

Glossary of Educational Research and Assessment Terminology

Reliability – Consistency of data. Inter-rater consistency, intra-rater consistency, estimate of the consistency of scores on a similar assessment. Reliability relates to DATA derived from an assessment tool, not the tool itself.

Validity – Established through the preponderance of evidence that the assessment tool assesses what is claimed that it assesses. Types of validity include predictive, concurrent criterion, construct

Cut score – a number that separates two categories of data. Typically, a cut score separates pass and fail, but it could also separate honors from pass.

Generalizability – similar to reliability, but includes specifics of particular assessment event. It is the estimated correlation between performance on the current measure and another measure of similar content, format and operations.

Compensatory grading – the policy of allowing good performance in one area to overcome poor performance in another area.

Non compensatory grading – the policy where sufficient performance must be achieved in each and every area.

Standard deviation – index of variability in a set of numbers. Computationally, this is the square root of the sum of deviations of each score and the mean, divided by the number of data points in the set.

Standard score – same as a z score; a deviation from the mean in standard deviation units

Scaled score – conversion of a set of scores to a distribution (scale) that has a selected mean and standard deviation, eg USMLE: mean is 200 and SD is 20.

Domain – Is the articulated breadth and scope of the intended learning from which assessment tasks are created. Typically, a sample of assessment tasks is presented to the learner from which one predicts or generalizes to performance on

all assessment tasks from the breadth and scope of expected learning.

IRT – “Item Response Theory.” The theory is that any given item (typically a multiple choice item) has 3 components that contribute the performance on that item: the item, the test-taker and the interaction of the item and test-taker. With sufficiently large data set on multiple items by the same test takers, formulas can partition a test-taker’s real ability from any influence of the item. Typically testing within an educational program has too few data points to compute stable IRT parameters.

Z score – a way of dividing the area under a distribution where the mean of the distribution = 0 based on the standard deviation of the set of numbers; also is a method for identifying where a person’s performance is relative to others in that group. ± 1.0 z score from the mean has 68% of the scores in that range. ± 1.96 z scores has 95% of the scores between those two numbers. A standard deviation of a group of numbers is equal to a z score of 1.0.

Deviation – the difference between two numbers; between the mean score and the lowest score, between the highest score and the next highest score, etc. Traditionally, it is the difference between any number in a group of numbers and the mean of that group.

Average deviation – the sum of each deviation of a number from the mean, divided by the count of the numbers in the group.

Mean – the average score for a set of numbers

Median – the middle-most number of a set of numbers. With an even number of scores, the median is the average of the two middle-most scores.

Mode – the most frequently occurring number in a set of numbers

Educational Research and Assessment Terminology

Normal curve/normal distribution – the “bell” curve. It is the assumed distribution of scores that underlies parametric statistics. The scores may be spread out “flat” or peaked around the mean. Educational assessment data rarely is in the form of a normal distribution.

Positively skewed – the descriptor for a set of numbers where most are bunched toward the high end with a “tail” of numbers toward the lower end

Negatively skewed – the descriptor for a set of numbers where most are bunched at the lower end with a “tail” of numbers reaching toward the high end.

Standard error of measurement (SEM) – This is the estimate of the variance of a person’s scores if the person took many tests of a similar size. It is computed with the reliability coefficient of a test and the standard deviation of the set of obtained scores.

Standard error of the mean – This is the estimate of variance of a group’s mean score, if many same-sized samples were selected and their means computed.

Confidence interval – a way of expressing the range within which a “real score” exists. This recognizes that there is “noise” in the measurement of a person’s performance or characteristics and gives a range and likelihood that the real score is in that range. An example is, a student scores 80% on a test. If the student were to take several parallel tests her scores would not be identical. We would say that her “true” ability would fall within the range of ± 1.96 SDs, 95% of the time. We would not use z scores, but rather actual scores to indicate the high and low ends of the range.

Methods for setting a cut score – The following were developed for the multiple-choice item format as a method to determine what the pass/fail score should be.

Nedelsky – one or more people judge the number of test-takers in a “borderline” group (usually 10) likely to get an item correct. The proportion for each item is added and then rounded to an integer. This is the cut score for pass/fail.

Angoff (modified) – one or more people judge which options of an answer a set of borderline students could eliminate as incorrect. The proportion of remaining options of the item that these borderline students would guess to be correct is determined. If borderline students could eliminate one option out of five for a given item, then they would have a 25% chance of guessing the correct answer from the remaining 4. These percentages are added for all items on the test and rounded to an integer to produce the cut score for the test. A “modified” Angoff procedure would then give the judges information about how many students would have failed the tests and an opportunity to modify their estimates to yield a more “reasonable” number of failing students.

Ebel – The method takes into account the difficulty and importance of each item when setting the cut score. First items are sorted into 6 cells formed from High, Medium and Low difficulty and High, Medium and Low importance. Judges then estimate the number of borderline students would likely get each item in a cell correct. The estimated proportions are added for all items, rounded to an integer and is the cut score. Data regarding the results from a group of students using this cut score may be provided and judges be allowed to modify their estimates.

Hofstee (compromise method) – This method combines two frames of reference to determine a cut score that compromises between a normative and criterion referenced cut score. Judges estimate the acceptable number of persons failing and the highest number of items answered correctly, yet the student would still fail. The answers to these and two other questions

Educational Research and Assessment Terminology

are plotted on the cumulative distribution of scores to determine the number of items, if failed, compromises the number of students who could fail and number of items that could be missed.

Norm-referenced – A framework for comparing test-takers where the interest is knowing the highest (or lowest) performance, regardless of the absolute score. An example is looking for the top 10 students' scores on a test in order to give awards for the "best in the group."

Criterion-referenced (domain-referenced) – A framework for comparing test-takers' performance with an expected performance based on the kinds of items or tasks in the assessment. Mastery learning calls for this kind of comparison where one or more instructors decide on a (minimum) level of performance to pass and all test-takers are judged by that score. All could pass or all could fail.

Summative assessment – An assessment administered at the end of a learning segment, course or program to determine whether each learner has learned a sufficient amount to pass. This is often referred to as a "go – no go" assessment.

Formative assessment – An assessment administered during a course of study to provide the learners and instructors an idea of how well each learner is achieving what is expected. Feedback is given to each learner about strengths and weaknesses so that they can improve.

Item analysis – several characteristics of (usually) multiple-choice items that indicate the quality of the item and of the whole test. The implicit frame of reference is normative. Item characteristics include the proportion of high-scoring test takers getting the item correct, the number of low-scoring test takers getting the item correct, an index of discrimination for each item, a difficulty index for each item, the

number of test-takers selecting each option of each question, and the non-distracters. The test's reliability is also reported.

Discrimination index – Index of the item's ability to separate generally high scoring test-takers from generally low scoring test-takers. It is the point biserial correlation of high and low total scores with getting the item correct or not. Usually the upper and lower 27% of test-takers, based on total test scores, are used for this analysis to maximize the variance between high and low performance. For norm referenced tests, a discrimination index of .50 is desired.

Difficulty index – This is simply the proportion of all test-takers that got an item correct.

Distracters – options for multiple-choice items that are incorrect, but close enough to the correct answer to be selected by test-takers who do not know as much as those who select the correct option. These are also called "foils."

Classification – Putting test-takers in categories on the basis of test scores, eg, they could be classified as "fail" or "pass." "Index of classification" is an estimate of the consistency of classifying test-takers on a parallel set of items.

Sensitivity – the ability of a test to identify everyone in a particular group and may include some people who should not be that group.

Specificity – identifying only those that belong in a certain group, but may miss some who should be in that group

KR20 – one of many KR formulas for estimating test reliability based on item statistics. Assumes item difficulties & inter-item correlations are relatively equal. Conceptually, this formula subtracts all of the individual item variances from the variance of the total test to yield an estimate of the test's reliability (internal consistency). There is a correction for tests where item difficulties are not similar called

Educational Research and Assessment Terminology

Horst's modification and is based on an estimate of the maximum variance possible for a test with a given range of item difficulties.

$$\text{KR20: } r_{tt} = \left[\frac{N}{n-1} \right] \left[\frac{\sigma_t^2 - \sum pq}{\sigma_t^2} \right]$$

Where p = proportion passing an item and $q = 1-p$ (proportion not passing the item) and σ_t^2 is the variance of the whole test.

$$\text{Max. variance est. } \sigma_m^2 = 2R_i p_i - M_t \left[1 - M_t \right]$$

Where R_i is an item's rank order when all items are ordered from easiest (1) to most difficult. M_t is the mean of the total test and p_i is the mean score of an item. The maximum variance term - σ_m^2 - is then used in the Horst modification formula.

$$\text{Horst Mod. } r_{tt} = \left(\frac{\sigma_t^2 - \sum pq}{\sigma_m^2 - \sum pq} \right) \left(\frac{\sigma_m^2}{\sigma_t^2} \right)$$

Relationship of variance to stand. deviation -

Variance = the sum of the squared deviations of each item from the test mean divided by the number of items in the test.

Standard deviation = the square root of the variance.

$$\text{Variance } \sigma^2 = \frac{\sum (\bar{X} - X)^2}{n}$$

Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$