Glossary for Reliability

Term	Definition
Accuracy	Appropriate educational decisions depend on accuracy of educational assessments. Accurate assessments will improve the quality of decisions whereas inaccurate assessments will do the opposite.
Alternate-form reliability	Consistency of results among two or more different forms of a test.
Classical test theory	Is a body of related psychometric theory that predicts outcomes of psychological testing such as the difficulty of items or the ability of test-takers. Generally speaking, the aim of classical test theory is to understand and improve the reliability of psychological tests.
Correlation coefficient	Reflects the degree of similarity between students' scores on the two tests in a test-retest stability reliability.
Equivalence reliability	The extent to which measurement on two or more forms of a test is consistent.
Equivalent (parallel) forms	Two or more forms of a test covering the same content whose item difficulty levels are similar.
Generalizability theory	G Theory provides a framework for conceptualizing, investigating, and designing reliable observations. The most important factor you can consider when judging potential tasks for performance assessments. With performance assessment it is often more difficult to generalize accurately about what skills and knowledge are possessed by the student.
Internal consistency reliability	Is the consistency in the way an assessment instrument's items function.
Inter-rater agreement and inter-rater reliability	Inter-rater agreement and inter-rater reliability are two indices that are used to ensure scoring consistency. The most popular method used for testing inter-rater reliability is correlation. Correlation tests the relationship between the scores of two raters.
Kuder Richardson Coefficient (K-R 20)	The Kuder Richardson Coefficient of reliability (K-R 20) is used to test the reliability of binary measurements such as exam questions, to see if the items within the instruments obtained the same binary (no/yes, right/wrong) results over a population of testing subjects.
Reliability procedures	The consistency with which an assessment procedure measures what it is measuring.
Reliability types	1. Test-retest reliability is a measure of reliability obtained by administering the same test twice over a period of time to a group of individuals. The scores from Time 1 and Time 2 can then be correlated in order to evaluate the test for stability over time. *Example:* A test designed to assess student learning in psychology could be given to a group of students twice, with the second administration perhaps coming a week after the first. The

	obtained correlation coefficient would indicate the stability of the scores. 2. Parallel forms reliability is a measure of reliability obtained by administering different versions of an assessment tool (both versions must contain items that probe the same construct, skill, knowledge base, etc.) to the same group of individuals. The scores from the two versions can then be correlated in order to evaluate the consistency of results across alternate versions. *Example:* If you wanted to evaluate the reliability of a critical thinking assessment, you might create a large set of items that all pertain to critical thinking and then randomly split the questions up into two sets, which would represent the parallel forms. 3. Inter-rater reliability is a measure of reliability used to assess the degree to which different judges or raters agree in their assessment decisions. Inter-rater reliability is useful because human observers will not necessarily interpret answers the same way; raters may disagree as to how well certain responses or material demonstrate knowledge of the construct or skill being assessed. *Example:* Inter-rater reliability might be employed when different judges are evaluating the degree to which art portfolios meet certain standards. Inter-rater reliability is especially useful when judgments can be considered relatively subjective. Thus, the use of this type of reliability would probably be more likely when evaluating artwork as opposed to math problems. 4. Internal consistency reliability is a measure of reliability used to evaluate the degree to which different test items that probe the same construct produce similar results. A. Average inter-item correlation is a subtype of internal consistency reliability. It is obtained by taking all of the items on a test that probe the same construct (e.g., reading comprehension), determining the correlation coefficient for each *pair* of items, and finally taking the average of all of these correlation coefficients. This final step yields the average inter-item correlation. B. Split-half reliability is another subtype of internal consistency reliability. The process of obtaining split-half reliability is begun by "splitting in half" all items of a test that are intended to probe the same area of

	knowledge (e.g., World War II) in order to form two "sets" of items. The *entire* test is administered to a group of individuals, the total score for each "set" is computed, and finally the split-half reliability is obtained by determining the correlation between the two total "set" scores.
Reliability and validity	**Reliability** is the degree to which an assessment tool produces stable and consistent results. **Validity** refers to how well a test measures what it is purported to measure.
Score reliability	The consistency with which two or more individuals would score the same response to a test item.
Spearman-Brown Formula	A formula for estimating reliability if test length is changed.
Split-half reliability	A procedure for estimating test reliability by which a test is divided into two comparable halves and the scores on the halves are then correlated.
Stability	The extent to which measurement on the same test is consistent over time.
True score variance	Are the scores that the students would obtain if the test had perfect reliability.
Test reliability	There are three types of evidence of test reliability, stability, alternate form, and consistency. Reliability is the consistency with which a test measures with whatever a test is supposed to measure (Popham).
Text-based inference	A judgment that is made related to a test result. The appropriateness of an inference is the basis for validity.