



Glossary of Assessment Terms

Unless otherwise noted, these terms and topics were taken and combined from three sources: the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) website at <http://cresst96.cse.ucla.edu/CRESST/pages/glossary.htm>, the Measurement Excellence and Training Resource Information Center (METRIC) website at <http://www.measurementexperts.org/learn/terms.asp>, and the National Academy Press publication entitled, "Uncommon Measures: Equivalence Among Educational Tests."

Accommodations and Adaptations - To enable **students with disabilities (SWD)** and limited English proficient students to take the assessment, changes in the way assessments are designed or administered are often made. Examples of the types of modifications that might be made include Braille forms for blind students or tests in languages other than English for students whose primary language is not English.

Alignment - The process of linking content and performance standards to assessment, instruction, and learning in classrooms. One typical alignment strategy is the step-by-step development of (a) content standards, (b) performance standards, (c) assessments, and (d) instruction for classroom learning. Ideally, each step is informed by the previous step or steps, and the sequential process is represented as follows:

Content Standards - Performance Standards - Assessments - Instruction for Learning

In practice, the steps of the alignment process will overlap. The crucial question is whether classroom teaching and learning activities support the standards and assessments. System alignment also includes the link between other school, district, and state resources. Alignment supports the goals of the standards, i.e., whether professional development priorities and instructional materials are linked to what is necessary to achieve the standards.

Alternative Assessment - (also authentic or performance assessment). An assessment that requires students to create a response to a question rather than choose from a set of responses provided to them. Exhibitions, investigations, demonstrations, written or oral responses, journals, and portfolios are examples of the assessment alternatives we think of when we use the term "alternative assessment." Ideally, alternative assessment requires students to actively accomplish complex and significant tasks, while bringing to bear prior knowledge, recent learning, and relevant skills to solve realistic or authentic problems. Alternative assessments are usually one key element of an **assessment system**.

Alternate forms reliability - A measure of reliability, in which alternate forms of the same measurement instruments are administered to the same subjects on separate occasions.

Analytic Scoring - Evaluating student work across multiple **dimensions** of performance rather than from an overall impression (**holistic scoring**). In analytic scoring, individual scores for each dimension are scored and reported. For example, analytic scoring of a history essay might include scores of the following dimensions: use of prior knowledge, application of principles, use of original source material to support a point of view, and composition. An overall impression of quality may be included in analytic scoring.



Assessment - The process of measuring one or several variables of interest in order to make decisions about individuals or inferences about a population. The process of gathering, describing, or quantifying information about performance.

Assessment System - The combination of multiple assessments into a comprehensive reporting format that produces comprehensive, credible, dependable information upon which important decisions can be made about students, schools, districts, or states.

Benchmark - A detailed description of a specific level of student performance expected of students at particular ages, grades, or development levels. Benchmarks are often represented by samples of student work. A set of benchmarks can be used as "checkpoints" to monitor progress toward meeting performance goals within and across grade levels, i.e., benchmarks for expected mathematics capabilities at Grades 4, 8, and graduation.

Bias - A source of systematic error in measurement or research. Unlike random error, bias may be controlled. Bias usually favors one group (e.g. race, age, gender) over another.

Calibration - Calibration is the process of putting scores from two different tests on the same scale. This process is used when tests that are created for different purposes, ones with different frameworks and test specifications, are to be compared.

Classical Test Theory - Classical test theory is the earliest theory of measurement that attempts to estimate the strength of the relationship between the observed score and the true score. The mathematic expression is: $X = T + E$, where X is the observed score, T is the true score and E is the error; also known as classical reliability theory. (From Suen HK. Principles of Test Theories. Lawrence Earlbaum Associates, Publishers. 1990)

Classroom Assessment - An assessment developed, administered, and scored by a teacher or set of teachers with the purpose of evaluating individual or classroom student performance on a topic. Classroom assessments may be aligned into an **assessment system** that includes **alternative assessments** and either a **norm-referenced** or **criterion-referenced assessment**. Ideally, the results of a classroom assessment are used to inform and influence instruction that helps students reach high standards.

Computer Adaptive Testing (CAT) - A computer-administered questionnaire based on the item response theory (IRT). The examinee responds to a sequence of questions, selected from an item pool. A program calculates statistics after each response and determines the subsequent order of items based on the examinee's responses to the previous items. It attempts to determine the examinee's true ability.

Confidence Interval - The likely range of values with a known probability of including the true value.

Confirmatory Factor Analysis - A form of theory-testing that confirms the existence of factors and offers a viable method of determining construct validity. The process tests a pre-determined



hypothesis and determines the specific variables that correlate to specific factors. It also pinpoints correlated factors.

Construct Validity - The degree to which an instrument accurately measures the construct of interest. Examples of construct validity include, but are not limited to, convergent and discriminant validity.

Content Standards - Broadly stated expectations of what students should *know and be able to do* in particular subjects and grade levels. Content standards define for teachers, schools, students, and the community not only the expected student skills and knowledge, but what schools should teach. An example of a Regents' standard is: "A **passing** student is able to demonstrate, on demand, the skills required for commencement-level achievement in each of the learning standards and key ideas assessed."

Content Validity - A method of establishing validity based on expert judgment that the content of the measure is consistent with what is to be measured.

Convergent Validity - A type of validity that is determined by hypothesizing and examining the overlap between two or more tests that presumably measure the same construct.

Correlation Coefficient - A value indicating the strength of a linear relationship between two variables in a sample.

Criteria - Guidelines, rules, characteristics, or **dimensions** that are used to judge the quality of student performance. Criteria indicate what we value in student responses, products or performances. They may be holistic, analytic, general, or specific. **Scoring rubrics** are based on criteria and define what the criteria mean and how they are used.

Criterion-Referenced Assessment - An assessment where an individual's performance is compared to a specific learning objective or **performance standard** and not to the performance of other students. Criterion-referenced assessment tells us how well students are performing on specific goals or standards rather than just telling how their performance compares to a norm group of students nationally or locally. In criterion-referenced assessments, it is possible that none, or all, of the examinees will reach a particular goal or performance standard. For example: "all of the students demonstrated proficiency in applying concepts from astronomy, meteorology, geology, oceanography, and physics to describe the forces that shape the earth."

Criterion-Related Validity - The degree to which a measure or test correlates with other measures or tests of the same construct assessed either concurrently or in the future; the ability of a test to predict a criterion. See also predictive validity.

Cronbachs Alpha - An index of reliability for a set of items that indicates the internal consistency of the items in a measure.



Cut score - A point on a score scale in which scores at or above the point are in a different category or classification than scores below the point (e.g. pass versus fail).

Dichotomous Response - Selections to items containing two main category types (e.g. male and female, pass or fail, etc.) Also known as a binary choice.

Dimensions - Desired knowledge or skills measured in an assessment and usually represented in a **scoring rubric**. For example, a measurement of student teamwork skills on a performance assessment might include 6 dimensions: adaptability (recognizing problems and responding appropriately), coordination (organizing team activities to complete a task on time), decision making (using available information to make decisions), interpersonal (interacting cooperatively with other team members), leadership (providing direction for the team), and communication (clearly and accurately exchanging information between team members).

Discriminant Validity - A type of validity that is determined by hypothesizing and examining differential relations between a test and measures of similar or different constructs. It is the opposite of convergent validity and is also known as divergent validity.

Ecological Validity - A type of validity that indicates the extent to which results will generalize to other settings.

Equating - Equating is the most technically supported method for linking two or more assessments. For equating to be possible, test content, format, purpose, administration, item difficulty, and populations must be equivalent. Some equating methods include linear equation and equipercentile equating.

Equipercentile Equating - Equipercentile equating involves adjusting the entire distribution of one test to the entire distribution of another test, for a given population. Scores at the same percentile on the two tests are equivalent.

Equity - Equity is the concern for fairness, i.e., that assessments are free from bias or favoritism. An assessment that is fair enables all children to show what they can do. At minimum, all assessments should be reviewed for (a) stereotypes, (b) situations that may favor one culture over another, (c) excessive language demands that prevent some students from showing their knowledge, and (d) the assessment's potential to include students with disabilities or limited English proficiency.

Evaluation - The process designed to answer an applied question related to the value or worth of something. When used for most educational settings, evaluation means to measure, compare, and judge the quality of student work, schools, or a specific educational program.

Exploratory Factor Analysis - A theory-generated procedure that identifies factors and/or factor patterns associated with variables or measures. It determines the number or nature of factors correlated to specific variables. This analysis seeks to discover factors when prior research analysis is not present.



Face Validity - The extent to which an instrument appears to be valid for making inferences on those completing it. Face validity is considered a weak form of validity and is often not considered validity at all.

Factor Analysis - A statistical technique used to examine the interrelations among a set of variables, or items, in order to identify an underlying structure to those items. This process can be confirmatory, which means that an underlying causal structure is hypothesized. Another method is exploratory factor analysis which is typically used to identify common underlying constructs among a group of variables. Factor analysis can provide evidence to show if the relations between items on a test are consistent with underlying theoretical construct or constructs.

False Negative - A person who has the condition in question, but tests negative for it. This is also considered a Type II error.

False Positive - A person who does not have the condition in question, but tests positive for it. This is also considered a Type I error.

Generalizability Theory - The basis of this approach to test theory is the ability to determine multiple sources of error in measurement using analysis of variance (ANOVA) techniques. This yields a generalizability (reliability) coefficient that may include multiple error sources, unlike Classical Test Theory, and also avoids the requirement of parallel tests. Instead, generalizability theory relies on a less restrictive assumption by randomly drawing items from the same pool of possible items.

Holistic Scoring - Evaluating student work in which the score is based on an overall impression of student performance rather than multiple dimensions of performance (**analytic scoring**).

Instrument - A device or procedure created to assess a trait or characteristic of individuals.

Internal Consistency - An index of the extent to which a set of items on a test measure the same trait or characteristic. See also Cronbach's alpha.

Inter-Rater Reliability - The extent to which multiple evaluators of a task or performance give identical ratings.

Item - An individual scenario, question or task designed to elicit a response from a test-taker. An individual question or exercise in an assessment or evaluative instrument.

Item Bank - A computerized or paper collection of test/questionnaire items, organized by subject matter, item difficulty, and question type (multiple choice, true/false, etc.). This bank assists in creating tests of questionnaires.

Item Response Theory - A measurement model that assumes the characteristic to be measured is a latent variable that causes the responses observed on a test or measure. The item characteristic curve



(ICC) graphically depicts the relation between the underlying latent characteristic and the likelihood of a given response on a given item by a given person. Estimation of the model parameters allows gives rise to estimation of the underlying characteristic for each respondent that is independent of the items chosen or the sample tested.

Kuder-Richardson Formulas - Types of measure of internal consistency for items that have only two possible answers. KR-21 gives a quick estimate of the lowest possible KR-20 for a given data set. These are special cases of Cronbach's alpha.

Likert-Type Response - An ordinal rating scale developed by Rensis Likert in efforts to measure attitude. (e.g. Strongly disagree – Strongly Agree).

Linear Equating - Linear equating involves adjusting the mean and standard deviation on one test so it is the same as on another test.

Linear Regression - Fitting a straight line into a data set to determine future outcomes. A technique used to fit a straight line into a data set. This determines the effect of an independent variable on a dependent variable.

Measurement - A logical rule for assigning numbers to observations to represent the quantity of a trait of characteristic possessed.

Moderation - This is the weakest form of linking and is often used when tests have different blueprints and are given to different groups of examinees. When procedures match distributions using scores, they are called statistical moderation links. When procedures match distributions using judgments, they are called social moderation links. Whether statistical or social, comparing scores based on moderation should only be interpreted carefully and should be used for making only general comparisons.

Norm - A standard of performance. Norms are frequently established by initially testing a large sample and then using those data to gauge the performance of individuals tested subsequently.

Norm-Referenced Assessment - An assessment where student performance or performances are compared to a larger group. Usually the larger group or "norm group" is a national sample representing a wide and diverse cross-section of students. Students, schools, districts, and even states are compared or rank-ordered in relation to the norm group. The purpose of a norm-referenced assessment is usually to sort students and not to measure achievement towards some criterion of performance.

On-Demand Assessment - An assessment that takes place at a predetermined time and place, usually under uniform conditions for all students being assessed. The SAT, district and state tests, and most in-class unit tests and final exams are examples of on-demand assessments.

Opportunity to Learn - To provide students with the teachers, materials, facilities, and instructional experiences that will enable them to achieve high standards. Opportunity to learn (OTL) is what takes



place in classrooms that enables students to acquire the knowledge and skills that are expected. OTL can include *what* is taught, *how* it is taught, *by whom*, and with *what* resources.

Ordinal scale - Rank order in which there is no zero point, and the differences between the ranks need not be equal. Used to answer items with scale responses. Compare with Likert-type response or dichotomous response.

Performance Standards - Explicit definitions of what students must do to demonstrate proficiency at a specific level on the **content standards**. For example, the performance level "exceptional achievement" on a **dimension** "communication of ideas" is reached when the student *examines the problem from several different positions and provides adequate evidence to support each position*.

Portfolio Assessment - A portfolio is collection of work, usually drawn from students' classroom work. A portfolio becomes a portfolio assessment when (1) the assessment purpose is defined; (2) criteria or methods are made clear for determining *what* is put into the portfolio, *by whom*, and *when*; and (3) criteria for assessing either the collection or individual pieces of work are identified and used to make judgments about performance. Portfolios can be designed to assess student progress, effort, and/or achievement, and encourage students to reflect on their learning.

Power - The probability that a statistical test will lead to a rejection of the null hypothesis when the null hypothesis is actually false. The additive inverse of the probability of a type II error.

Predictive Validity - A type of criterion related validity in which a test is correlated with a criterion occurring at a later point in time.

Principal Component - Uncorrelated variables that are derived from a mathematical procedure (principle component analysis) used to reduce the dimension of the data. This technique is similar to factor analysis, as it attempts to extract factors and aides in identifying new and meaningful underlying variables.

Program Evaluation - Assessment of the effectiveness of ongoing activities such as in health care delivery, education, rehabilitation, or other social programs or interventions.

Projection - This form of linking is used to predict scores on one test from scores on another test. Users of this technique need to realize that the same things are not being measured. Often statistical regression methods are used to project scores from one test to another.

Psychometrics - The study of how the measurement of psychological information is operationalized.

Qualitative Data - Information presented in narrative form.

Quantitative Data - Information presented in numerical form.



Questionnaire - An assessment device containing written questions that is used for the purpose of gathering data from an individual. See also survey.

Random Error - Measurement error that is attributed to random causes.

Rater Training - The process of educating **raters** to evaluate student work and produce dependable scores. Typically, this process uses **anchors** to acquaint raters with criteria and scoring rubrics. Open discussions between raters and the trainer help to clarify **scoring criteria** and **performance standards**, and provide opportunities for raters to practice applying the rubric to student work. Rater training often includes an assessment of rater reliability, that raters must pass in order to score actual student work.

Rater - A person who evaluates or judges student performance on an assessment against specific criteria.

Raw Score - The unstandardized score for a test or measure. It is typically not interpretable without additional information for reference.

Reliability - A degree of consistency of a measure; the degree to which a test is free of random error. A test or measure that produces consistent results is said to have high reliability. Reliability estimation is achieved through the correlation between the observed and true score, and is known as the reliability index. A high value for the correlation indicates high reliability. Two assumptions are important for reliability estimation. The first is the assumption of independence, which states that the true score is unrelated to, or independent of, the error score. This assumption allows estimation of the square of the reliability index (ratio of true variance to observed variance) through parallel testing. This leads to the parallel tests assumption: for two tests designed to test the same construct on the same subjects, the T remains the same. Specifically, scores on both tests have the same variance; and the errors on both tests are mutually independent.

The degree to which the results of an assessment are dependable and consistently measure particular student knowledge and/or skills. Reliability is an indication of the consistency of scores across **raters**, over time, or across different **tasks** or **items** that measure the same thing. Thus, reliability may be expressed as (a) the relationship between test items intended to measure the same skill or knowledge (item reliability), (b) the relationship between two administrations of the same test to the same student or students (test/retest reliability), or (c) the degree of agreement between two or more raters (rater reliability). An unreliable assessment cannot be valid.

Scale - Values given to student performance. Scales may be applied to individual items or performances, for example, *checklists*, i.e., yes or no; *numerical*, i.e., 1-6; or *descriptive*, i.e., the student presented multiple points of view to support her essay. Scaled scores occur when participants' responses to any number of items are combined and used to establish and place students on a single scale of performance.

Select Response - An assessment method in which testers are given predefined choices from which to choose in order to answer a given item.



Self-Administered - A method by which a measure is answered by the tester and which does not require a trained individual to administer or supervise its completion. Compare this method of test-taking to a measure that requires administration a licensed psychologist in order to rate the response.

Sensitivity - The ability of a test to detect a condition when, in fact, it is present.

Specificity - The ability of a test to detect that a condition is not present when it is, in fact, not present.

Split-Half Reliability - A measure of the reliability of a measures based on the correlation between halves of the measure. The estimate is corrected for test length using the Spearman-Brown formula, which assumes the halves are parallel, or by generalizability, which assumes the halves are essentially-tau-equivalent, or represent a random sample of content from the domain.

Stability - A measure of reliability, known as the test-retest approach. It is often used when alternate forms of tests are not available. Tests that yield the same results over time are said to be stable, and hence, reliable.

Standardization - A consistent set of procedures for designing, administering, and scoring an assessment. The purpose of standardization is to assure that all students are assessed under the same conditions so that their scores have the same meaning and are not influenced by differing conditions. Standardized procedures are very important when scores will be used to compare individuals or groups.

Standardized Item Alpha - The average correlation among all pairs items; an indicator of the reliability of the standardized items.

Standards - The broadest of a family of terms referring to statements of expectations for student learning, including **content standards**, **performance standards**, and **benchmarks**.

Standards-Based Reform - A program of school improvement involving setting high standards for all students and a process for adapting instruction and assessment to make sure all students can achieve the standards.

Stem - Also known as question stem and item prompt. It is the item or question used to elicit a multiple-choice response.

Students With Disabilities (SWD) - A broadly defined group of students with physical and/or mental impairments such as blindness or learning disabilities that might make it more difficult for them to do well on assessments without **accommodations** or **adaptations**.

Survey - A technique used to determine specific information about a sample of individuals. It may or may not contain measures of constructs.



Systemic Error - Error introduced into a measure that is consistent in form; also known as bias.

Tailored Testing - A testing method, used in computer adaptive testing, in which the test complexity is structured to the test taker's ability. It attempts to determine the examinee's true ability.

Task - An activity, exercise, or question requiring students to solve a specific problem or demonstrate knowledge of specific topics or processes.

Test-Retest Reliability - Under the assumptions of equal true scores and uncorrelated errors, the correlation between two administrations of a test given to the same individuals at two different points in time is an estimate of the test's reliability.

True score - In Classical Test Theory, the true score on a test is the expected score of an individual across all conditions.

Type I error - The error made when a true null hypothesis is rejected.

Type II error - The error made when a false null hypothesis is not rejected.

Validity - The degree to which an instrument measures what it is intended to measure. Note: reliability is a necessary but insufficient condition for validity. The extent to which an assessment measures what it is supposed to measure and the extent to which inferences and actions made on the basis of test scores are appropriate and accurate. For example, if a student performs well on a reading test, how confident are we that that student is a good reader? A valid standards-based assessment is aligned with the **standards** intended to be measured, provides an accurate and reliable estimate of students' performance relative to the standard, and is fair. An assessment cannot be valid if it is not reliable.