The two most important properties of an assessment are its validity and reliability. Validity refers to the meaningfulness of the interpretations and uses of a test score and is the most important property of an assessment. Reliability refers to the extent to which test scores are free from errors of measurement. Thus, validity examines the interpretations and uses that can reasonably be made from the consistent part of the test scores, whereas reliability is concerned with inconsistent or random errors of measurement. As a result, reliability is a necessary but not sufficient condition for validity.

That is, there needs to be some level of consistency to understand the meaningfulness of particular uses and interpretations of test scores, but measuring consistently does not guarantee the meaningfulness of the interpretations or uses.

Reliability and validity are not global properties of an assessment. Instead, they are properties of specific uses and interpretations that are made from a set of test scores. A test could be valid for a particular use or interpretation and not for another. For example, a test might measure the curriculum covered in a school without providing valid estimates of student performance because of the length of the tests or the nonequivalence of forms. The same is true for reliability. For example, a test might provide reliable scoring without being stable over time. In addition, reliability and validity are a matter of degree. Tests are not considered valid or invalid. Instead, they are valid to some degree. Similarly, a test is not considered reliable or unreliable, but is reliable to some degree.

Estimates of reliability are indices that quantify the amount of measurement error for a particular test use or interpretation for a specified population. Although reliability can be defined broadly in terms of consistency or generalizability, specific statistical indices of reliability will vary depending on the statistical model and the sources of error. The statistical model may be based on classical test theory, generalizability theory, or item response theory. Classical test theory and generalizability theory are based on total scores, whereas item response theory is based on an estimate of a latent trait. In this entry, only classical test theory and generalizability theory are considered. Within each theory, there are multiple indices of reliability based on multiple sources of measurement error, including item heterogeneity, equivalence of test forms, stability over time, and consistency of subjective ratings. Different sources of error would be of concern in different contexts. For example, the test score of a student writing an essay is affected by errors in scoring, whereas the test score from a student taking a multiple-choice test is affected by the heterogeneity of the items selected to measure the construct. In addition, a test score can be affected by multiple sources of error simultaneously. A student taking the GRE might be affected by the heterogeneity of the items, the form of the test, and the subjectivity of the scoring for the written portion of the test. Thus, there are many types of reliability that vary depending on the sources of error being considered as well as the statistical model or test theory being used. These varying definitions will be selected based on the particular test use or score interpretation being made, and one type of reliability should not be considered interchangeable with another.

Classical Test Theory and Estimates of Reliability

Classical test theory assumes that any observed test score, X, is the sum of a true score, T, and a random error, E. That is, $X = T + E$. The issue of defining the true score is a matter of validity, whereas the issue of defining the random error is a matter of reliability. According to classical test theory, the error is the sum of all random components, whereas the true score is the sum of all consistent effects. Thus, the error is undifferentiated with respect to different sources of randomness, unlike in generalizability theory. Broadly, two indices of reliability are commonly reported: the reliability coefficient and the standard error of measurement.

The reliability coefficient (ρ) is defined as the ratio of the true score variance to the observed score variance, or the ratio of the true score variance to the sum of the true score variance and the error variance. Hence, the value of the reliability coefficient is the proportion of variation in test scores that can be attributed to consistent measurement (i.e., the true score). The reliability coefficient ranges from 0.0 to 1.0, with higher values being preferred. At $\rho = 0.0$, there is no consistency in the measurement procedure and the observed score is equal to random error ($X = E$). At $\rho = 1.0$, the observed score has no error and is equal to the true score ($X = T$). In practice, the reliability coefficient will be somewhere between the two extreme values.

The standard error of measurement (SEM) is the standard deviation of the errors of measurement. The SEM ranges from 0.0 to the standard deviation of the observed scores, σ_x. When the SEM $= \sigma_x$, there is no consistency in the measurement procedures, the reliability coefficient is equal to 0.0, and the observed score is equal to the random error. When the SEM $= 0.0$, there is perfect consistency in the test scores, the observed score is equal to the true score, and the reliability coefficient is equal to 1.0. In practice, the SEM will fall somewhere between the two extreme values.

Miller, M. (2008). Reliability. In N. Salkind (Ed.), Encyclopedia of educational psychology. (pp. 847-853). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412963848.n230

The reliability coefficient is an easily interpreted index of the consistency of the test scores because it is in a standard range for all tests. Although the SEM is more difficult to interpret initially, it is in the metric of the test scores that allows for the interpretation of the individual test scores via confidence intervals. Another advantage of the SEM is that it is not based on the true scores, and consequently, it is not influenced by sampling errors. The reliability coefficient will be underestimated when the sample range of scores is restricted, whereas the SEM will be largely uninfluenced by sampling fluctuations.

Types of Reliability

Within the framework of classical test theory, there are several types of reliability coefficients based on the source of the random errors. The types of reliability discussed below are test-retest, alternate form, alternate form test-retest, interrater, split half, and internal consistency.

Test-retest reliability is used to examine the stability of the trait being measured over time. The reliability coefficient is the correlation between test scores for a sample taking the same test on two occasions. Generally, test-retest reliability is higher when the time span between test administrations is shorter. However, the test-retest reliability should be estimated with a time interval that mirrors the actual use of the test rather than trying to maximize the value of the coefficient.

Alternate form reliability is used to measure the equivalence of test scores across two (parallel) forms of a test. The reliability coefficient is the correlation between test scores on the two forms of the test taken by the same sample without a substantial time lag. Usually, half of the sample receives one form first (e.g., Form A), and the other half of the sample receives the other form first (e.g., Form B) so that there is no order effect. Then, examinees take the form they have not taken yet. Alternate form reliability is higher when care is taken to make sure that the two forms are equivalent in content and statistical properties (i.e., mean, standard deviation, and distribution shape).

Alternate form test-retest reliability follows the same procedure as with the alternate form reliability except that there is a time lag between test administrations. In this case, the errors of measurement include stability over time and equivalence of the forms. In general, this type of reliability will be lower than alternate form or test-retest reliability, which target only one type of random error.

Interrater reliability is used to measure the consistency of ratings from subjective scoring. The reliability coefficient is the correlation between the ratings from two raters on the same sample of writings/essays. Interrater reliability is higher when standardized procedures are used by the raters to score the writings. At a minimum, the standardized procedures should include training of the raters and clearly defined rubrics. Large-scale assessments further standardize the procedures to include benchmark writings, monitoring the process, intervening when ratings disagree, and other procedures to check the rating process.

Split half reliability is used to measure the consistency within a single administration of a test by examining the relationship between two halves of the same test. The procedure for split half reliability is to administer a single form of the test to a sample. The reliability estimate is then based on the correlation between two halves of the test adjusted for test length. That is, the test is divided into two equivalent halves based on test content and item statistics (often, this can be accomplished by using odd- and even-numbered items to form the halves), and the halves are correlated. However, the reliability will be less for half of a test than it is for the full-length test. Consequently, the correlation between the halves is adjusted upward using the Spearman-Brown prophecy formula. Split half reliability will be higher when the equivalence of the two forms is higher in terms of content and item statistics. However, the matching of the two halves should not be completed on the basis of the sample statistics because random sampling fluctuations could inflate the value of the reliability. Instead, careful matching should be completed based on content and item statistics from a prior data collection.

Internal consistency is used to measure the consistency of items within a single test form. The procedure for internal consistency is to administer a single form of the test to a sample and estimate the internal consistency using item and test statistics with an internal consistency formula. The formula for internal consistency has many equivalent forms in the literature, including the Kuder-Richardson 20 (KR20) formula for dichotomously scored items and Cronbach's alpha. Internal consistency is also easy to compute with most standard statistical software (e.g., SPSS or SAS).

Internal consistency is higher when the items are more homogeneous.

Below is a summary of the reliability coefficients and their major sources of error that are reported in classical test theory:

Reliability Type	*Source of Error*
Test-retest	Stability over time
Alternate form	Equivalence across forms
Alternate form test-retest	Stability over time and equivalence across forms
Interrater	Consistency of ratings
Split half	Equivalence across halves
Internal consistency	Equivalence and item homogeneity

Each of the reliability coefficients above differs in its data collection procedure, computation, and major source of error. The "appropriate" reliability coefficient should match the intended use or interpretation of the test. For example, when subjective measurements are part of the assessment procedure, interrater reliability is needed. When multiple items are being used (which should be the case), internal consistency or split half reliability should be used. In short, the reliability estimate(s) should include all sources of error that will be part of the test use or interpretation. One type of reliability should not substitute for another.

Standard Error of Measurement

The reliability coefficient is used to quantify the precision of an assessment for a particular use or interpretation. The index is simple to interpret because it is always based on the same scale (0.0–1.0). However, the reliability coefficient fails to show the amount of error that might be expected in an individual's test score. The SEM is the standard deviation of the errors of measurement and can be used to create confidence intervals for examinee scores. Assuming a normal distribution, 68% of the observed scores will be within one SEM of their true score, and 95% of the observed scores will be within 1.96 SEMs of their true score. For example, if SEM = 2.00 and an examinee's true score was 25, upon repeated measurements, 68% of the scores for that examinee would be between 23 and 27. Note that the confidence interval is around the true score and not the observed score, which leads to the interpretation that 68% of the time that a confidence interval based on one SEM is constructed, it will contain the true score.

As with the reliability coefficient, a SEM can be created for different types of measurement error. In fact, the SEM is calculated using the appropriate reliability coefficient so that the appropriate source of error is being used. Thus, the table mentioned earlier can be used for the SEM or the reliability coefficient so that the SEM can be created with each of the different sources of error.

Magnitude of Reliability

The literature does not provide definitive guidance on acceptable levels of reliability. However, it is clear that what constitutes an acceptable level of reliability is determined by the use of the test. Uses of the test with higher stakes require higher levels of reliability. For example, reliability for a test being used in theoretical research may not require the same level of consistency as would be required for high-stakes uses of tests such as high school graduation, certification, or licensure.

How to Increase Reliability

It is important to be able to increase reliability when developing instruments. In general, there are two ways that should always be considered when increasing reliability: greater standardization and increasing the number of items. Test administration and test development procedures should be standardized so that no random errors are introduced. The effect of the standardization will not only globally affect each type of reliability, but it will also have specific effects on certain types of reliability. Standardization includes methods to create equivalent forms of a test (alternate form), methods to create homogeneous pools of items (internal consistency), or equivalent halves of tests (split half). Standardization also includes methods to create consistency in scoring through the development of rubrics and standardized scoring procedures (interrater).

Another key element to increasing reliability is increasing the length of the test. The Spearman-Brown formula is based on the principle that longer tests are

more reliable. Assuming that the conditions of testing do not change with increased length (i.e., fatigue, boredom, or item quality), increasing the number of items always leads to more reliable tests. Thus, short forms of a test or subscores are generally less reliable than the full form. As a consequence, subscores and short forms are also more difficult to interpret. Thus, a test that provides a reliable total score for accountability may not be useful when examining the subscores that might be needed for diagnostic interpretations.

Reliability and Aggregation

Increasing the number of items will increase the reliability of test scores because the scores are averaged over more data. Similarly, increasing the number of examinees and averaging across their scores will reduce the SEM and increase reliability. That is, the reliability of the mean will be higher than the reliability of an examinee. This applies to estimates for the full sample as well as aggregates such as classrooms, schools, or states. Whether averaging across items or examinees, the estimate becomes more stable. Indeed, the SEM will almost always be lower for group means than it is for individual means. (Note: under some conditions, the reliability coefficient could be lower for the group means when the true score variance in the groups is restricted in range. However, even under these conditions, the SEM will typically be lower and the group means will be more stable.)

Reliability and Growth Scores

The reliability of growth or difference scores, defined as posttest minus pretest, has received considerable attention in the literature. Some have argued that the growth score is unreliable and that growth is negatively correlated with the pretest. That is, growth will be higher for examinees with low pretests. However, other researchers have pointed out that low reliability for the difference scores does not necessarily result in less power for comparisons among groups, and the difference score may be the construct of interest. At any rate, caution should be used in interpreting growth scores at the examinee level.

Relationship of Reliability and Validity

Classical test theory assumes that an examinee's test score is the composite of a true score and random error. Validity addresses the true score by examining its uses and interpretations. Thus, any systematic error or bias is part of the true score, whereas only random errors are addressed in the reliability analysis. As discussed above, reliability is a necessary but not sufficient condition for validity. This means that there needs to be a true score (with some level of reliability) to examine validity, but that the existence of a true score does not guarantee that it is not a biased estimate of the construct of interest as a result of some systematic error.

In addition to this relationship of validity and reliability, there may be a tension between the two psychometric properties of the test. Higher reliability can be attained by standardizing the testing procedure, which has the potential to reduce the breadth of the construct being measured and, thus, to decrease the validity. For example, higher internal consistency is attained by increasing item homogeneity. To the extent that the construct requires heterogeneity of the items, this will create a tension between reliability and validity. As a second example, interrater reliability is increased by standardizing the scoring procedure (e.g., clearly defined rubrics and training). This standardization can limit the definition of good writing by not including some types of writing in the rubric and thus, as a consequence, limit the breadth of the construct. As a result, it is important to consider the impact of any standardization on the validity of the test as well as the reliability so that the construct is still clearly being measured.

Generalizability Theory and Estimates of Reliability

Classical test theory examines errors of measurement with a single undifferentiated error that may represent multiple sources of error (e.g., alternate form test-retest reliability). In addition, classical test theory examines reliability only from a norm-referenced perspective. That is, the methods rely solely on correlations that focus on rank ordering of scores. The correlations are sensitive to differences in rank ordering but not to shifts in scale. Thus, the reliability can be high even when the scales for the two forms, raters, and so on, are substantially different. For example, Rater A could rate 10 points higher than Rater B, and the reliability would be equal to 1.0 as long as every essay was scored exactly 10 points higher by Rater B than Rater A.

Generalizability theory solves each of these issues by (a) modeling multiple sources of error and (b) differentiating between errors based on rank ordering (i.e., relative error) and errors based on point estimation (i.e., absolute error). Generalizability theory assumes that each examinee has a universe score that is his or her average score across all conditions in the universe of admissible observations. That universe is composed of measurement facets (e.g., raters, items) with a particular level of a facet being a condition (e.g., selected raters or items). The potential measurement conditions selected from the study are then considered to be the universe of generalization. Generalizability theory is more complex statistically than classical test theory and is done in two stages: the G-study and the D-study. In the G-study (generalizability study), random effects analysis of variance (ANOVA) is used to estimate variance components for each of the effects in the model. The ANOVA, with the associated variance components, can be estimated with one or more facets (e.g., persons by items, or persons by items by raters). In the D-study (decision study), alternative measurement models can be examined to optimize the measurement procedures or to examine a reasonable set of measurement conditions. The results of the D-study identify the universe of generalization.

Similar to classical test theory, there are two types of indices computed in the D-study that show the consistency of the measurement procedure. The generalizability coefficient, or dependability index, is the ratio of the universe score variance to the universe score variance plus the error variance, and it is analogous to a reliability coefficient, whereas the second index is analogous to the SEM. The generalizability coefficient shows the ratio when using relative error variance and emphasizes the rank ordering of scores. Thus, it would be used for norm-referenced score reporting. The dependability index shows the ratio when using absolute error variance and emphasizes the absolute magnitude of the test scores. Thus, it would be used for criterion-referenced score reporting. The second index is the standard error. The standard error also can be computed for relative or absolute score reporting. The use and interpretations of the indices in generalizability theory are analogous to the indices in classical test theory.

Absolute error is always greater than or equal to relative error. Consequently, the generalizability coefficient is less than or equal to the dependability index, and the absolute standard error is greater than or equal to the relative standard error. As a result, more conditions (items, raters, etc.) may be needed when estimating examinee scores absolutely rather than relative standing.

Consistency of Classification

Each of the reliability coefficients, whether from classical test theory or generalizability theory, is based on continuous variables. Often, the measurement procedure is based on the classification of examinees. For example, the National Assessment of Educational Progress (NAEP) classifies students as Advanced, Proficient, Basic, and Below Basic. Clearly, when the data are nominal or categorical, a different statistical procedure must be used to examine consistency. Decision consistency is a method of examining reliability and the exact agreement across measurement conditions when the data are categorical. Decision consistency can be calculated for each of the sources of error described above in the section on classical test theory.

Two indices commonly reported for decision consistency are the proportion agreement and Cohen's Kappa. Proportion agreement shows the proportion of the examinees that are classified the same across forms, time, and so on. For example, if two forms were given to a sample for the NAEP reading assessment in Grade 8, the proportion agreement would be equal to the sum of the proportions that were Advanced on both forms, Proficient on both forms, Basic on both forms, and Below Basic on both forms. The same calculations could be done with other sources of error (e.g., ratings are the same across two raters). Cohen's Kappa is the proportion agreement after statistically adjusting for the expected agreement. Thus, Cohen's Kappa shows the agreement above and beyond chance and generally has lower values than the proportion agreement.

M. David Miller

See also Assessment; Descriptive Statistics; Evaluation; Testing

Further Readings

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

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.
Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.