CHAPTER 3

# RELIABILITY

by Dimiter Dimitrov
George Mason University

In B. T. Erford (Ed.) *Assessment for Counselors*
2007 Boston: Houghton-Mifflin/Lahaska Press
(pp. 99-122)

# *Reliability*

## Chapter Outline

## What is Reliability?

Measurements in counseling, education, and related behavioral fields are not completely accurate and consistent. There is always some error involved due to person's conditions (e.g., mood, fatigue, and momentary distraction) and/or external conditions such as noise, temperature, light, etc., that may randomly occur during the measurement process. The instrument of measurement (e.g., tests, inventories, or raters) may also affect the accuracy of the scores (observations). For example, it is unlikely that the scores of a person on two different forms of an anxiety test would be equal. Also, different scores are likely to be assigned to a person when different counselors evaluate a specific attribute of this person. In another scenario, if a group of persons take the same test twice within a short period of time, one can expect the rank order of their scores on the two test administrations to be somewhat similar, but not exactly the same. In other words, one can expect a relatively high, yet not perfect, positive correlation of test-retest scores for the group of examinees. Inconsistency occurs also in different criterion-referenced classifications (e.g., pass-fail or mastery- nonmastery) based on measurements obtained through testing or subjective expert judgments.

In measurement parlance, the higher the accuracy and consistency of measurements (scores, observations), the higher their reliability. Thus, the *reliability* of measurements indicates the degree to which the measurements are *accurate*, c*onsistent*, and *repeatable* when (a) different people conduct the measurement, (b) using different instruments that purport to measure the same trait (e.g., proficiency, attitude, anxiety), and (c) there is incidental variation in measurement conditions. The reliability is a key condition for quality measurements with tests, inventories, or individuals (raters, judges, observers, etc.). Most importantly, reliability is a necessary (yet, not sufficient) condition for the *validity* of measurements. To remind, validity has to do meaningfulness, accuracy, and appropriateness of interpretations and decisions based on measurement data.

It is important to note that reliability refers to the measurement data obtained with an instrument and not to the instrument itself. Previous studies and recent editorial policies of professional journals (e.g., Dimitrov, 2002; Sax, 1980; Thompson & Vacha-Haase, 2000) emphasize that it is more accurate to talk about reliability of measurement data than reliability of tests (items, questions, and tasks). Tests cannot be accurate, stable, or unstable, but observations (scores) can. Therefore, any reference to "reliability of a test" should be interpreted to mean the "reliability of measurement data derived from a test".

## CLASSICAL MODEL OF RELIABILITY

### True Score

Measurements with performance tests, personality inventories, expert evaluations, and even physical measurements, are not completely accurate, consistent, and replicable. For example, although the height of a person remains constant throughout repeated measurements within a short period of time (say, 15 minutes) using the same scale, the observed values would be scattered around this constant due to imperfection in the visual acuity of the measurer (same person or somebody else). Thus, if $T$ denotes the person's constant (*true*) height, then the observed height, $X$, in any of the repeated measurements will deviate from $T$ with an *error of measurement*, $E$. That is,

$$X = T + E. \tag{1}$$

To grasp what is meant by *true score* in classical test theory, imagine that a person takes a standardized intelligence test each day for 100 days in a raw. The person would likely obtain a number of different observed scores over these occasions. The mean of all observed scores would represent an approximation of the person's true score, $T$, on the standardized intelligence test. In general, the true score, $T$, is the mean of the theoretical distribution of $X$ scores that would be observed in repeated independent measurements of the same person with the same test. Evidently, the true score, $T$, is a hypothetical concept because it is not practically possible to test the same person infinity times in independent repeated measurements (i.e., each testing does not influence any subsequent testing).

It is important to note that the error in Equation 1 is assumed to random in nature. Possible *sources of random error* are: (1) fluctuations in the mood or alertness of persons taking the test due to fatigue, illness, or other recent experiences, (2) incidental variation in the measurement conditions due, for example, to outside noise or inconsistency in the administration of the instrument, (3) differences in scoring due to factors such as scoring errors, subjectivity, or clerical errors, and (4) random guessing on response alternatives in tests or questionnaire items. Conversely, *systematic errors* that remain constant from one measurement to another do not lead to inconsistency and, therefore, do not affect the reliability of the scores. Systematic errors will occur, for example, when a counselor X assigns two points lower than a counselor Y to each person in the stress evaluation of a group of individuals. So, again, *the reliability of any measurement is the extent to which the measurement results are free of random errors*.

**Classical Definition of Reliability**

Equation 1 represents the classical assumption that any *observed score*, X, consists of two parts*: true score*, *T*, and *error of measurement*, *E*. Because errors are random, it is assumed that they do not correlate with the true scores (i.e., $r_{TE} = 0$). Indeed, there is no reason to expect that persons with higher true scores would have systematically larger (or systematically smaller) measurement errors than persons with lower true scores. Under this assumption, the following is true for the variances of observed scores, true scores, and errors for a population of test-takers:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2, \qquad (2)$$

i.e., the observed score variance is the sum of true score variance and error variance. Given this*, the reliability of measurement indicates what proportion of the observed score variance is true score variance*. The analytic translation of this definition is

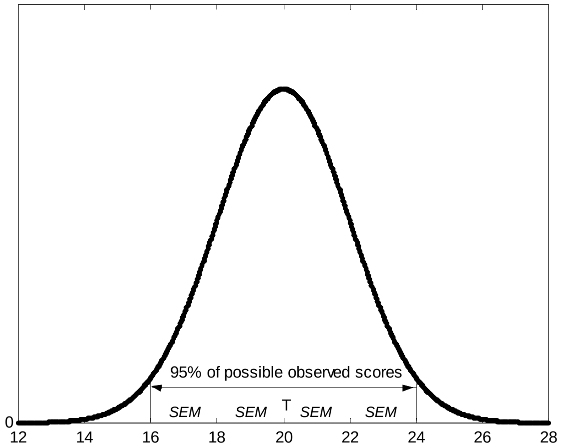
$$r_{xx} = \frac{\sigma_T^2}{\sigma_X^2}. \qquad (3)$$

**Note**: The notation for reliability, $r_{xx}$, stems from the equivalent definition that the reliability is also the correlation between the observed scores on two *parallel tests* (i.e., tests with equal true scores and equal error variances for every population of persons taking the two tests; e.g., Allen & Yen, p. 73). The reliability can also be represented as the *squared correlation between observed scores and true scores*: $r_{XX} = r_{XT}^2$.

Any definition of reliability (e.g., Equation 3) implies that the reliability may take values from 0 to 1, with $r_{xx} = 1$ indicating *perfect reliability* - this is possible only when the total observed score variance is true score variance ($\sigma_X^2 = \sigma_T^2$) or, equivalently, when the error variance is zero ($\sigma_E^2 = 0$). The closer $r_{xx}$ to zero, the lower the score reliability.

**Standard Error of Measurement (*SEM*)**

Classical test theory also assumes that (a) the distribution of observed scores that a person may have under repeated independent testings is normal and (b) the standard deviation of the normal distribution, referred to as *standard error of measurement* (*SEM*), is the same for all persons taking the test. Under these assumptions, Figure 1 represents the (hypothetical) normal distribution of observed scores for repeated measurements of one person with the same test. The mean of this distribution is, in fact, the person's true score ($T = 20$) and the standard deviation is the standard error of measurement ($SEM = 2$).



``````

**Figure 1.** Normal distribution of observed scores for repeated testings of one person.
``````

Based on basic statistical properties for normal distributions, Figure 1shows that (a) almost all possible observed scores for this person are expected to fall in the interval from $T - 3(SEM)$ to $T - 3(SEM)$, which in this case is from 14 to 26, and (b) about 95% of these observed scores are expected to fall in the interval from $T - 2(SEM)$ to $T + 2(SEM)$, which in this case is from 16 to 24. The latter property may be used reversely to construct (approximately) a 95% confidence interval of a person's true score, $T$, given the observed score, $X$, of the person in a real testing:

$$X - 2(SEM) < T < X + 2(SEM) \tag{4}$$

For example, if $X = 23$ is the person's observed score in a single real testing, then his/her true score is expected (with about 95% confidence) to be in the interval from $X - 2(SEM)$ to $X + 2(SEM)$. Thus, with $X = 23$ and $SEM = 2$, the 95% confidence interval for the true score of this person is from 23 – 2(2) to 23 + 2(2), i.e., from 19 to 27.

Evidently, smaller *SEM* will produce smaller confidence intervals for the person's true score thus leading to higher accuracy of measurement. Given that *SEM* is inversely related to reliability, one can infer that the higher the reliability, the higher the accuracy of measurements. As some previous studies indicate, however, although the reliability coefficient is a convenient unitless number between 0 and 1, the *SEM* relates directly to the meaning of the original scale of measurement (e.g., number-right correct answers) and is therefore more useful for score interpretations (e.g., Feldt & Brennan, 1989; Thissen, 1990). Using the following equation, one can determine the *SEM* from the reliability, $r_{xx}$, and the standard deviation of the observed scores:

$$SEM = \sigma_X \sqrt{1 - r_{XX}}\,. \tag{5}$$

(*Note*: One can easily derive Equation 5 from Equations 2 and 3, taking into account that $SEM = \sigma_E$.)

For example, if the reliability is .90 and the standard deviation of the persons' observed scores is 5, then the standard error of measurement is: $SEM = 5\sqrt{1-.9} = 5(.3162) = 1.581.$

**Caution**: The concept of SEM is based on two assumptions: (a) *normality* – the distribution of possible observed scores under repeated independent testings is normal and (b) *homoscedasticity* – the standard deviation of this normal distribution of possible observed scores is the same for all persons taking the test. These assumptions, however, are generally not true, particularly for persons with true scores that are far away (higher or lower) from the average true score for a sample of examinees. Therefore, results based on the classical *SEM* (e.g., confidence intervals for true scores) should be perceived only as overall rough estimations and interpreted with caution. There are more sophisticated (yet, mathematically more complex) measurement methods that provide higher accuracy in estimating (conditional) standard errors of measurement for persons with different true scores. Brief notes on such methods are provided later in this chapter.

### Standard Error of Estimation

As shown in the previous section, $X \pm 2SEM$ is a 95% confidence interval for the true score, *T*, of a person, given the person's observed score, *X*, and the standard error of measurement, *SEM* (see Equation 5). Still within classical test theory, an estimation of a person's true score from his/her observed score can be obtained by simply regressing *T* on *X*. In fact, the regression coefficient in predicting *T* from *X* is equal to the reliability of the test scores, $r_{XX}$ (Lord & Novick, 1968, p. 65). Specifically, if μ is the population mean of test scores, the regression equation for estimating true scores from observed scores is

$$\hat{T} = r_{XX}X + (1 - r_{XX})\mu. \quad (6)$$

**Note**. Equation 6 shows also that the estimated (predicted) true score, $\hat{T}$, is closer to the observed score when the reliability, $r_{XX}$, is high and, conversely, closer to the mean, μ, when the reliability is low. In the extreme cases, (a) $\hat{T} = X$, with perfectly reliable scores ($r_{XX} = 1$), and (b) $\hat{T} = \mu$, with totally unreliable scores ($r_{XX} = 0$).

All persons with the same observed score, *X*, will have the same predicted true score, $\hat{T}$, obtained with Equation 6, but not necessarily the same actual true scores, *T*. The standard deviation of the estimation error ($\varepsilon = T - \hat{T}$) is referred to as *standard error of estimation*, $\sigma_{\varepsilon}$ (or, *SEE*), and is evaluated as follows:

$$SEE = \sigma_X \sqrt{r_{XX}(1 - r_{XX})}, \tag{7}$$

where $\sigma_X$ is the standard deviation of the observed scores and $r_{XX}$ is the score reliability.

The standard error of estimation, obtained with Equation 7, is always smaller than the standard error of measurement, obtained with Equation 5 (*SEE* < *SEM*). Thus, when estimating true scores is of primary interest, the regression approach (Equation 6) provides more accurate estimation of a person's true score, *T*, compared to confidence intervals for *T* based on *SEM*.

**Caution**. Keep in mind that the *SEM* is an overall estimate of differences between observed and true scores (*X* – *T)*, whereas the *SEE* is an overall estimate of differences between actual and predicted true scores ($T - \hat{T}$). Also, the estimation of true scores using Equation 6 requires information about the population mean of observed scores, μ, (or at least the sample mean, $\overline{X}$, for a sufficiently large sample), whereas obtaining confidence intervals for true scores using *SEM* (e.g., $X \pm 2SEM$) does not require such information.

**Example 1**. Given the standard deviation, $\sigma_X = 5$, and the reliability, $r_{XX} = .91$, for the observed scores, *X*, one can use Equation 5 to obtain the *SEM* and Equation 7 for the *SEE*. Specifically, $SEM = 5\sqrt{1 - (.91)^2} = 2.073$ and $SEE = 5\sqrt{(.91)(1 - .91)} = 1.431$; (note that *SEE* < *SEM*). If the observed score for a person is *X* = 30, the interval $X \pm 2SEM$ (in this case, 30 ± 2 x 2.073), shows that the true score of this person is somewhere between 25.854 and 34.146. Given the mean of the observed scores (say, μ = 25), a more accurate estimate of the true score is obtained using Equation 6: $\hat{T} = (.91)(30) + (1 - .91)(25) = 29.55$ (why is the true score estimate, $\hat{T}$, close to the observed score, *X* = 30, in this example?)

# TYPES OF RELIABILITY

The reliability of test scores for a population of examinees is defined as the ratio of their true score variance to observed score variance (Equation 3). Equivalently, the reliability can also be represented as the squared correlation between true and observed scores (e.g., Allen & Yen, 1979, p. 73). In empirical research, however, true scores cannot be directly determined and therefore the reliability is typically estimated by coefficients of internal consistency, test-retest, alternate forms, and other types of reliability estimates adopted in the measurement literature. It is important to note that different types of reliability relate to different sources of measurement error and, contrary to some common misconceptions, are generally not interchangeable.

## Internal Consistency

*Internal consistency* estimates of reliability are based on the average correlation among items within a test or scale. A widely known method for determining internal consistency of test scores yields a *split-half reliability* estimate. With this method, the test is split into two halves which are assumed to be parallel (i.e., the two halves have equal true scores and equal error variances). The score reliability of the whole test is estimated then by the *Spearman-Brown formula*:

$$r_{XX} = \frac{2r_{12}}{1 + r_{12}}, \tag{8}$$

where $r_{12}$ is the Pearson correlation between the scores on the two halves of the test. For example, if the correlation between the two test halves is 0.6, then the split-half reliability estimate is: $r_{XX}$ = 2(0.6)/(1+0.6) = 0.75.

One commonly used approach to forming test halves, called the *odd/even* method, is to assign the odd-numbered test items to one half and the even-numbered test items to the other half of the test. A more recommended approach, called *matched random subsets*, involves three steps. First, two statistics are calculated for each item: (a) the proportion of individuals who answered the item correctly and (b) the point-biserial correlation between the item and the total test score. Second, each item is plotted on a graph using these two statistics as coordinates of a dot representing the item. Third, items that are close together on the graph are paired and one item from each pair is randomly assigned to one half of the test.

The Spearman-Brown formula is not appropriate when there are indications that the test halves are not parallel (e.g., when the two test halves do not have equal variances). In such cases, the internal consistency of the scores for the whole test can be estimated with the Cronbach's *coefficient α* (Greek letter alpha) using the formula (Cronbach, 1951):

$$\alpha = \frac{2[\mathrm{VAR}(X) - \mathrm{VAR}(X_1) - \mathrm{VAR}(X_2)]}{\mathrm{VAR}(X)}, \tag{9}$$

where VAR($X$), VAR($X_1$), and VAR($X_2$), represent the sample variance of the whole test, its first half, and its second half, respectively. For example, it the observed score variance for the whole test is 40 and the observed variances for the two test halves are 12 and 11, respectively, then coefficient alpha is: $\alpha = 2(40 - 12 - 11)/40 = 0.85$.

**Caution**. With *speed* tests, the split-half correlation coefficient would be close to zero since most examinees would answer correctly almost all items in the first half and (running out of time) will miss most items in the second half of the test.

In the general case, the coefficient α is calculated for more than two components of the test. Each test component is an item or a set of items. The formula for coefficient α is simply an extension of Formula 9 for more than two components of the test.

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum \mathrm{VAR}(X_i)}{\mathrm{VAR(X)}}\right], \tag{10}$$

where $n$ is the number of test components,

$X$ is the observed score for the whole test,

$X_i$ is the observed score on the $i$th test component (i.e., $X = X_1 + X_2 + \ldots + X_n$),

VAR($X$) is the variance of $X$,

VAR($X_i$) is the variance of $X_i$, and

Σ (Greek capital letter "sigma") is the summation symbol.

If each test component is a dichotomous item (1= correct, 0 = incorrect), the coefficient α can be calculated by an equivalent formula, called *Kuder-Richarson formula 20*, with the notation *KR20* (or *α-20*) for the coefficient of internal consistency:

$$KR20 = \frac{n}{n-1}\left(1 - \frac{\sum p_i(1-p_i)}{\mathrm{VAR}(X)}\right), \tag{11}$$

where $n$ is the number of dichotomous test items,

$X$ is the observed score for the whole test,

VAR($X$) is the variance of $X$

$p_i$ is the proportion of persons who answered correctly item $i$,

$p_i(1 - p_i)$ is the variance of the observed binary scores on item $i$ ($X_i$ = 1 or 0),

that is VAR($X_i$) = $p_i(1 - p_i)$.

**Example 2**: Table 1 illustrates the calculation of the *KR20* coefficient (Formula 11) for the observed scores of 50 persons on a test of four dichotomous items ($n$ = 4) given that the variance of the total observed scores on the test is 1.82 [i.e., VAR($X$) = 1.82].

**Table 1**

| Item | $N_i$ | $p_i$ | $1 - p_i$ | $p_i(1 - p_i)$ |
|---|---|---|---|---|
| 1 | 7 | 7/50 = .14 | .86 | .14 x .86 = .1204 |
| 2 | 12 | 12/50 = .24 | .76 | .24 x .76 = .1824 |
| 3 | 18 | 18/50 = .36 | .64 | .36 x .64 = .2304 |
| 4 | 13 | 13/50 = .26 | .74 | .26 x .74 = .1924 |

Summation: Σ $p_i(1 - p_i)$ = 0.7526

$$KR20 = \frac{4}{3}\left(1 - \frac{0.7526}{1.82}\right) = .782$$

*Note*. $N_i$ = number of persons responding correctly on item $i$ ($i$ = 1, 2, 3, 4)

**Caution:** It is important to note that coefficient α (or *KR20*) is an accurate estimate of reliability, $r_{XX}$, only if there is no correlation among measurement errors and the test components (if not parallel) are at least *essentially tau-equivalent*. By definition, test components are *essentially tau-equivalent* if the persons' true scores on the components differ by a constant. Tau-equivalency implies also that the test components measure the same trait (e.g., anxiety) and their true scores have equal variances in the population of respondents. When measurement errors do not correlate, but the test components are not essentially tau-equivalent, coefficient α will underestimate the actual reliability ($\alpha < r_{XX}$). If, however, the measurement errors with some test components correlate, coefficient α may substantially overestimate the reliability ($\alpha > r_{XX}$). Correlated errors may occur, for example, (a) with items related to a common stimulus (e.g., same paragraph or graph) and (b) with tests presented in a speeded fashion.

### Test-Retest Reliability

When test developers and practitioners are interested is assessing the extent to which persons consistently respond to the same test, inventory, or questionnaire administered on different occasions, this is a question of *test-retest reliability* (stability) of test data. Test-retest reliability is estimated by the correlation between the observed scores of the same people taking the same test twice. The resulting correlation coefficient is referred to also as *coefficient of stability*.

The major problem with test-retest reliability estimates is the potential for carry-over effects between the two test administrations. Readministration of the test within a short period of time (e.g., a few days or weeks) may produce carry-over effects due to memory and/or practice. For example, students who take a history test may look up some answers they were unsure of after the first administration of the test thus changing their true knowledge on the history content measured by the test. Likewise, the process of completing an anxiety inventory could trigger an increase in the anxiety level of some people thus causing their true anxiety scores to change from one administration of the inventory to the next.

If the construct (attribute) being measured varies over time (e.g., cognitive skills, depression), a long period of time between the two administrations of the instrument may produce carry-over effects due to biological maturation, cognitive development, changes in information, experience, and/or moods. Thus, test-retest reliability estimates are most appropriate for measurements of traits that are stable across the time period between the two test administrations (e.g., visual or auditory acuity, personality, and work values). In addition to carry-over effect problems with estimates of test-retest reliability, there is also a practical limitation to retesting because it is usually time-consuming and/or expensive. Therefore, retesting solely for the purpose to estimate score stability may be impractical.

**Caution**: *Test-retest reliability* and *internal consistency* are independent concepts. Basically, they are affected by different sources of error and, therefore, it may happen that measures with low internal consistency have high temporal stability and vice versa. Previous research on stability showed that the test-retest correlation coefficient can serve well as a surrogate for the classical reliability coefficient if an essentially tau-equivalent test model with equal error variances or a parallel test model is present (Tisak & Tisak, 1996).

### Alternate Form Reliability

If two versions of an instrument (test, inventory, or questionnaire) have very similar observed-score means, variances, and correlations with other measures, they are called *alternate forms* of the instrument. In fact, any decent attempt to construct parallel tests is expected to result in alternate test forms as it is practically impossible to obtain perfectly parallel tests (i.e., equal true scores and equal error variances). Alternate forms usually are easier to develop for instruments that measure, for example, intellectual abilities or specific academic abilities than those that measure constructs that are more difficult to represent with measurable variables (e.g., personality, motivation, temperament, anxiety).

*Alternate form reliability* is a measure of the consistency of scores on alternate test forms administered to the same group of individuals. The correlation between observed scores on two alternate test forms, referred to also as *coefficient of equivalence*, provides an estimate of the reliability of either one of the alternate forms. Estimates of alternate form reliability are subject to carry-over effects as test-retest reliability coefficients, but in lesser degree due to the fact that the persons are not tested twice with the same items. A recommended rule-of-thumb is to have a 2-week time period between administrations of alternate test forms.

Whenever possible, it is important to obtain both internal consistency coefficients and alternate forms correlations for a test. If the correlation between alternate forms is much lower than the internal consistency coefficient (e.g., a difference of 0.20 or more), this might be due to (a) differences in content, (b) subjectivity of scoring, and (c) changes in the trait being measured over time between the administrations of alternate forms. To determine the relative contribution of these sources of error, it is usually recommended to administer the two alternate forms on the same day for some respondents and then within a 2-week time interval for others. If the correlation between the scores on the alternate forms for the same-day administration is much higher than the correlation for the 2-week time interval, then variation in the trait being measured is a major source of error. For example, it is likely that measures of mood will change over a 2-week time interval and thus the 2-week correlation will be lower than the same-day correlation between the alternate forms of the instrument. However, if the two correlations are both low, the persons' scores may be stable over the 2-week time interval but the alternate forms probably differ in content.

**Caution**: When scores on alternate forms of an instrument are assigned by raters (e.g., counselors or teachers), one may check for scoring subjectivity by using a three-step procedure: (1) randomly split a large sample of persons, (2) administer the alternate forms on the same day for one group of people, and (3) administer the alternate forms within a 2-week time interval for the other group of people. If the correlations between raters are high for both groups, there is probably little scoring error due to subjectivity. If the correlation over the 2-week time interval and the same-day correlation are both consistently low across different raters, it is difficult to determine the major sources of scoring errors. Such errors can be reduced by training the raters in using the instrument and providing clear guidelines for scoring behaviors or traits being measured.

## Criterion-Referenced Reliability

Criterion-referenced measurements report how the examinees stand with respect to an external criterion. The criterion is usually some specific educational or performance objective such as “know how to apply basic algebra rules” or “being able to recognize patterns”. Because a criterion-referenced test may cover numerous specific objectives (criteria), each objective should be measured as accurately as possible. When the results of criterion-referenced measurements are used for dichotomous classifications related to mastery or nonmastery of the criterion, the reliability of such classifications is often referred to as *classification consistency*. This type of reliability shows the consistency with which classifications are made, either by the same test administered on two occasions or by alternate test forms.

Two classical indices of classification consistency are (a) $P_o$ = the observed proportion of persons consistently classified as masters/nonmasters and (b) Cohen’s κ (Greek letter *kappa*) = the proportion of nonrandom consistent classifications. Their calculation is illustrated for the two-way data layout in Table 2 where the entries are proportions of persons classified as masters/nonmasters by two alternate test forms of a criterion-referenced test (Test A and Test B). Specifically, $p_{11}$ is the proportion of persons classified as masters by both test forms, $p_{12}$ is the proportion of persons classified as masters by test A and nonmasters by test B, and so on. Also, $P_{A1}$ , $P_{A2}$, $P_{B1}$, and $P_{B2}$ are notations for marginal proportions, that is: $P_{A1} = p_{11} + p_{12}$, $P_{B1} = p_{11} + p_{21}$, and so on. Thus, the observed proportion of consistent classifications (masters/nonmasters) is

$$P_o = p_{11} + p_{22} \tag{12}$$

**Table 2.**

*Contingency Table for Mastery-Nonmastery Classifications*

| | | Test B | | |
|---|---|---|---|---|
| | | Master | Nonmaster | |
| Test A | Master | $p_{11}$ | $p_{12}$ | $P_{A1}$ |
| | Nonmaster | $p_{21}$ | $p_{22}$ | $P_{A2}$ |
| | | $P_{B1}$ | $P_{B2}$ | |

However, $P_o$ can be a misleading indicator of classification consistency because part of it may occur by chance. Cohen's *kappa* takes into account the proportion of consistent classification that is (theoretically) expected to occur by chance, $P_e$, and provides a ratio of nonrandom consistent classifications

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{13}$$

where $P_e$ is (theoretically) the sum of the crossproducts of the marginal proportions in Table 2: $P_e = P_{A1}P_{B1} + P_{A2}P_{B2}$. In Formula 13, the numerator ($P_o$ - $P_e$) is the proportion of nonrandom consistent classification being detected, whereas the denominator (1 - $P_e$) is the maximum proportion of nonrandom consistent classification that may occur. Thus, Cohen's *kappa* indicates what proportion of the maximum possible nonrandom consistent classifications is found with the data.

**Example 3.** Let us use specific numbers for the proportions in Table 2: $p_{11}$ = 0.3, $p_{12}$ = 0.2, $p_{21}$ = 0.1, and $\mathrm{p}_{22}$ = 0.4. The marginal proportions are: $P_{A1}$ = 0.3 + 0.2 = 0.5, $P_{A2}$ = 0.1 + 0.4 = 0.5, $P_{B1}$ = 0.3 + 0.1 = 0.4, and $P_{B2}$ = 0.2 + 0.4 = 0.6. With these data, the observed proportion of consistent classification is $P_o$ = 0.3 + 0.4 = 0.7 (Formula 12).

The proportion of consistent classifications that may occur by chance in this hypothetical example is: $P_e = (0.5)(0.4) + (0.5)(0.6) = 0.5$. Using Formula 13, the Cohen's kappa ratio is: $\kappa = (0.7 - 0.5)/(1 - 0.5) = 0.2/0.5 = 0.4$. Thus, the initially obtained 70% of observed consistent classifications ($P_o = 0.7$) is reduced to 40% consistent classifications after taking into account the proportion of consistent classifications that may occur by chance. Given the conservative estimation provided by *kappa*, it is reasonable to report that the classification consistency is somewhere between .40 and .70 (i.e., between $\kappa$ and $P_o$).

**Note**: For practical purposes, it is recommended to report both $P_o$ and the Cohen's *kappa* as the latter is very conservative thus underestimating the actual rate of consistent classifications. Previous research (e.g., Chase, 1996; Peng & Subkoviak, 1980) provides additional procedures for estimating classification consistency, including scenarios with a single test administration or prior to the initial application of the test.

### Inter-rater Reliability

The chances of measurement error usually increase when the scores are based on subjective judgments of the person(s) doing the scoring. Such situations occur, for example, with classroom assessment of essays or portfolios where the teacher is, in fact, the "instrument" of assessment. In another scenario, involving some projective tests of personality, the scorer (e.g., counselor or psychotherapist) should decide if the person's responses suggest normal functioning or some form of psychopathology. Also, subjective judgments of raters (experts, judges) are often used for classification purposes (e.g., to determine a "minimum level of competency" in pass/fail decisions). The person doing the scoring is referred to in this section as a *rater* (scorer, expert, judge). In all cases of rater scoring, it is important to estimate to the degree to which the scores are unduly affected by the subjective judgment of the rater(s). Such estimation is provided by coefficients of *inter-rater reliability* (called also inter-scorer reliability, inter-judge reliability, or inter-rater agreement).

Depending on the context of measurement, there are different methods of estimating inter-rater reliability. Most frequently used classical measures of inter-rater reliability are the Person correlation coefficients and the Cohen's *kappa* coefficient (or some extended versions of *kappa*). For example, the two indexes of classification consistency illustrated with Table 2 (observed proportion of consistent classification, $P_{\text{o}}$, and Cohen's *kappa*) can be used as estimates of inter-rater reliability if two raters (instead of two test forms) classify persons as masters or nonmasters. One can use Formula 13 to calculate Cohen's kappa when persons (or their products) are classified by two raters into more than two categories, but $P_{\text{o}}$ and $P_e$ should be calculated with a contingency table for the respective number of categories. Thus, with classifications into three categories (e.g., low, medium, and high performance), we have: $P_{\text{o}} = p_{11} + p_{22} + p_{33}$ and $P_e = P_{\text{A}1}P_{\text{B}1} + P_{\text{A}2}P_{\text{B}2} + P_{\text{A}3}P_{\text{B}3}$.

If, however, two raters independently assign scores to portfolios of students, then the Pearson correlation coefficient for the two sets of scores can be used as an estimate of inter-rater agreement. The higher the correlation coefficient, the lower the error variance due to scorer differences and thus the higher the inter-rater agreement.

When scoring with alternate forms of a measurement instrument is done by (two or more) raters, one can check for measurement error due to subjectivity of scoring by administering the alternate forms (a) one the same day for one group of subjects and (b) with a 2-week delay for another group of subjects. If the correlations between raters are high for both groups, there is probably little error due to subjectivity of scoring. If, however, the correlation over the 2-week time interval and the same-day correlation are both consistently low across different raters, it is difficult to say what is the major source of unreliability (subjectivity of scoring or, say, differences in content for the two alternate forms of the instrument). The inter-rater reliability can be improved by training the raters in the use of the instrument and providing clear guidelines for scoring.

## RELIABILITY OF COMPOSITE SCORES

In many situations, scores from two or more scales are combined into *composite scores* to measure and interpret a more general dimension (trait or proficiency) related to these scales. Composite scores are often used with test battery for achievement, intelligence, aptitude, depression, or eating disorders. For example, the scores on nine scales (factors) with the Symptom Checklist-90-Revised (SCL-90-R; Derogatis, …) are combined into three "global" (composite) scores in measuring current psychological symptom status. One frequently reported composite score is the sum of verbal and quantitative scores of the Graduate Record Examination (GRE).

Although the composite score may be simply the sum of several scale scores, its reliability is usually not just the mean of the reliabilities for the scales being combined. The issue of reliability estimation for composite scores is addressed in this section when the composite score is (a) the sum of two scale scores, (b) the difference (*gain*) score for pretest to posttest measurements, and (c) the sum of three or more scale scores.

**Reliability of Sum of Scores**

Let us have two scale scores, $X_1$ and $X_2$, and a composite score which is the sum of these two scores: $Y = X_1 + X_2$. For example, with the GRE scoring, the composite score is the sum of the verbal and quantitative scores. The formula for estimating the *reliability of the composite score*, $r_{YY}$, is a special case (for two scale scores) of a more general formula provided in pervious research (e.g., Nunnally & Bernstein, 1994, p. 268):

$$r_{YY} = 1 - \frac{\sigma_1^2(1 - r_{11}) + \sigma_2^2(1 - r_{22})}{\sigma_Y^2}, \tag{14}$$

where $\sigma_1^2$ is the variance of $X_1$, that is: $\sigma_1^2$ = VAR($X_1$)],

$\sigma_2^2$ is the variance of $X_2$ , that is: $\sigma_2^2$ = VAR($X_2$),

$\sigma_Y^2$ is the variance of the composite score Y, that is: $\sigma_Y^2$ = VAR($Y$),

$r_{11}$ is the reliability of $X_1$, and

$r_{22}$ is the reliability of $X_2$.

**Example 4**: The estimation of the reliability for a composite score, Y = $X_1 + X_2$, is illustrated in this example with data from a real study on attitudes and behaviors of students related to their sexual activities. Specifically, $X_1$ is the score on a scale labeled "Love as Justification for Sexual Involvement" and $X_2$ is the score on a scale labeled "Sex for Approbation". With the notations adopted in Formula 12, the following results were obtained from the study data for (a) the variances of $X_1$, $X_2$, and $Y$: $\sigma_1^2 = 13.750$, $\sigma_2^2 = 10.433$, $\sigma_Y^2 = 38.5992$ and (b) the reliabilities of $X_1$ and $X_2$: $r_{11} = .8334$, $r_{22} = .8217$.

Replacing these components for their values in Formula 12, we obtain:

$$r_{YY} = 1 - \frac{13.750(1-.8334) + 10.433(1-.8217)}{38.592} = .892.$$

Thus, the reliability estimate of the composite score $Y$ (.892) in this example is higher than the reliability estimates of its components, $X_1$ (.8334) and $X_2$ (.8217). This, however, is not always the case.

**Caution**: Although not explicitly present in Formula 14, the correlation between $X_1$ and $X_2$, denoted hereafter $r_{12}$, affects the reliability of the composite score; (in the above example, $r_{12} = .598$). In fact, when $X_1$ and $X_2$ do not correlate ($r_{12} = 0$), the reliability of their sum (Y $= X_1 + X_2$) is the average of their reliabilities: $r_{YY} = (r_{11} + r_{22})/2$.

In many cases, the scores that are combined into a composite score come from scales with different units of measurement (e.g., 3-point and 5-point survey scales). Therefore, to present the measurements on a common scale (and for some technical reasons), the raw scores are often converted into standard scores (z- scores) before being summed. This is done, for example, with the raw scores of the primary psychological symptoms measured with the self-report symptom inventory SCL-90-R. For the special case of standard (*z*-) scores, Formula 14 is converted into a much simpler equivalent form

$$r_{YY} = 1 - \frac{2 - (r_{11} + r_{22})}{\sigma_{Y_z}^2}, \tag{15}$$

where $\sigma^2_{Y_z}$ is the variance of the sum of the *z*-scores for $X_1$ and $X_2$ (i.e., $Y_z = z_1 + z_2$),

$r_{11}$ is the reliability of $X_1$, and

$r_{22}$ is the reliability of $X_2$.

For the data in Example 4, $\sigma^2_{Y_z}$= 3.203 and, as before, $r_{11}$ = .8334, $r_{22}$ = .8217. With this, using Formula 15, we obtain the same value for the reliability of the composite score $Y = X_1$ and $X_2$ (or, equivalently, for $Y_z = z_1 + z_2$):

$$r_{YY} = 1 - \frac{2 - (.8334 + .8217)}{3.203} = .892.$$

**Note**. Formula 15 follows directly from Formula 14, taking into account that the variance of the standard (*z*-) scores for any variable is 1 and, thus, $\sigma^2(z_1) + \sigma^2(z_2) = 2$.

Formulas 14 and 15 can be readily extended when the composite score is a sum of more than two scale scores. For example, when $Y = X_1 + X_2 + X_3$, the reliability of the composite score *Y* can be estimated by extending Formula 15 as follows:

$$r_{YY} = 1 - \frac{3 - (r_{11} + r_{22} + r_{33})}{\sigma^2_{Y_z}}, \qquad (16)$$

where $\sigma^2_{Y_z}$ is the variance of the sum of the standard (*z*-) scores for $X_1$, $X_2$, and $X_3$, that is $Y_z = z_1 + z_2 + z_2$ ; ($r_{11}$, $r_{22}$, and $r_{33}$ are the reliabilities for $X_1$, $X_2$, and $X_3$, respectively).

**Reliability of Difference Scores**

The difference between two observes scores for the same person, called *difference score*, is widely used in behavioral research primarily (a) to measure the person's growth across time points and (b) to compare the person's scores on academic, psychological, or personality variables. For example, measurement of *change* using the person's difference (or *gain*) score from pretest to posttest is used to assess the effect of specific educational programs, counseling treatments, and rehabilitation services or allied health interventions.

Clearly, the quality of the results and the validity of interpretations in studies on change and profile analysis depend, among other things, on the reliability of difference scores.

Technically, the difference of two scores, $Y = X_2 - X_1$, is a composite score of the sum $Y = X_2 + (- X_1)$. Therefore, the reliability of the difference score, $Y$, can be estimated with Formula 14 (or its z-score version, Formula 15).

**Example 5:** As in Example 4, the data in this example also come from the study on attitudes and behaviors of students related to their sexual activities. However, instead of summing the scores on two scales, the composite score is now the difference (gain) from pretreatment to posttreatment measurements on a scale labeled "Self-affirmation", that is, $Y = X_2 - X_1$, where $X_1$ is the pretreatment score and $X_2$, the posttreatment score on this scale. With these data, the variance of the difference $Y_z = z_2 - z_1$ (where $z_1$ and $z_2$ are the standard values for $X_1$ and $X_2$ ) was $\sigma^2_{Y_z} = 0.786$. The Cronbach's alpha reliability estimates for $X_1$ and $X_2$ were $r_{11} = .8282$ and $r_{22} = .8374$, respectively. Using Formula 15, the reliability of the difference scores is

$$r_{YY} = 1 - \frac{2 - (.8282 + .8374)}{0.786} = .575.$$

Evidently, the reliability of the difference score (.575) is smaller than the reliability of the scores entering the difference (.8282 and .8374). As noted earlier, the reliability of the difference score, $r_{YY}$, is (implicitly) influenced by the correlation between $X_1$ and $X_2$ (in this case, $r_{12} = .606$) because this correlation affects the value of $\sigma^2_{Y_z}$ in Formula 15.

**Caution**. The use of difference (gain) scores in measurement of change has been criticized because of the (generally false) assertion that the difference between scores is less reliable than the score themselves (e.g., Cronbach & Furby, 1970; Linn & Slindle, 1977; Lord, 1956). This assertion is true, however, if the prettest scores and the posttest scores have equal variances and equal reliability. When this is not the case, which may happen in many situations, the reliability of the gain score is reasonably high (e.g., Overall & Woodward, 1975; Zimmerman & Williams, 1982). The relatively low reliability of gain scores does not preclude valid testing of the null hypothesis of zero mean gain score in a population of examinees, but it is not appropriate to correlate the gain score with other variables for these examinees (Mellenbergh, 1999). An important practical implication is that, without ignoring the caution urged by some authors, researchers should not always discard gain score and should be aware when gain scores are useful.

**Reliability of Weighted Sums**

Let the scores from two tests, $X_1$ and $X_2$, have different "weights" (say $w_1$ and $w_2$, respectively) in a composite score, $Y = w_1 X_1 + w_2 X_2$. To estimate the reliability of the composite score, $Y$, given the reliabilities of $X_1$ and $X_2$, one can (for simplicity) use the weighted composite score, $Yz$, of the standardized variables $Z_1$ and $Z_2$ which are obtained by transforming the raw scores of $X_1$ and $X_2$ into *z*- scores. That is,

$$Yz = w_1 Z_1 + w_2 Z_2 .$$

With this, the reliability of the composite score, $Y$ (or $Yz$ ) is given by the formula

$$r_{YY} = 1 - \frac{(1 - r_{11})w_1^2 + (1 - r_{22})w_2^2}{\sigma_{Yz}^2}, \qquad (17)$$

where $r_{YY}$ is the reliability of the composite score $Y$ (or $Yz$),

$r_{11}$ is the reliability of $X_1$,

$r_{22}$ is the reliability of $X_2$, and

$\sigma_{Yz}^2$ is the variance of the composite score $Yz$ (the weighed sum of $Z_1$ and $Z_2$).

**Example 6**. The reliability estimates (e.g., Cronbach's *alpha* coefficients) for the scores from two tests, $X_1$ and $X_2$, are $r_{11}$ = .72 and $r_{22}$ = .80, respectively. The scores on the two tests are summed into a composite score, $Y$, with $X_1$ given an importance of 40 percent ($w_1$ = 0.4) and $X_2$, an importance of 60 percent ($w_2$ = 0.6): $Y = (0.4)X_1 + (0.6)X_2$.

After transforming the scores on $X_1$ and $X_2$ into *z*- scores to obtain the standardized variables $Z_1$ and $Z_2$, respectively, the variance of $Yz = (0.4)Z_1 + (0.6)Z_2$ was found to be $\sigma^2_{Yz}$ = 1.27. Using Formula 17 with this information, the reliability of the composite score *Y* is

$$r_{YY} = 1 - \frac{(1-.72)(0.4)^2 + (1-.80)(0.6)^2}{1.27} = .908.$$

Formula 17 can be easily extended to estimate the reliability of a weighted sum for the scores of more than two tests. In case of three tests, Formula 17 for the reliability of the composite score $Y = w_1X_1 + w_2X_2 + w_3X_3$ extends to

$$r_{YY} = 1 - \frac{(1 - r_{11})w_1^2 + (1 - r_{22})w_2^2 + (1 - r_{33})w_3^2}{\sigma^2_{Yz}}, \qquad (18)$$

where $\sigma^2_{Yz}$ is the variance of $Yz = w_1Z_1 + w_2Z_2 .+ w_3Z_3$. Formulas 17 and 18 (as well as their extensions for more than three tests) apply equally well when some of the weights are negative numbers.

## DEPENDABILITY IN GENERALIZABILITY THEORY

Generalizability theory (GT) is an extension of classical measurement theory and takes into account all available error sources (*facets*), such as items, raters, test forms, and occasions, that influence the reliability for either relative (norm-referenced) or absolute (criterion-referenced) interpretations (e.g., Brennan, 2001; Shavelson & Web, 1991). Classical test theory estimates only one source of error at a time and provides estimates of reliability only for relative decisions. Indeed, both coefficients of internal consistency (e.g., Cronbach's alpha) and Pearson correlations for test-retest (or alternate forms) reliability are based on the *relative* standing of persons to each other on the measurement scale. Thus, classical reliability coefficients do not provide information on dependability of the *absolute* performance of a person regardless of relative performance of this person compared to other persons. Such information is provided, instead, with dependability coefficients in GT by estimating separately multiple sources of measurement error in a single analysis. Generally, this is done by representing the overall error variance as a sum of variance components related to different sources of measurement error using statistical methods in the framework of analysis of variance (ANOVA).

**Caution**. As noted earlier, the *true score* in classical theory is the mean of observed scores that a person may obtain under numerous independent administrations of the same test. As a person may have different true scores for different sets of items, the classical true score theory does not provide information about how generalizable the person's score is over a "universe" of admissible test items. Thus, the test items represent a potential source of error in generalization referred to as measurement *facet*. In GT, the universe of person's admissible observations can be defined also by facets such as raters and occasions taken together. The accuracy of generalizing from a person's observed score in a measurement to his/her *universe score* under all admissible testing conditions (e.g., items, raters, and occasions) is referred to in GT as *dependability*.

## Dependability with One-Facet Crossed Design

Let us examine the traditional measurement scenario in which persons are scored on the items of an instrument. In GT, this is referred to as *persons-by-items* (*p* x *i*) crossed design (ach person is scored on each item). Persons (*p*) are the *object of measurement* and items (*i*) represent a *facet* of the measurement (a potential source of generalization error). With the ANOVA analysis of the (*p* x *i*) factorial design, in which persons (*p*) and items (*i*) are two random and crossed factors, the *total variance* of the observed scores, $\sigma_X^2$, can be represented as a sum of three variance components: (a) variance for persons, $\sigma_p^2$, (b) variance for items, $\sigma_i^2$, and (c) variance for "person-by-item" interaction which is confounded with the variance of other (unaccounted for) error of measurement (*e*), $\sigma_{pi,e}$. The variance of "person-by-item" interaction (*pi*) cannot be separated from the variance of other possible sources of error because there is only one observation (score) per cell in the two-way (*p* x *i*) ANOVA data layout. The GT equation for the total observed variance as a sum of variance components with the one-facet crossed design (*p* x *i*) is

$$\sigma_X^2 = \sigma_p^2 + \sigma_i^2 + \sigma_{pi,e}^2. \tag{19}$$

A diagram presentation of Equation 19 is provided in Figure 2. As "persons" are the object of measurement, their variance component, $\sigma_p^2$, is not related to random error of measurement. The other two variance components in Equation 19, however, are error related. Specifically, the variance component $\sigma_{pi,e}^2$ is contributing to *relative error* of measurement because the interaction between persons and items (*pi*) affects the relative standings of persons on the scale of measurement (see Figure 2a). The larger the variance component $\sigma_{pi,e}^2$, the more the persons' relative standings change from item to item.

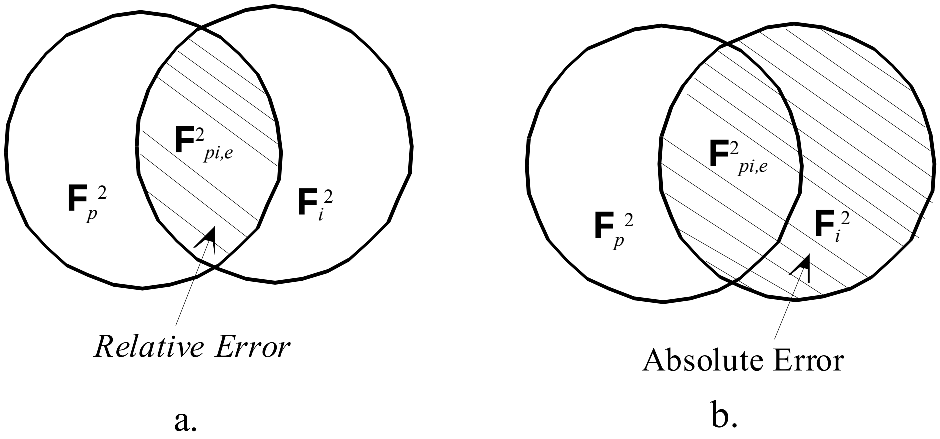


**Figure 2**. Sources of error for relative and absolute interpretations with the *p* x *i* design.

The variance component $\sigma_i^2$ indicates the degree to which the items differ in their average difficulty for all persons, but it does not reflect changes in the persons' relative standings and, therefore, does not relate to relative error of measurement. Indeed, items may vary in difficulty but the relative standing of persons may remain the same - if it does not, this would be reflected in large $\sigma^2_{pi,e}$, not in large $\sigma^2_i$.

For absolute (criterion-related) interpretations of the persons' performance, both $\sigma^2_i$ and $\sigma^2_{pi,e}$ are taken into account (see Figure 2b) because the absolute performance of a person would depend on the items chosen in the test. The item selection (more difficult or easier items), however, affects both the average item difficulty and possible changes in the persons' relative standings across items (i.e., both $\sigma^2_i$ and $\sigma^2_{pi,e}$, respectively).

Given the variance components, one can estimate the *relative error variance*

$$\sigma^2_{\mathrm{Re}l} = \frac{\sigma^2_{pi,e}}{n_i} \tag{20}$$

and the *absolute error variance*

$$\sigma^2_{Abs} = \frac{\sigma^2_i + \sigma^2_{pi,e}}{n_i}, \tag{21}$$

where $n_i$ is the number of items. It is important to note that $n_i$ can be different from the number of items used in ANOVA to estimate the variance components $\sigma^2_i$ and $\sigma^2_{pi,e}$. Thus, Equations 20 and 21 can be used to "predict" the relative and absolute error variances for a test with given number of items, $n_i$.

In GT, the score reliability for relative (norm-referenced) interpretations is referred to as *generallizability* (*G*-) *coefficient*:

$$G = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{\mathrm{Re}l}}. \tag{22}$$

A reliability-like coefficient for absolute (criterion-referenced) interpretation is referred to as *index of dependability* denoted Φ (phi):

$$\Phi = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{Abs}}. \tag{23}$$

**Caution.** With the "person-by-item" ($p$ x $i$) design, or any other one-faceted design (e.g., "person-by-rater" or "person-by-occasion"), the G-coefficient is comparable to the classical reliability coefficient, both suitable for relative (norm-referenced) interpretations only. The G-coefficient and Cronbach's *alpha* are expected to have the same value when calculated with the same data. Also, the comparison of Formulas 20 and 21 (see Figures 2a. and 2b.) shows that the relative error variance, $\sigma^2_{\mathrm{Re}l}$, is always smaller than the absolute error variance, $\sigma^2_{Abs}$. As these two error variances are in the denominator of the ratio for G and Φ (Equations 20 and 21 respectively), it follows that the dependability index is always smaller than the G-coefficient (Φ < G).

**Example 7**. Table 3 provides the scores of 20 persons on four items. The goal is first to estimate G and Φ with the data in Table 3 and then to "predict" them for a 20-item test. The procedure for achieving this goal is described in four steps.

*Step 1*: The ANOVA for the person-by-item ($p$ x $i$) design with the data in Table 3 provides the following estimates of the sample variance ("mean square", MS) for *persons* ($\mathrm{MS}_p = 0.718$), *items* ($\mathrm{MS}_i = 1.613$), and the *interaction* between them confounded with other possible sources of error ($\mathrm{MS}_{pi,e} = 0.437$).

*Step 2*: Using the rules for *expected mean squares* with the ($p$ x $i$) design (e.g., Shavelson & Webb, 1991, p. 29), we estimate the variance components (with a "hat" notation, $\hat{\sigma}^2$, indicating that these are sample estimates, not population variances):

$$\hat{\sigma}^2_p = \frac{MS_p - MS_{pi,e}}{n_i} = \frac{0.718 - 0.437}{4} = 0.0702,$$

$$\hat{\sigma}^2_i = \frac{MS_i - MS_{pi,e}}{n_i} = \frac{1.613 - 0.437}{4} = 0.2940, \text{ and}$$

$$\hat{\sigma}^2_{pi,e} = MS_{pi,e} = 0.437.$$

*Step 3*: Using Formulas 18 and 19 with the values for the variance components in Step 2, we estimate the relative and absolute error variances, respectively

$$\hat{\sigma}^2_{\mathrm{Re}\,l} = \frac{.437}{4} = 0.1093 \;;\; \hat{\sigma}^2_{Abs} = \frac{.437 + .2940}{4} = 0.1828.$$

*Step 4*: Finally, using Formulas 20 and 21, we obtain

$$G = \frac{0.0702}{0.0702 + 0.1828} = .3911 \;;\quad \Phi = \frac{0.0702}{0.0702 + 0.1828} = .2775.$$

**Table 3**

| Person | *Item 1* | *Item 2* | *Item 3* | *Item 4* |
|---|---|---|---|---|
| 1 | 2 | 3 | 5 | 5 |
| 2 | 5 | 5 | 4 | 4 |
| 3 | 4 | 3 | 4 | 4 |
| 4 | 3 | 3 | 5 | 5 |
| 5 | 3 | 3 | 4 | 5 |
| 6 | 3 | 4 | 4 | 4 |
| 7 | 4 | 5 | 5 | 5 |
| 8 | 4 | 4 | 5 | 5 |
| 9 | 4 | 5 | 5 | 5 |
| 10 | 4 | 4 | 3 | 3 |
| 11 | 4 | 4 | 5 | 5 |
| 12 | 5 | 5 | 4 | 4 |
| 13 | 4 | 4 | 4 | 4 |
| 14 | 4 | 3 | 5 | 5 |
| 15 | 4 | 4 | 5 | 5 |
| 16 | 3 | 3 | 4 | 5 |
| 17 | 4 | 5 | 4 | 4 |
| 18 | 5 | 5 | 5 | 5 |
| 19 | 5 | 5 | 4 | 4 |
| 20 | 4 | 4 | 4 | 4 |

The Cronbach's *alpha* coefficient for the data in Table 3 is $\alpha = .3911$ and thus, as expected, equal to the GT estimate of reliability for relative interpretations ($G = .3911$). This, however, is true only with the "person-by-item" design (or other one-facet designs such as "person-by-rater" or "person-by-occasion"). Let us remind, however, that the estimation of the G-coefficient is possible with more than one sources of measurement error (facets), whereas classical reliability estimates such as Cronbach's alpha or test-retest correlations are provided only with one-facet designs. Also, the GT index Φ for dependability of absolute decisions is not provided with the classical framework.

The second part of the assignment with this example relates to estimating *G* and Φ when the test consists of 20 items (ideally, parallel to the initial four items). To do this, we replace $n_i$ for 20 (instead of 4) in Steps 3 and 4 with the above calculations thus obtaining $G = .7622$ and $\Phi = .6579$. The "predicted" values of *G* and Φ will further increase with the increase of the test length (e.g., with $n_i = 40$ or $n_i = 60$).

**Note**. In classical test theory (one-faceted designs such as "person-by-item), the "prediction" of reliability related to changes (increase/decrease) in the test length is provided by the *Spearman-Brown formula*

$$r_{YY} = \frac{kr_{XX}}{1 + (k - 1)r_{XX}}, \tag{24}$$

where *k* indicates *how many times* the number of the items of a test *X* (with reliability $r_{XX}$) is increased or decreased ($k > 1$ or $k < 1$, respectively) to obtain a test *Y* with "predicted" reliability $r_{YY}$. For example, if the reliability of a test *X* is $r_{\text{XX}} = .65$, increasing the length of test *X* three times ($k$ =3) would increase the reliability for the resulting test, *Y,* to $r_{YY}$ = 3(.65)/[ 1 + 3(.65)] = .848. The Spearman-Brown formula produces classical reliability coefficients under assumptions that are difficult to satisfy in real measurements: (a) the items being added are parallel to the initial items and (b) the items have equal variances.

## Dependability with Two-Facet Crossed Design

Precision of measurements can be estimated in GT when two or more facets are taken into account. For example, if each person is evaluated by each of several raters on each of several occasions, the GT design "*person-by-rater-by-occasion*" (*p* x *r* x *o*) includes two facets, *rater* and *occasion*. The total variance of observed scores is then a sum of the variance component for persons, $\sigma_p^2$, and error related variance components for (a) raters, $\sigma_r^2$, (b) occasions, $\sigma_o^2$, (c) interaction "person-by-rater", $\sigma_{pr}^2$, (d) interaction “person-by-occasion”, $\sigma^2_{po}$, (e) interaction “rater-by-occasion”, $\sigma^2_{ro}$, and (f) interaction *person*-by-*rater*-by *occasion* confounded with other sources of error, $\sigma^2_{pro,e}$; (confounding occurs because there is only one observation for the within cell error variance with the ANOVA design “*p* x *r* x *o* “).

**Example 8**: Table 4 provides data for a GT design “*p* x *r* x *o*” where each of 10 persons is evaluated by each of three raters on each of two occasions. The estimation of reliability for relative decisions (*G*-coefficient) and dependability for absolute decisions (Φ) is illustrated in four steps.

**Table 4**

| | | Rater | | |
|---|---|---|---|---|
| Person | Occasion | 1 | 2 | 3 |
| 1 | 1 | 1 | 1 | 6 |
| | 2 | 1 | 4 | 2 |
| 2 | 1 | 4 | 7 | 6 |
| | 2 | 2 | 5 | 7 |
| 3 | 1 | 2 | 5 | 2 |
| | 2 | 5 | 6 | 6 |
| 4 | 1 | 2 | 7 | 6 |
| | 2 | 6 | 4 | 7 |
| 5 | 1 | 6 | 3 | 3 |
| | 2 | 5 | 6 | 7 |
| 6 | 1 | 4 | 3 | 3 |
| | 2 | 5 | 3 | 5 |
| 7 | 1 | 5 | 3 | 6 |
| | 2 | 5 | 2 | 2 |
| 8 | 1 | 3 | 7 | 4 |
| | 2 | 7 | 7 | 6 |
| 9 | 1 | 4 | 2 | 6 |
| | 2 | 2 | 3 | 5 |
| 10 | 1 | 3 | 2 | 2 |
| | 2 | 3 | 5 | 2 |

*Step 1*: The three-way (*person* x *rater* x *occasion*) ANOVA with the data in Table 4 provides estimates of the sample variances ("mean square", MS) for the factors *person*, *rater*, and *occasion*, as well as the interactions between them: $\mathrm{MS}_p$ = 6.891, $\mathrm{MS}_r$ = 4.067, $\mathrm{MS}_o$ = 4.817, $\mathrm{MS}_{pr}$ = 2.974, $\mathrm{MS}_{po}$ = 3.113, $\mathrm{MS}_{ro}$ = 0.067, and $\mathrm{MS}_{pro,e}$ = 2.863 (variance for the interaction between person, rater, and occasion, confounded with other possible sources of measurement error).

*Step 2*: Using the rules for *expected mean squares* with the (*p* x *i*) design (Shavelson & Webb, 1991, p. 33), we obtain estimates of the variance components. When negative variance components occur, they are set to zero [≅ 0] because this is due to sampling error (variances cannot be negative).

$$\hat{\sigma}^2_{pro,e} = MS_{pro,e} = 2.863,$$

$$\hat{\sigma}^2_{pr} = \frac{MS_{pr} - MS_{pro,e}}{n_o} = \frac{6.891 - 2.863}{2} = 2.014,$$

$$\hat{\sigma}^2_{po} = \frac{MS_{po} - MS_{pro,e}}{n_r} = \frac{3.113 - 2.863}{3} = 0.0833,$$

$$\hat{\sigma}^2_{ro} = \frac{MS_{ro} - MS_{pro,e}}{n_p} = \frac{0.067 - 2.863}{10} = -0.2796\ [\cong 0],$$

$$\hat{\sigma}^2_{p} = \frac{MS_{p} - MS_{pr} - MS_{po} + MS_{pro,e}}{n_r n_o} = \frac{6.891 - 2.974 - 3.113 + 2.863}{6} = 0.6112,$$

$$\hat{\sigma}^2_{r} = \frac{MS_{r} - MS_{pr} = MS_{ro} + MS_{pro,e}}{n_p n_o} = \frac{4.067 - 2.974 - 0.067 + 2.863}{20} = 0.1945, \text{ and}$$

$$\hat{\sigma}^2_{o} = \frac{MS_{o} - MS_{po} - MS_{ro} + MS_{pro,e}}{n_p n_r} = \frac{4.817 - 3.113 - 0.067 + 2.863}{30} = 0.150.$$

*Step 3*: Using the values for the variance components in Step 2 in the GT formulas for relative and absolute error variances with the "*p* x *r* x o" design (e.g., Shavelson & Web, 1991, p. 96), we obtain

$$\hat{\sigma}^2_{\mathrm{Re}l} = \frac{\sigma_{pr}}{n_r} + \frac{\sigma_{po}}{n_o} + \frac{\sigma_{pro,e}}{n_r n_o} = \frac{2.014}{3} + \frac{0.0833}{2} + \frac{2.863}{6} = 1.1901, \text{ and}$$

$$\hat{\sigma}^2_{Abs} = \frac{\sigma^2_r}{n_r} + \frac{\sigma^2_o}{n_o} + \frac{\sigma^2_{ro}}{n_r n_o} + \sigma^2_{\mathrm{Re}l} = \frac{0.1945}{3} + \frac{0.150}{2} + \frac{0}{6} + 1.1901 = 1.3298.$$

*Step 4*: Using Formulas 20 and 21, we obtain the *G*-coefficient for reliability of *relative* (norm-referenced) interpretations and index Φ for dependability of *absolute* (criterion-related) interpretations, respectively:

$$G = \frac{0.6112}{0.6112 + 1.1901} = .3393; \quad \Phi = \frac{0.6112}{0.6112 + 1.3299} = .3149.$$

One can "predict" the values of *G* and Φ for any number of raters and/or occasions. As an exercise, the reader may do this for 10 raters and 4 occasions, by using $n_r = 10$ and $n_o = 4$ in Steps 2 and 3 and then conduct Step 4 with the obtained estimates for $\sigma^2_p, \sigma^2_{\mathrm{Re}l}$, and $\sigma^2_{Abs}$.

**Note**. The above example deals with a two-facet *crossed* design (*p* x *r* x *o*) where each person is evaluated by each rater on each occasion. However, if different persons are evaluated on different occasions by each rater, then raters are still *crossed* with both persons and occasions, but this time occasions are *nested* within persons (denoted as *o*:*p*). This is referred to in GT as a *partially nested* (*o*:*p*) x *r* design. Theoretical and practical discussions of with various (crossed, nested, and partially nested) designs are provided in numerous sources on generalizability theory (e.g., Brennan, 2001, 2000; Shavelson & Webb, 1991). Practical applications of such designs can be tremendously facilitated by the use of the computer program for GT analysis GENOVA (Crick & Brennan, 1983).

## Dependability of Cutting-Score Classifications

In many measurement scenarios, criterion-referenced decisions (e.g., pass/fail or mastery/nonmastery) are based on a cutting score indicating the proportion of items answered correctly. In the context of GT, Brennan and Kane (1977) introduced the following *dependability index*, Φ(λ), for criterion-referenced decisions based on a cutting score, λ:

$$\Phi(\lambda) = \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \sigma_{Abs}^2}, \tag{25}$$

where μ is the theoretical mean of persons' scores (proportion of items correct),

λ is the cutting score (proportion of items correct),

$\sigma_p^2$ is the variance component for persons, and

$\sigma_{Abs}^2$ is the absolute error variance (see Equation 21).

Practical estimations of Φ(λ) are obtained by replacing the population parameter μ and the variance components by their sample-based estimates. Specifically, an unbiased estimate of $(\mu - \lambda)^2$ is provided by $(\overline{X} - \lambda)^2 - \hat{\sigma}^2(\overline{X})$, where $\hat{\sigma}^2(\overline{X})$ is the estimated variability produced by replacing the population mean μ for its sample estimate, $\overline{X}$. For example, the calculation of $\hat{\sigma}^2(\overline{X})$ with the "*person* x *item*" (*p* x *i*) design discussed in the previous section is

$$\hat{\sigma}^2(\overline{X}) = \frac{\hat{\sigma}_p^2}{n_p} + \frac{\hat{\sigma}_i^2}{n_i} + \frac{\hat{\sigma}_{pi,e}^2}{n_p n_i}. \tag{26}$$

**Example 9**: The purpose of this example is to illustrate the estimation of Φ(λ) for the data in Table 5 with the "*person* x *item*" (*p* x *i*) design when the cutting score for "pass/fail" decision is λ = .80 (i.e., 80 percent items correct are required for a person to pas*s* the test). It is necessary first to evaluate $\hat{\sigma}^2(\overline{X})$ with Equation 26 and then Φ(λ) with Equation 25. The estimation procedure is described in five steps.

**Table 5**

| | Item | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Person | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 3 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 |
| 4 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 5 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 8 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 10 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 11 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 12 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 13 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 |
| 14 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 19 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| 20 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*Step 1*: With the data in Table 5, $n_p = 20$, $n_i = 10$, and the sample mean (proportion correct) over all persons and items is $\overline{X} = 0.45$.

*Step 2*: By following the first three steps described in Example 7 for the "*person-by-item*" ($p$ x $i$) design, we obtain estimates of the components involved in calculations with Equations 26 and 25: $\hat{\sigma}^2_p = 0.0628$, $\hat{\sigma}^2_i = 0.0366$, $\hat{\sigma}^2_{pi,e} = 0.155$, and $\hat{\sigma}^2_{Abs} = 0.0192$.

*Step 3*: Using Equation 24 with the estimates reported in Step 2, we obtain

$$\hat{\sigma}^2(\overline{X}) = \frac{0.0628}{20} + \frac{0.0366}{10} + \frac{0.155}{200} = 0.0076.$$

*Step 4*: The term $(\mu - \lambda)^2$ in the right-hand side of Equation 25 is estimated as

$$(\mu - \lambda)^2 = (\overline{X} - \lambda)^2 - \hat{\sigma}^2(\overline{X}) = (.45 - .80)^2 - 0.0076 = 0.1149.$$

*Step 5*: Using Equation 25 with the values obtained in the previous steps, one can determine $\Phi(\lambda)$ for any (proportion correct) cutting score, λ, as

$$\Phi(\lambda) = \frac{0.0628 + (.45 - \lambda)^2}{0.0628 + (.45 - \lambda)^2 + 0.0192}. \tag{27}$$

For λ = .8 in Equation 27, we obtain $\Phi(\lambda) = .906$. Thus, the dependability for "pass/fail" (or mastery/nonmastery) classifications based on a cutting score λ = .8 (i.e., 80 percent items correct) is $\Phi(\lambda) = .906$.

Figure 3 displays values of $\Phi(\lambda)$ obtained with Equation 27 for different (proportion correct) cutting scores, λ, that vary from 0 to 1. The lowest dependability is reached when the cutting score equals the mean of proportion correct scores over all persons and items (in this case, λ = μ = 0.45). Specifically, replacing λ for .45 in Equation 27, we obtain the lowest dependability index with the data in this example: $\Phi(\lambda) = .766$ (see Figure 3). The implication is that the less reliable strategy in pass/fail (mastery/nonmastery) decisions is to use the mean domain score as a cutting score.

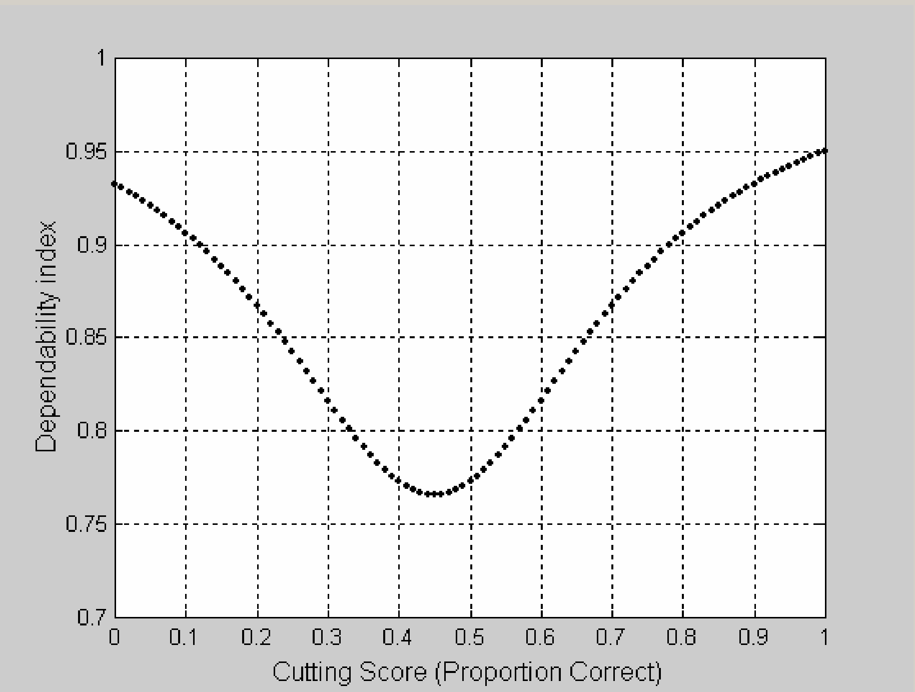


**Figure 3**. Values of the dependability index, $\Phi(\lambda)$, as a function of the cutting score, λ, for the data in Example 9.

## SUMMARY

This chapter introduces the concept of reliability, types of reliability, different methods of estimating reliability, and principles in interpreting and comparing reliability coefficients. Some major points are summarized here in the form of pith-laden responses to questions addressed in this chapter.

### What is reliability?

Generally, reliability of measurements (e.g., test scores and survey ratings) indicates their accuracy and consistency under random variations in measurement conditions. Such variations may be produced by person's conditions (e.g., fatigue or mood) and/or external sources (e.g., noise, temperature, different raters, and different test forms). It is important to emphasize that reliability relates to the measurement data obtained with an instrument, not to the instrument itself. Therefore, accidental reference to "reliability of a test" should be interpreted as "reliability of measurement data derived from a test".

In classical test theory, the *true score* of a person is defined as the theoretical mean of the observed scores that this person may have under numerous independent testings with the same test. A basic assumption is that the person's observed score is a sum of his/her true score and an error. Tests with equal true scores and equal error variances, for any population of examinees, are referred to as *parallel tests*. The *reliability* of test scores is defined as the correlation between observed scores on parallel tests or, equivalently, as the ratio of the true score variance to observed score variance for the same test.

*Standard error of measurement* (*SEM*) is the standard deviation of the (assumed normal) distribution of the difference between the persons' observed scores and their true scores. *Standard error of estimation* (*SEE*) is the standard deviation of the differences between the persons' actual true scores and estimated true scores when observed scores are used to predict true scores in a simple linear regression. The *SEE* is always smaller than the *SEM*. Therefore, when the estimation of true scores is of primary interest, the regression prediction of true scores using *SEE* should be preferred to confidence intervals for true scores using *SEM*.

## What are the classical types of reliability?

Five types of classical reliability were discussed in this chapter: internal consistency, test-retest reliability, alternate form reliability, classification consistency, and inter-rater reliability.

*Internal consistency* estimates of reliability are based on the average correlation among items within an instrument. If the instrument consists of different scales, internal consistency should be estimated for each scale. Widely used estimates of internal consistency are the *split-half* reliability coefficient and the Cronbach's coefficient *alpha* (or its equivalent version, *KR20*, for dichotomously scored items). The split-half method is appropriate under the strong assumption that the two halves of test being used represent parallel (sub)tests. A weaker assumption, *tau-equivalency* of test components, is required with Cronbach's alpha (or *KR20*). Internal consistency estimates are appropriate mostly with achievement tests. It is always useful to report the internal consistency of test scores even when other types of reliability are of primary interest. With *speed* tests, however, it would be misleading to report estimates of internal consistency.

*Test-retest reliability* indicates the extent to which persons consistently respond to the same test, inventory, or questionnaire administered on different occasions. It is estimated by the correlation between the observed scores of the same people taking the same test twice. The resulting correlation coefficient is referred to also as *coefficient of stability*. The major problem with test-retest reliability estimates is the potential for carry-over effects between the two test administrations (e.g., due to biological maturation, cognitive development, changes in information, experience, and/or moods). Thus, test-retest reliability estimates are most appropriate for measurements of traits that are stable across the time period between the two test administrations (e.g., personality and work values). Basically, test-retest reliability and internal consistency are affected by different sources of error and, therefore, it may happen that measures with low internal consistency have high temporal stability and vice versa.

*Alternate form reliability* relates to the consistency of scores on alternate test forms administered to the same group of individuals. It is estimated by the correlation between observed scores on two alternate test forms, referred to also as *coefficient of equivalence*. Estimates of alternate form reliability are also subject to carry-over effects but not as much as test-retest reliability coefficients because the persons are not tested twice with the same items. A recommended rule-of-thumb is to have a 2-week time period between administrations of alternate test forms.

*Criterion-referenced reliability* shows the consistency with which decisions about mastery-nonmastery of a specific objective (criterion) are made, using either the same test administered on two occasions or alternate test forms. Widely used classical indices of classification consistency are the observed proportion of consistent classifications and the (more conservative) Cohen's *kappa* coefficient which takes into account consistent classifications that may occur by chance.

*Inter-rater reliability* refers to the consistency (agreement) in subjective judgments of raters (experts, judges) used for classification purposes (e.g., to determine a "minimum level of competency" in pass/fail decisions) or scoring rubrics in alternative assessments (e.g., portfolios, projects, and products). Depending on the measurement case, frequently used estimates of inter-rater reliability are correlation coefficients, observed proportion of consistent classifications, and Cohen's *kappa* coefficient (or *kappa*-like coefficients).

## What is reliability of composite scores?

The person's scores from two or more scales of some instruments are combined into *composite scores* to measure and interpret a more general dimension (trait or proficiency) related to these scales. Composite scores are often used with test battery for achievement, intelligence, aptitude, depression, or eating disorders. Although the composite score may be simply the sum of several scale scores, its reliability is usually not just the mean of the reliabilities for the scales being combined. In this chapter, the reliability estimation for composite scores is addressed for cases when the composite score is a sum (or difference) of two scale scores or a weighted sum of scores.

**How to improve classical reliability of measurements?**

Researchers and test users can reduce measurement error thus improving reliability by (1) writing items clearly, (2) providing complete and understandable test instructions, (3) administering the instrument under prescribed conditions, (4) reducing subjectivity in scoring, (5) training raters and providing them with clear scoring instructions, (6) using heterogeneous respondent samples to increase the variance of observed scores, and (7) increasing the length of the test by adding items which are (ideally) parallel to those that are already in the test. The general principle behind improving reliability is *to maximize the variance of relevant individual differences and minimize the error variance*.

**What is dependability in generalizability theory?**

In **c**lassical (true-score) test theory, reliability estimation is based on a single source of measurement error (*facet*) – most frequently, *items* or *raters*. Also, classical estimates of reliability (e.g., test-retest correlation, coefficient alpha, or *KR20*) provide consistency information about the *relative* (norm-referenced) standing of persons to each other on the measurement scale, but not about their *absolute* (criterion-referenced) performance. As a person may have different true scores for different sets of items, the classical true score theory does not provide information about how generalizable the person's score is over a "universe" of admissible test items.

Generalizability theory (GT) is an extension of classical measurement theory and takes into account all available error sources (*facets*), such as items, raters, test forms, and occasions, that influence the reliability for either relative or absolute interpretations. This is done by representing the total error variance as a sum of variance components related to different sources of measurement error using ANOVA-based statistical designs. In GT, the accuracy of generalizing from a person's observed score in a measurement to his/her *universe score* under all admissible testing conditions (e.g., items, raters, and occasions) is referred to as *dependability*.

The section on GT dependability in this chapter illustrates the estimation of (relative and absolute) error variance, G-coefficient for *relative* (norm-referenced) interpretations, and index Φ(λ) for dependability of *absolute* (criterion-referenced) interpretations based on a (proportion correct) cutting score, λ. The lowest value of Φ(λ) is reached when the cutting score equals the mean of proportion correct scores over all persons and items. In this case (lowest possible dependability for absolute decisions with the test data), Φ(λ) is referred to simply as *dependability index* Φ. The logic with the GT designs illustrated in this chapter, "person-by-item" (*p* x *i*) and "person-by-item-by-occasion" (*p* x *r* x *o*), is efficiently applied in GT with various (crossed and/or nested) designs for estimation of dependability of relative and absolute interpretations of measurements (e.g., Brennan, 2001; Shavelson & Webb, 1991).

**Why is reliability important?**

The most important characteristic of (objective or subjective) of any measurement is its validity, that it, the degree to which measurement data lead to correct, meaningful, and appropriate interpretations. To allow for such interpretations, however, the scores should be accurate and consistent (i.e., reliable). Criterion-related validity of an entrance exam test, for example, is assessed by the correlation between the persons' scores on this test and their scores on a criterion (e.g., GPA at the end of the first academic year). However, the observed test scores on a test cannot correlate higher with any other (criterion) scores than they correlate with the true scores on the test. On the other side, the squared value of the correlation between the observed and true scores on a test represents the reliability of the test scores. Thus, *a criterion-related validity coefficient of test scores cannot exceed the square root of their reliability*. In other words, the reliability of scores predetermines a "ceiling" for their criterion-related validity. However, how closely this ceiling will be approached depends on other factors as well. Therefore, the reliability is a necessary (albeit not sufficient) condition for validity.

Reliability of measurement data is also an important assumption in hypothesis testing with statistical methods. For example, many research cases involve comparisons of two or more groups on a posttest while controlling for pre-test differences among the groups. The extent to which the pretest scores (i.e., covariate) are unreliable, the groups being compared will not be truly equated on the pretest and the results will be misleading. Using analysis of covariance assumes highly reliable scores on the control variables. Cliff (1987) compared the effects of lacking reliability and validity in measures to "effects that resemble tuberculosis as it occurred a generation or two ago: hey are widespread, the consequences are serious, the symptoms are easily overlooked, and most people are unaware of their etiology or treatment." (p. 129).

Yet, reliability estimates and/or measurement errors for the data at hand are still seldom reported in behavioral research. As Thompson (1992) pointed out, "one reason why researchers give too little attention to measurement considerations is that researchers often incorrectly presume that the characteristics of reliability inures to tests, when in fact reliability is a characteristic of a given set of data collected at a given time form a given set of subjects using a given protocol." (p. xii). The quality, accuracy, consistency, and meaningfulness of measurements should be in the focus of researchers and practitioners in counseling, education, and related fields. An important factor in this process is gaining a thorough understanding of the concept of reliability and skills in its estimation and proper interpretations.

## . STUDY QUESTIONS AND PROBLEMS

1. Provide examples of incidental person's conditions that may affect the observed score of the person on a test.
2. Provide examples of incidental external conditions that may affect the observed score of the person on a test.
3. What do *X*, *T*, and *E* stand for in classical test theory? How do they relate?
4. Would the person's true score be the same on any two tests that measure the same trait (e.g., verbal proficiency or anxiety)? Explain.
5. How are reliability and error of measurement related?

6. Provide (up to three) equivalent definitions of reliability in classical test theory.
7. Which method of estimating internal consistency reliability, *split-half* method or Cronbach's *alpha*, works under stronger (more difficult to satisfy) assumptions?
8. If a person has 53 points on a test and the standard error of measurement is two points, what is (approximately) the 95% confidence interval for the person’s true score on this test?
9. If the variance of the observed scores on a test is 36 and the Cronbach’s *alpha* coefficient is .85, what is (a) the standard error of measurement, (b) the standard error of estimation, and (c) the predicted true score for a person with an observed score of 24 if the population mean is reported to be 30.
10. Which classical assumption requires the test components to measure the same trait and have equal true score variances? What is the effect of violating this assumption on estimates of internal consistency (Cronbach’s *alpha* or *KR20*)?
11. Which coefficient, *alpha* or *KR20*, is appropriate for estimating the internal consistency of scores obtained on a 5-point survey scale?
12. Are the *alpha* coefficient of reliability and the correlation coefficient for test-retest reliability interchangeable?
13. What may cause the correlation between the scores on alternate test forms to be much lower (say, 0.20 or more) than the internal consistency of these scores?
14. What is important to take into account when adding new items to a test to increase its reliability as predicted by the Spearman-Brown formula?
15. Is a large variability of observed scores more important for internal consistency or classification consistency (e.g., in “mastery/nonmastery” decisions)?
16. How would you check for subjectivity in scoring when two alternate forms of a measurement instrument are used by (two or more) raters?
17. Is the reliability of a composite score obtained by summing the person’s scores on two scales equal to the average reliability for the two scales?

18. Is the reliability of the pretest to posttest difference (gain score) usually smaller or larger than each of the reliabilities of the pretest and posttest scores?
19. What is the conceptual difference between the classical *true score* of a person and the person's *universe score* in generalizability theory (GT)?
20. In which measurement scenario (design) the classical coefficient *alpha* and the generalizability coefficient (G-coefficient) are comparable?
21. What type of error variance (relative, absolute, or both) is affected by the variance of the interaction between students and raters in a measurement where each student is evaluated by each rater ( "student -by-rater" design)?
22. Why is the relative error variance always smaller than the absolute error variance in GT? (*Hint*: see Figure 2.)
23. What are, in general, the advantages of dependability coefficients in GT over reliability coefficients in classical test theory?
24. Determine the degree of agreement between two teachers (inter-rater reliability) given the frequencies of their *mastery-nonmastery* classifications of 100 essays. (*Hint*: see Example 3.)

| | | Teacher B | |
|---|---|---|---|
| | | Mastery | Nonmastery |
| Teacher A | Mastery | 40 | 10 |
| | Nonmastery | 15 | 35 |

25. A study investigates the effect of a counseling treatment for reducing road rage as measured on a driving anger scale. The *alpha* reliability coefficient for the participants' scores on this scale was found to be .88 before the treatment and .79 after the treatment. The treatment effect was measured by a *gain* score obtained by subtracting the standard (z-) scores on the pre-treatment from the standard (z-) scores on the post-treatment measurements. Estimate the reliability of the gain score given that its variance is 1.42. (*Hint*: see Example 5.)

26. What is the reliability for the sum of the person's scores on two different scales of cognitive processing ability, given that (a) the score variances for the two scales are 8.5 and 10.2, respectively, (b) the reliability estimates for the two scales are .77 and .82, respectively, and (c) the variance of the sum (composite score) is 14.5. (Hint: use Formula 14).

27. What is the reliability of a composite score, *Y*, obtained as a sum of the person's scores on three subscales of mental health disorders (e.g., anxiety, depression, and sleep disturbances), denoted $X_1$, $X_2$, and $X_3$, respectively. Given below are the variances and the *alpha* coefficients of reliability for the three scales and the variance of the composite score. (*Hint*: Extend Formula 14 for $X_1$, $X_2$, and $X_3$.)

| | Subscale | | | Composite |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | score, *Y* |
| Variance | 35 | 28 | 42 | 90 |
| Reliability | .80 | .77 | .85 | |

28. In a study on aggressive behavior among middle school students, 40 students were evaluated by four raters (teachers and school counselors) on a 4-point scale. As each rater evaluated each student, the data were analyzed with the "*student-by-rater*" (*s* x *r*) design in GT. The analysis of variance (ANOVA) results for the sample variances ("mean square") for students (s), raters (*r*), and their interaction confounded with other possible sources of error, *e*) were: $MS_s = 2.76$, $MS_r = 5.71$, and $MS_{sr,e} = 1.85$. Using these results for the sample of 40 students ($n_s = 40$) and four raters ($n_r = 4$), calculate the estimates for (a) relative error variance, $\sigma^2_{\mathrm{Re}l}$, (b) absolute error variance, $\sigma^2_{Abs}$, (c) G-coefficient for relative (norm-referenced) interpretations, and (d) dependability index, Φ, for absolute (criterion-referenced) interpretations. (*Hint*: see Example 7.)

## SUGGESTED REFERENCES FOR ADDITIONAL READING

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (2000). Performance assessment from the perspective of generalizability theory. *Applied Psychological Measurement, 24*, 339-353.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14(3)*, 277-289.

Chase, C. (196). Estimating the reliability of criterion-referenced tests before administration. *Mid-Western Educational Researcher, 9*, 2-4.

Cliff, N. (1987). *Analyzing multivariate data*. San Diego, CA: Harcourt Brace Jovanovich.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system*. Iowa City, IO: ACT.

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

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 27(6)*, 440-458.

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62(5)*, 783-801.

Feldt, L. S., & Brennan, R. L. (1989). *Reliability*. In R.L. Linn (Ed.) Educational Measurement (3rd ed., pp. 105-146). New York: Macmillan.

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement, 17*, 510-521.

Mellenbergh, G. J. (1999). A note on simple gain score precision. *Applied Psychological Measurement, 23*, 87-89.

Nunnally, J. C., & Bernstein, I.. H. (1994). *Psychometric theory* (3rd ed.) New York: McGraw-Hill.

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82*, 85-86.

Raykov, T. (1998). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement, 22*, 369-374.

Reckase, M. D. (1995). Portfolio assessment: A theoretical estimate of score reliability. *Educational Measurement: Issues and Practice, 14(1)*, 12-14, 31

Sax, G. (1980). *Principles of educational and psychological measurement and evaluation* (2nd ed.). Belmont, CA: Wadsworth.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Subkoviak, M. J.(1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25(1)*, 47-55.

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 161-185). Hillsdale, NJ: Lawrence Erlbaum.

Thompson, B. (1992). Editorial comment: Misuse of ANCOVA and related "statistical control" procedures. *Reading Psychology: An International Quarterly, 13*, iii-xviii.

Thompson, B. &Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19*, 149-154.