on the test. Comparing Equation 7.24 for coefficient alpha, we see that the numerator within the brackets involves summation of the variance of all test items. The primary difference between the two equations is that in KR_{20} the variance for dichotomous items is based on multiplying proportions, whereas in coefficient alpha the derivation of item variance is not restricted to multiplying the proportion correct times the proportion incorrect for an item because items are allowed to be scored on an ordinal or interval level of measurement (e.g., Likert-type scales or continuous test scores on an interval scale). Finally, where all test items are of equal difficulty (e.g., the proportion correct for all items are equal), the KR_{21} formula applies and is provided in Equation 7.27.

For a detailed exposition of the KR_{20}, KR_{21}, and coefficient alpha formulas with sample data, see the Excel file titled "Reliability_Calculation_Examples.xlsx" located on the companion website (*www.guilford.com/price2-materials*).

Equation 7.27. Küder–Richardson formula 21

$$k\,R_{21} = \frac{k}{k-1}\left[1 - \frac{\hat{\mu}(k-\hat{\mu})}{k\hat{\sigma}_X^2}\right]$$

- k = number of items.
- $\hat{\mu}$ = total score on the test.
- $\hat{\sigma}_X^2$ = total test score variance.

7.15 RELIABILITY ESTIMATION USING THE ANALYSIS OF VARIANCE METHOD

Another useful and general approach to estimating the reliability of test scores is the analysis of variance (Hoyt, 1941). Consider the formulas for coefficient alpha, KR_{20} and KR_{21}. Close inspection reveals that the primary goal of these formulas is the partitioning of (1) variance attributed to individual items and (2) total variance collectively contributed by all items on a test. Similarly, in the analysis of variance (ANOVA), one can partition the variance among persons and items, yielding the same result as coefficient alpha. The equation for the ANOVA method (Hoyt, 1941) is provided in Equation 7.28.

To illustrate Equation 7.28 using example data, we return to the data used in the examples for coefficient alpha. Restructuring the data file as presented in Table 7.14 ensures the correct layout for running ANOVA in SPSS. Note that Table 7.14 only provides a partial listing of the data (because there are 25 items on the test) used in the example results depicted in Table 7.15.

The data layout example in Table 7.14 continues until all persons, items, and scores are entered. Next, the following SPSS syntax is used to produce the mean squares required for calculation of the reliability coefficient.

SPSS syntax to produce Table 7.15

```
UNIANOVA score BY person item
/METHOD=SSTYPE(3)
/CRITERIA=ALPHA(.05)
/DESIGN=person item person*item.
```

Inserting the mean squares for persons and the person by items interaction yields a reliability coefficient of .82—the same value as that which resulted using the formula for coefficient alpha. Applying the person and person by item mean squares to the ANOVA approach yields $\rho_{XX'}$ = .847 – .156/.847 = .82.

Equation 7.28. ANOVA method for estimating the coefficient of reliability

$$\rho_{XX'} = \frac{MS_{\text{persons}} - MS_{\text{persons*items}}}{MS_{\text{persons}}}$$

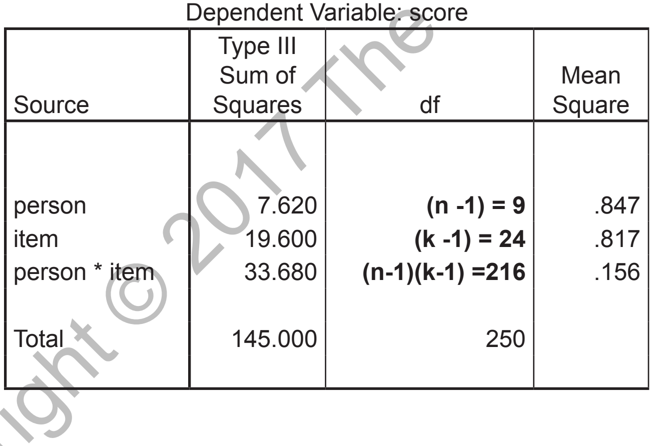
- $\rho_{XX'}$ = coefficient of reliability.
- MS_{persons} = variability attributed to persons.
- $MS_{\text{persons*items}}$ = variability attributed to persons and items together.

TABLE 7.14. Data Layout for Reliability Estimation Using SPSS ANOVA

Person	Item	Score
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	0
7	1	1
8	1	1
9	1	1
10	1	1

Note. Table consists of 10 persons, the first item out of 25, and persons' scores on item 1.

TABLE 7.15. ANOVA Output: Tests of Between-Subjects Effects



Dependent Variable: score

Source	Type III Sum of Squares	df	Mean Square
person	7.620	**(n -1) = 9**	.847
item	19.600	**(k -1) = 24**	.817
person * item	33.680	**(n-1)(k-1) =216**	.156
Total	145.000	250	

7.16 RELIABILITY OF DIFFERENCE SCORES

An important aspect of score reliability for certain types of research relates to how change over time affects the reliability of scores (Linn & Slinde, 1977; Zimmerman & Williams, 1982; Rogosa, Brandt, & Zimowski, 1982). For example, consider the case where a difference score based on fluid intelligence and crystallized intelligence is of interest for diagnostic reasons. Although the primary research question may be about whether the change in score level is statistically different, a related question focuses on how reliability is affected by the change in score level. To address the previous question, we consider the reliability of change scores as a function of (1) the reliability of the original scores used for computation of the difference score, and (2) the correlation between the scores obtained on the two tests. Based on these two components, the usefulness of calculating the reliability of change scores depends on the psychometric quality of the measurement instruments.

The research design of a study plays a crucial role in the application and interpretation of the reliability of change scores. For example, if groups of subjects selected for a study are based on a certain range of pretest score values, then the difference score will be a biased estimator of reliable change (e.g., due to restricted range of pretest scores). Elements of the research design also play an important role when using change scores. For example, random assignment to study groups provides a way to make inferential statements that are not possible when studying intact groups. Equation 7.29 provides the formula estimating the reliability of difference scores based on pretest to posttest change. Note that Equation 7.29 incorporates all of the elements of reliability theory presented thus far in this chapter. Within the true score model, one begins with the fact that it is theoretically possible to calculate a difference score. Given this information, the usual true score algebraic manipulation (i.e., true scores to observed scores) applies. Equation 7.29 illustrates the reliability of difference scores.

To illustrate the use of Equation 7.29, we use crystallized (serving as test 1) and fluid intelligence (serving as test 2) subtest total scores. In Equation 7.30, application of our score data is applied.The following information is obtained from the GfGc.sav dataset and is based on the total sample (N = 1,000).

Equation 7.29. Reliability of difference scores

$$\rho_{DD'} = \frac{\hat{\rho}_{X_1X_1'}\sigma^2_{X_1'} + \hat{\rho}_{X_2X_2'}\sigma^2_{X_2} - 2\rho_{X_1X_2}\sigma_{X_1}\sigma_{X_2}}{\sigma^2_{X_1} + \sigma^2_{X_2} - 2\rho_{X_1X_2}\sigma_{X_1}\sigma_{X_2}}$$

- $\rho_{DD'}$ = reliability of a difference score.
- $\hat{\rho}_{X_1X_1}$ = reliability of test 1.
- $\hat{\rho}_{X_2X_2}$ = reliability of test 2.
- $\sigma^2_{X_1}$ = variance of scores on test 1.
- $\sigma^2_{X_2}$ = variance of scores on test 2.
- $2\rho_{X_1X_2}$ = two times the correlation between tests 1 and 2.
- $\sigma_{X_1}\sigma_{X_2}$ = product of the standard deviation of test 1 and test 2.
- $\hat{\rho}_{X_1X_1'}$ = reliability of test 1.
- $\hat{\rho}_{X_2X_2}$ = reliability of test 2.

Equation 7.30. Application of the equation for the reliability of difference scores using statistics in Table 7.16

$$\rho_{DD'} = \frac{.95(502.21) + .89(129.50) - 2(.463)(22.41)(11.38)}{502.21 + 129.50 - 2(.463)(22.41)(11.38)}$$

$$= \frac{477.10 + 115.25 - (.926)(255.02)}{631.71 - (.926)(255.02)}$$

$$= \frac{592.35 - 236.15}{631.71 - 236.15}$$

$$= \frac{356.20}{395.56}$$

$$= .90$$

TABLE 7.16. Descriptive Statistics and Reliability Estimates for Crystallized and Fluid Intelligence Tests

	Crystallized intelligence subtest total score (test 1)	Fluid intelligence subtest total score (test 2)
Mean	81.57	33.00
Standard deviation	22.41	11.38
Reliability	0.95	0.89

7.17 APPLICATION OF THE RELIABILITY OF DIFFERENCE SCORES

To ensure the existence of highly reliable difference scores, the following conditions should be present. Both tests (i.e., scores) should exhibit high reliability but be correlated with each other at a low to moderate level (e.g., .30–.40). This situation produces reliability of difference scores that are high. Finally, the psychometric quality of the tests used to derive difference scores for the analysis of change is crucial to produce reliable change scores. The concept of the reliability of change scores over time can also be extended beyond the analysis of discrepancy between different constructs (e.g., crystallized and fluid intelligence presented here) or basic pretest to posttest analyses to analyze change over time. For example, analytic techniques such as longitudinal item response theory (IRT; covered in Chapter 10) and hierarchical linear and structural equation modeling provide powerful frameworks for the analysis of change (Muthen, 2007; Zimmerman, Williams, & Zumbo, 1993; Raudenbush, 2001; Card & Little, 2007).

7.18 ERRORS OF MEASUREMENT AND CONFIDENCE INTERVALS

Reliability has been presented so as to provide information regarding the consistency or stability of test scores. Alternatively, it is also useful to view how "unreliable" test scores are. Such unreliability is regarded as a discrepancy between observed scores and true scores and is expressed as the error of measurement relative to scores on a test. In this section, three different approaches to deriving estimates of errors of measurement are presented along with the interpretation of each using example data. These three approaches are from Lord and Novick (1968, pp. 67–68). The first technique presented is the **standard error of measurement**, $\hat{\sigma}_{X.T} = \hat{\sigma}_E = \sigma_X\sqrt{1-\hat{\rho}_{XX'}}$, and is based on the error in predicting a person's observed score given the person's true score on randomly parallel tests. The second technique is the **standard error of estimation**, $\hat{\sigma}_{T.X} = \sigma_X\sqrt{\hat{\rho}_{XX'}(1-\hat{\rho}_{XX'})}$, and is based on the error in predicting a person's true score from his or her observed score. It is useful for establishing confidence limits and intervals for *true scores* based on observed scores (i.e., based on the standard deviation of the errors of estimation of true score given an observed score). The third technique is the **standard error of prediction**, $\hat{\sigma}_{Y.X} = \sigma_Y\sqrt{1-\rho^2_{XX'}}$, and is useful for predicting scores on test form Y from parallel test form X. The next section provides application of the SEM and the standard error of prediction.

7.19 STANDARD ERROR OF MEASUREMENT

The standard error of measurement (*SEM*; $\hat{\sigma}_E$) provides an estimate of the discrepancy between a person's true score and observed score on a test of interest. Measurement error for test scores is often expressed in standard deviation units, and the *SEM* indexes the *standard deviation of the distribution of measurement error*. Formally, the *SEM* ($\hat{\sigma}_E$) is defined as the standard deviation of the discrepancy between a person's true score and observed score over infinitely repeated testing occasions. Gulliksen (1950b, p. 43) offered an intuitive definition of the *SEM* as "the error of measurement made in substituting the observed score for the true score." Equation 7.31 illustrates the standard error of measurement.

Equation 7.31. Population *SEM*

$$\hat{\sigma}_E = \sigma_X\sqrt{1-\hat{\rho}_{XX'}}$$

- $\hat{\sigma}_E$ = population standard error of measurement.
- σ_X = observed score population standard deviation.
- $\hat{\rho}_{XX'}$ = coefficient of reliability based on scores on a test.

When applying Equation 7.31 to score data, sample estimates rather than population parameters are typically used to estimate the *SEM*.

The *SEM* provides a single index of measurement error for a set of test scores. It can be used for establishing **confidence limits** and developing a **confidence interval** around a person's *observed score given the person's estimated true score*. Within classical test theory, a person's true score is *fixed* (or constant), and it is the observed and error scores that randomly fluctuate over repeated testing occasions (Lord & Novick, 1968, p. 56). One can derive confidence limits and an associated interval for observed scores using the *SEM*. However, because a person's true score is of primary interest in the true score model, one should first estimate the true score for a person prior to using Equation 7.31 to derive confidence intervals.

Two problems occur when not accounting for true score: (1) a regression effect (i.e., the imperfect correlation between observed and true scores, which produces a regression toward the group mean), and (2) the impact of **heteroscedastic** (nonuniform) errors across the score continuum (Nunnally & Bernstein, 1994, p. 240). Consequently, simply using the *SEM* has the effect of overcorrecting owing to larger measurement error in observed scores as compared to true scores. Confidence intervals established without estimating true scores will lack symmetry (i.e., lack the correct precision across the score scale) around observed scores. To address the issue of regression toward the mean due to errors of measurement, Stanley (1970), Nunnally and Bernstein (1994), and Glutting, McDermott, and Stanley (1987) note that one should first estimate *true scores* for a person and then derive estimated true score–based confidence intervals that can be used with observed scores. This step, illustrated in Equation 7.32, overcomes the problem of lack of symmetry from simply applying the *SEM* to derive confidence intervals for observed scores.

As an example, consider estimating a true score for a person who obtained an observed score of 17. Returning to Tables 7.3 and 7.4, we see that the mean is 11.50, the

Equation 7.32. Estimated true score derived using a deviation-based observed score multiplied by the reliability estimate corrected in relation to the group mean

$$\hat{T} = \hat{\rho}_{XX'}(X_i - \bar{X}_j) + \bar{X}_j$$

- $\hat{T}$ = estimated true score.
- X_i = observed score for a person.
- $\bar{X}_j$ = mean score for a group of persons.
- $X_i - \bar{X}_j$ = deviation score for person *i*.
- $\hat{\rho}_{XX'}$ = coefficient of reliability.

standard deviation of observed scores is 4.3, and the reliability is .82. Application of this information to Equation 7.33 provides the following result.

As noted earlier, lack of symmetry for confidence intervals derived with an *SEM* without first estimating true scores neglects accounting for a regression effect. The regression effect causes biased scores either upward or downward, depending on their location relative to the group mean. For example, high observed scores are typically further away from the mean of the group (i.e., they exhibit an upward bias effect), and low scores are typically biased downward lower than the actual observed score. For these reasons, *it is correct to establish confidence intervals or probable ranges for a person's observed score given their (fixed or regressed) true score*. Using the estimated true score for a person, one can apply Equation 7.33 to Equation 7.34a to derive a symmetric confidence interval for true scores that can be applied to a person's *observed scores*. Equation 7.34a can be expressed as $\hat{\sigma}_{X.T}$ to show that applying the *SEM* to estimated true scores yields the prediction of

Equation 7.33. Estimated true score expressed as a regressed observed score using reliability of .82, observed score of 17, and group mean of 11.50

$$\begin{aligned}\hat{T} &= .82(17-11.5)+11.5\\ &= .82(5.5)+11.5\\ &= 4.51+11.5\\ &= 16.01\end{aligned}$$

Equation 7.34a. *SEM* expressed as the prediction of observed score on true score

$$\hat{\sigma}_{X.T} = \sigma_X\sqrt{1-\hat{\rho}_{XX'}}$$

- $\hat{\sigma}_{X.T}$ = standard error of measurement as the prediction of observed score from true score.
- σ_X = observed score population standard deviation.
- $\hat{\rho}_{XX'}$ = coefficient of reliability based on scores on a test.

Equation 7.34b. Illustration of Equation 7.34a

$$
\begin{aligned}
\hat{\sigma}_{X.T} &= \sigma_X\sqrt{1-\hat{\rho}_{XX'}}\\
&= 4.3\sqrt{1-.82}\\
&= 4.3(.42)\\
&= 1.82
\end{aligned}
$$

- $\hat{\sigma}_{X.T}$ = standard error of measurement as the prediction of observed score from true score.
- σ_X = observed score population standard deviation.
- $\hat{\rho}_{XX'}$ = coefficient of reliability based on scores on a test.

observed scores from true scores. The resulting confidence intervals will be symmetric about a person's true score but asymmetric about their observed score. *This approach to developing confidence intervals is necessary in order to account for regression toward the mean test score.*

Equation 7.35a provides the following advantages. First, Stanley's method is based on a score metric that *is expressed in estimated true score units* (i.e., $\hat{T} - T'$, the T' = predicted true score) (Glutting et al., 1987). Second, as Stanley demonstrated (1970), his

Equation 7.35a. Stanley's method for establishing confidence limits—expressed in true score units—based on estimated true scores

$$\hat{T} \pm (z)(\hat{\sigma}_{X.T})(\hat{\rho}_{XX'})$$

- $\hat{T}$ = estimated true score.
- z = standard normal deviate (e.g., 1.96).
- $\hat{\sigma}_{X.T}$ = standard error of measurement as the prediction of observed score from true score.
- $\hat{\rho}_{XX'}$ = coefficient of reliability.

Equation 7.35b. Application of Stanley's method for establishing a 95% confidence interval for observed scores based on estimated true score of 16.01

$$\begin{aligned}&\hat{T} \pm (1.96)(1.82)(.82)\\&= (1.96)(1.5)\\&= 16.01 \pm 2.94\\&= 13.07 - 18.95\end{aligned}$$

- $\hat{T}$ = estimated true score (16.01).
- z = standard normal deviate (e.g., 1.96).
- $\hat{\sigma}_{X.T}$ = standard error of measurement as the prediction of observed score from true score (1.82).
- $\hat{\rho}_{XX'}$ = coefficient of reliability (.82).

method adheres to the classical true score model assumption that states, *for a population of examinees, errors of measurement exhibit zero correlation with true scores.*

Interpretation

To facilitate understanding that a person's true score will fall within a confidence interval based on that person's observed score, consider the following scenario. First, using the previous example, let's assume that a person's true score is 16, the reliability is .82, and the standard error of measurement is 1.82. Next, let's assume that this person is repeatedly tested 1,000 times. Of the 1,000 repeated testing occasions, 950 (95%) would lie within 2.94 points of their true score (e.g., between 13.07 and 18.95). Fifty scores would fall outside of the interval 13.07 to 18.95. Finally, if a confidence interval is derived for each of the person's 1,000 observed scores, 950 of the intervals would be generated around observed scores between 13.07 and 18.95 (each interval would contain the person's true score). From the previous explanation, we see that 5% of the time the person's true score would not fall within the interval 13.07 to 18.95. However, there is a 95% chance that the confidence interval generated around the observed score of 16 will contain the person's true score.

A common alternate approach to establishing confidence limits and intervals offered by Lord and Novick (1968, pp. 68–70) does not always meet the classical true score model requirement of zero correlation between true and error scores—unless the reliability of the test is perfect (i.e., 1.0). Lord and Novick's (1968, p. 68) approach is expressed in *obtained score units* (e.g., $\hat{T} - T$) and is provided in Equation 7.36a.

Continuing with Lord and Novick's approach, we will next illustrate the probability that a person's true score will fall within a confidence interval based on their observed score. Again, we assume that a person's true score is 16 and that the standard error of measurement is 1.82. Next, let's assume that this person is repeatedly tested 1,000 times. Of the 1,000 repeated testing occasions, 950 (95%) would lie within 3.25 points of their true score (e.g., between 12.76 and 19.26). Notice that the confidence interval is wider in Lord and Novick's method (see Equation 7.36a) because the product of the *z*-ordinate and the estimated standard error is multiplied by the *square root of the reliability*. Fifty scores would fall outside of the interval 12.76 to 19.26. Finally, if a confidence interval was derived for each of the person's 1,000 observed scores, 950 of the intervals would be generated around observed scores between 12.76 and 19.26 (each interval would contain the person's true score). It is apparent from the previous explanation that 5% of the time the person's true score would not fall within the interval 12.76 to 19.26. However, there is a 95% chance that the confidence interval generated around the observed score of 16 will contain the person's true score.

Equation 7.36a. Lord and Novick's method for establishing confidence limits—expressed in obtained score units—based on estimated true scores

$$\hat{T} \pm (z)(\hat{\sigma}_{X.T})\sqrt{\hat{\rho}_{XX'}}$$

- $\hat{T}$ = estimated true score.
- z = standard normal deviate (e.g., 1.96).
- $\hat{\sigma}_{X.T}$ = standard error of measurement as the prediction of observed score from true score.
- $\sqrt{\hat{\rho}_{XX'}}$ = square root of coefficient of reliability or the reliability index.

Equation 7.36b. Application of Lord and Novick's method for establishing a 95% confidence interval for observed scores based on estimated true score of 16.01

$$\hat{T} \pm (1.96)(1.82)(.91)$$

$$= (1.96)(1.66)$$

$$= 16.01 \pm 3.25$$

$$= 12.76 - 19.26$$

7.20 STANDARD ERROR OF PREDICTION

The standard error of prediction is useful for predicting the probable range of scores on one form of a test (e.g., *Y*), given a score on an alternate parallel test (e.g., *X*). For example, using the crystallized intelligence test example throughout this chapter, one may be interested in what score one can expect to obtain on a parallel form of the same test. To derive an error estimate to address this question, Equation 7.37a is required.

Equation 7.37a. Standard error of prediction expressed as the prediction of test score *Y* on parallel test score *X*

$$\hat{\sigma}_{Y.X} = \sigma_Y \sqrt{1 - \rho^2_{XX'}}$$

- $\hat{\sigma}_{Y.X}$ = standard error of prediction.
- σ_Y = standard deviation of test *Y*.
- $\rho^2_{XX'}$ = reliability of test *X* squared.

Equation 7.37b. Derivation of the standard error of prediction

$$\begin{aligned}\sigma_{Y.X} &= 4.3\sqrt{1 - .82^2} \\ &= 4.3\sqrt{.327} \\ &= 4.3(.572) \\ &= 2.46\end{aligned}$$

Equation 7.37c. Application of standard error of prediction for establishing a 95% confidence interval for observed scores based on an estimated true score of 16.01

$$\begin{aligned}&\hat{T} \pm (1.96)(2.46) \\ &= \ 4.82 \\ &= \ 16.01 \pm 4.82 \\ &= \ 11.19 - 20.83\end{aligned}$$

Applying the same example data as in Equations 7.32 and 7.33 to Equation 7.37a yields the error estimate in Equation 7.37b.

Next, we can apply the standard error of prediction derived from Equation 7.37c to develop a 95% confidence interval.

Interpretation

Using the standard error of prediction, the probability that a person's true score will fall within a confidence interval based on that person's observed score is illustrated next. Again we assume that a person's true score is 16, the standard deviation of test *X* is 4.3, and the reliability estimate is .82. Next, we assume that this person is repeatedly tested 1,000 times. Of the 1,000 repeated testing occasions, 950 (95%) would lie within 4.82 points of the person's true score (e.g., between 11.19 and 20.83). Notice that the confidence interval is wider in the previous examples. Fifty scores would fall outside of the interval 11.19 to 20.83. Finally, if a confidence interval was derived for each of the person's 1,000 observed scores, 950 of the intervals would be generated around observed scores between 11.19 and 20.83 (each interval would contain the person's true score). It is apparent from the previous explanation that 5% of the time the person's true score would not fall within the interval 11.19 to 20.83. However, there is a 95% chance that the confidence interval generated around the observed score of 16 will contain the person's true score.

7.21 SUMMARIZING AND REPORTING RELIABILITY INFORMATION

Summarizing and reporting information regarding measurement error is essential to the proper use of any instrument. More broadly, any assessment procedure that uses some form of instrumentation or measurement protocol for the assessment of knowledge, skill, or ability is prone to error. Ideally, the optimal way to evaluate the quality of the reliability of scores is to conduct independent replication studies that focus specifically on reliability (AERA, APA, & NCME, 1999; 2014, p. 27). The following points are essential in reporting errors of measurement: (1) sociodemographic details about the study group or examinee population, (2) sources of error, (3) magnitude of errors, (4) degree of generalizability across alternate or parallel forms of a test, and (5) degree of agreement among raters or scorers. Information on the reliability of scores may be reported in terms of one or more coefficients (depending on the use of the scores) such as (1) stability—test–retest, (2) equivalence—alternate forms, and (3) internal consistency—coefficient alpha or split-half. When decisions are based on judgment, coefficients of interscorer or rater consistency are required.

Errors of measurement and reliability coefficients involving decisions based on judgments have many sources. For example, evaluator biases, scoring subjectivity, and between-examinee factors are all sources of error. To meet these additional challenges, when errors of measurement and reliability are being reported for decisions based on judgments resulting in classifications, **generalizability theory** (Cronbach et al., 1972) provides a comprehensive (presented next in Chapter 8) framework that allows for many types of applied testing scenarios. Reliability information may also be reported in terms of error variance or standard deviations of measurement errors. For example, when test scores are based on classical test theory, the standard error of measurement should be reported along with confidence intervals for score levels. For IRT, information on functions should be reported because they provide the magnitude of error across the score range. Also, when a test is based on IRT, information on the individual item characteristic functions should be reported along with the test characteristic curve. The item characteristic and test functions provide essential information regarding the precision of measurement at various ability levels of examinees. Item response theory will be covered thoroughly in Chapter 10.

Whenever possible, reporting conditional errors of measurement is also encouraged because errors of measurement are not uniform across the score scale and this has implications for the accuracy of score reporting (AERA, APA, & NCME, 1999, p. 29). For approaches to estimating conditional errors of measurement see Kolen, Hanson, and Brennan (1992), and for conditional reliability, see Raju, Price, Oshima, and Nering (2007).

When comparing and interpreting reliability information obtained from using a test for different groups of persons, consideration should be given to differences in variability of the groups. Also, the techniques used to estimate the reliability coefficients should be reported along with the sources of error. Importantly, it is essential to present the theoretical model by which the errors of measurement and reliability coefficients were derived (e.g., classical test theory, IRT, or generalizability theory). This step is critical because interpretation of reliability coefficients varies depending on the theoretical model used for estimation.

Finally, test score precision should be reported according to the type of scale by which they have been derived. For example, raw scores or IRT-based scores may reflect different errors of measurement and reliability coefficients than standardized or derived scores. This is particularly true at different levels of a person's ability or achievement. Therefore, measurement precision is substantially influenced by the scale in which the test scores are reported.

7.22 SUMMARY AND CONCLUSIONS

Reliability refers to the degree to which scores on tests or other instruments are free from errors of measurement. This dictates their level of consistency, repeatability, or reliability. Reliability of measurement is a fundamental issue in any research endeavor because some form of measurement is used to acquire data—and no measurement process is error free. Identifying and properly classifying the type and magnitude of error is essential to estimating the reliability of scores. Estimating the reliability of scores according to the classical true score model involves certain assumptions about a person's observed, true, and error scores. Reliability studies are conducted to evaluate the degree of error exhibited in the scores on a test (or other instrument). Reliability studies involving two separate test administrations include the alternate form and test–retest methods or techniques.

The internal consistency approaches are based on covariation among or between test item responses and involve a single test administration using a single form. The internal consistency approaches include (1) split-half techniques with the Spearman–Brown correction formula, (2) coefficient alpha, (3) the Küder–Richardson 20 formula, (4) the Küder–Richardson 21 formula, and (5) the analysis of variance approach. The reliability of scores used in the study of change is an issue important to the integrity of longitudinal research designs. Accordingly, a formula was presented that provides a way to estimate the reliability of change scores.

It is also useful to view how "unreliable" test scores are. The unreliability of scores is viewed as a discrepancy between observed scores and true scores and is expressed as the error of measurement. Three different approaches to deriving estimates of errors of measurement and associated confidence intervals were presented, along with the interpretation of each using example data. The three approaches commonly used are (1) the standard error of measurement, (2) the standard error of estimation, and (3) the standard error of prediction.

KEY TERMS AND DEFINITIONS

Attributes. Identifiable qualities or characteristics represented by either numerical elements or categorical classifications of objects that can be measured.

Classical test theory. Based on the true score model, a theory concerned with observed, true, and error score components.

Classical true score model. A model-based theory of properties of test scores relative to populations of persons based on true, observed, and error components. Classical test theory is based on this model.

Coefficient alpha. An estimate of internal consistency reliability that is based on item variances and covariances and that does not require strictly parallel or true score equivalence between its internal components or half tests. The alpha coefficient is the mean of all possible randomly split-half tests using Rulon's formula. In relation to theoretical or true score estimates of reliability, alpha produces a lower-bound estimate of score reliability.

Coefficient of equivalence. Calculated as the correlation between scores on two administrations of the same test.

Coefficient of reliability. The ratio of true score variance to observed score variance.

Coefficient of stability. Correlation coefficient between scores on two administrations of the same test on different days; calculated using the test–retest method.

Composite score. The sum of responses to individual items where a response to an item is a discrete number.

Confidence interval. A statistical range with a specified probability that a given parameter lies within the range.

Confidence limits. Either of two values that provide the endpoints of a confidence interval.

Congeneric tests. Axiom specifying that a person's observed, true, and error scores on two tests are allowed to differ.

Constant error. Error of measurement that occurs systematically and constantly due to characteristics of the person, the test, or both. In the physical or natural sciences, this type of error occurs by an improperly calibrated instrument being used to measure something such as temperature. This results in a systematic shift based on a calibration error.

Deviation score. A raw score subtracted from the mean of a set of scores.

Essential tau-equivalence. Axiom specifying that a person's observed score random variables on two tests are allowed to differ but only by the value of the linking constant.

Generalizability theory. A highly flexible technique for studying error that allows for the degree to which a particular set of measurements on an examinee are generalizable to a more extensive set of measurements.

Guttman's equation. An equation that provides a derivation of reliability estimation equivalent to Rulon's method that does not necessarily assume equal variances on the half-test components. This method does not require the use of the Spearman–Brown correction formula.

Heteroscedastic error. A condition in which nonuniform or nonconstant error is exhibited in a range of scores.

Internal consistency. Determines whether several items that propose to measure the same general construct produce similar scores.

Item homogeneity. Test items composed of similar content as defined by the underlying construct.

Küder–Richardson Formula 20 (KR-20). A special case of coefficient alpha that is derived when items are measured exclusively on a dichotomous level.

Küder–Richardson Formula 21 (KR-21). A special case of coefficient alpha that is derived when items are of equal difficulty.

Measurement precision. How close scores are to one another and the degree of measure of error on parallel tests.

Parallel tests. The assumption that when two tests are strictly equal, true score, observed, and error scores are the same for every individual.

Random error. Variability of errors of measurement function in a random or nonsystematic manner.

Reliability. The consistency of measurements based on repeated sampling of a sample or population.

Reliability coefficient. The squared correlation between observed scores and true scores. A numerical statistic or index that summarizes the properties of scores on a test or instrument.

Reliability index. The correlation between observed scores and true scores.

Rulon's formula. A split-half approach to reliability estimation that uses difference scores between half tests and that does not require equal error variances on the half tests. This method does not require the use of the Spearman–Brown correction formula.

Spearman–Brown formula. A method in which tests are correlated and corrected back to the total length of a single test to assess the reliability of the overall test.

Split-half reliability. A method of estimation in which two parallel half tests are created, and then the Spearman–Brown correction is applied to yield total test reliability.

Standard error of estimation. Used to predict a person's score on one test (Y) based on his or her score on another parallel test (X). Useful for establishing confidence intervals for predicted scores.

Standard error of measurement. The accuracy with which a single score for a person approximates the expected value of possible scores for the same person. It is the weighted average of the errors of measurement for a group of examinees.

Standard error of prediction. Used to predict a person's true score from his or her observed score. Useful for establishing confidence intervals for true scores.

Tau-equivalence. Axiom specifying that a person has equal true scores on parallel forms of a test.

True score. Hypothetical entity expressed as the expectation of a person's observed score over repeated independent testing occasions.

True score model. A score expressed as the expectation of a person's observed score over infinitely repeated independent testing occasions. True score is only a hypothetical entity due to the implausibility of actually conducting an infinite number of independent testing occasions.

Validity. The degree to which evidence and theory support the interpretations of test scores entailed by proposed use of a test or instrument. Evidence of test validity is related to reliability, such that reliability is a necessary but not sufficient condition to establish the validity of scores on a test.

Guilford Publications
370 Seventh Avenue, Suite 1200
New York, NY 10001
212-431-9800
800-365-7006
www.guilford.com