

The ABCs of Assessment

A parameter:

In item response theory (IRT), the A parameter is a number that indicated the discrimination of a test item. If the A parameter for a test item is large, the probability that the test taker will answer the item correctly increases sharply within a fairly narrow range of ability. If the A parameter is small, the probability of a correct answer increases gradually over a wide range of ability.

Ability:

The knowledge, skills, or other characteristics of a test taker measured by the test.

Adaptive testing:

Type of testing in which the questions presented to the test taker are selected on the basis of the test taker's previous responses. Good performance by the test taker leads to harder questions; poor performance leads to easier questions. The purpose of adaptive testing is to use testing time more efficiently by not giving test takers questions that are much too easy or too difficult.

Alpha (or α) coefficient:

A statistic that is used to estimate the reliability of scores on a test. What alpha actually measures is internal consistency – the extent to which the test takers performed similarly on all of the items. Under some reasonable assumptions, alpha also indicated the extent to which those test takers would perform similarly on two different forms of the same test. The alpha coefficient is commonly used to indicate the reliability of scores in tests in which the questions all measure the same general type of knowledge skill.

Analytic scoring:

A procedure for scoring responses on a constructed-response test, in which the scorer awards points separately for specific features of the response.

Anchor test:

For equating the scores on two test forms of a test that are taken by different groups of test takers, it is necessary to know how those groups differ in the ability measured by the test. An anchor test is a test given to both groups to obtain this information. The anchor test can be a set of test questions appearing in both forms (called "common items"), or it can be a separate test taken by both groups.

Assessment, test, examination:

Terms that refer to devices or procedures for getting information about knowledge, skills, or other characteristics of the people being assessed, tested, or examined. The three terms are often used interchangeably, but there are some differences between them. Assessment is the broadest of the three terms, examination is the narrowest.

B parameter:

In item response theory (IRT), the B parameter is a number that indicates the difficulty of a test question. In general, a higher B parameter indicates a more difficult test item.

C parameter:

In item response theory (IRT), the C parameter is a number that indicates the probability that a test taker with no knowledge of the subject will answer the question correctly – determined empirically, not simply assumed to be a particular value.

Calibration:

The meaning of the term depends on the context. In item response theory (IRT), calibration refers to the process of estimating the numbers (called parameters) that describe the statistical characteristics of each test question. In the scoring of a constructed response test, calibration refers to the process of checking to make sure that each scorer is applying the scoring standards correctly.

Classical test theory:

A statistical theory that forms the basis for many calculations done with test scores, especially those involving reliability. The theory is based on partitioning a test taker's score into two components: a component called the "true score" that generalizes to other occasions of testing with the same test, and a component called "error of measurement" that does not generalize. The size of the error of measurement component is indicated by the standard error of measurement.

Comparable:

Two scores are comparable if they can be meaningfully compared. Raw scores on different forms of a test are not comparable because the questions on one form can be more difficult than questions on another form. Scaled scores on different forms of a test are comparable if the process of computing them includes equating. Percentile scores are comparable if they refer to the same group of test takers.

Computer-adaptive testing:

Adaptive testing that is conducted with the aid of a computer. For practical and logistical reasons, most adaptive tests are delivered by computer.

Confidence interval:

A range of possible values for an unknown number (such as a test taker's score), computed in such a way as to have a specified probability of including the unknown number. That specified level is called the confidence level and is usually a high number, typically 90 to 95 percent.

Constructed-response item:

A test question that requires the test taker to supply the answer instead of choosing it from a list of possibilities.

Constructed-response test:

Any test in which the test taker must supply the answer to each question, instead of choosing it from a list of possibilities. The term constructed-response test usually refers to a test that calls for responses that can be written on a paper or typed into a computer. Tests calling for

responses that cannot be written on paper or typed into a computer are usually referred to as performance assessments.

Converted score:

A test score that has been converted into something that is not a raw score. One common type of converted score is a “scaled score” - a score that has been transformed onto a different set of number from those of the raw scores, usually after equating to adjust for the difficulty of the test questions. Another common type of converted score is a percentile score. Instead of converted score, the term derived score is often used.

Correlation:

A statistic that indicates how strongly two measures, such as test scores, tend to vary together. If the correlation between scores on two tests is high, test takers tend to have scores that are about equally above average (or equally below average) on both tests. The correlation will range from -1.00 to +1.00. When there is no tendency for the scores to vary together, the correlation is .00.

Criterion referencing:

Making test scores meaningful without indicating the test taker’s relative position within a group. On a criterion-referenced test, each individual test taker’s score is compared with a fixed standard, rather than with the performance of the other test takers. Criterion referencing is often done in terms of proficiency levels. The test score required to attain each proficiency level is specified in advance. The percentages of test takers at the different proficiency levels are not fixed; they depend on how well the test takers perform on the test.

Cut score:

Also known as a passing score, the cut score is the score that a candidate must achieve to obtain a certain classification, such as basic, proficient, or advanced.

Distractors:

Distractors are the incorrect options of a multiple-choice item. A distractor analysis is an important part of psychometric review, as it helps determine if one is acting as a keyed response.

Equating:

The process of determining comparable scores on different forms of an examination. Equating makes it possible to report scaled scores that are comparable across different forms of the test.

Form:

A specific set of items that are administered together for a test. For example, if a test included a certain set of 100 items this year, and a different set of 100 items next year, these would be two distinct forms.

Item:

The basic component of a test, often colloquially referred to as a “question,” but items are not necessarily phrased as a question. They can be as varied as true/false statements, rating scales, and performance task simulations, in addition to the ubiquitous multiple-choice item.

Item Bank:

A repository of items for a testing program, including items at all stages, such as newly written, reviewed, pretested, active, and retired.

Item Banker:

A specialized software program that facilitates the maintenance and growth of an item bank by recording item stages, statistics, notes, and other characteristics.

Item Response Theory (IRT):

A comprehensive approach to psychometric analysis and test development that utilizes complex mathematical models. This provides several benefits, including the ability to design CATs, but requires larger sample sizes. A common rule of thumb is 100 candidates for the one-parameter model and 500 for the three-parameter model.

Key:

The key is the correct response to an item.

Mean (of test scores):

The average, computed by summing the test scores of a group of test takers, then dividing by the number of test takers in the group.

Median (of test scores):

The point on the score scale that separates the upper half of the test takers from the lower half. The median has a percentile rank of 50.

Non-cognitive assessment:

Attempts to measure traits and behaviors other than the kinds of knowledge and skills measured by traditional academic texts – traits such as perseverance, self-confidence, self-discipline, communication skills, et al.

Norm-Referenced:

A test score (not a test) is norm-referenced if it is interpreted with regard to the performance of other candidates. Percentile rank is an example of this, because it does not provide any information regarding how many items the candidate got correct.

Normalization:

Transforming test scores onto a score scale so as to produce a score distribution that approximates the symmetric, bell-shaped distribution, called a normal distribution. Normalization is a type of scaling.

Norms:

Statistics that describe the performance of a group of test takers for the purpose of helping test takers and test users interpret the scores. Norms information is often reported in terms of percentile ranks.

P-value (or P+):

A classical index of item difficulty, presented as the proportion of candidates who correctly responded to the item. A value above 0.90 indicates an easy item, while a value below 0.50 indicates a relatively difficult item. Note that it is inverted; a higher value indicates *less* difficulty.

Percentile score rank:

A test score that indicates the test taker's relative position in a specified group. A test taker's percentile score (or percentile rank) is a number from 1 to 100, indicating the percent of the group with scores no higher than the test taker's score. The most common way to compute the percentile score is to compute the percent of the group with lower scores, plus half the percent with exactly the same score as the test taker.

Performance level descriptor:

A statement of the knowledge and skills a test taker must have to be classified at a particular performance level, such as basic, proficient, or advanced.

Point-Biserial Correlation:

A classical index of item discrimination, calculated as the Pearson correlation between the item score and the total test score. If below 0.0, low-scoring candidates are actually doing better than high-scoring candidates, and the item should be revised or retired. Low positive values are marginal, higher positive values are ideal.

Pretest (or Pilot) Item:

An item that is administered to candidates simply for the purposes of obtaining data for future psychometric analysis. The results on this item are not included in the score. It is often prudent to include a small number of pretest items in a test.

Rasch model:

A set of assumptions for item response theory. The Rasch model assumes that a test taker's probability of answering a test question correctly depends on the test taker's ability and on only one characteristic of the test question – its difficulty.

Raw score:

A test score that has not been adjusted to be comparable with scores on other forms of the test and is not expressed in terms of the performance of a group of test takers. The most common types of raw scores are the number of questions answered correctly, the percentage of questions answered correctly, and, on a constructed-response test, the sum of the ratings assigned by scores to a test taker's responses.

Reliability:

A measure of the repeatability or consistency of the measurement process. It is the tendency of test scores to be consistent on two or more occasions of testing. If there is no real change in the test taker's knowledge. If a set of scores has high reliability, the test takers' scores would tend to agree strongly with their scores on another occasion of testing.

Rubric:

A set of rules for scoring the responses on a constructed response test. Also called a scoring guide.

Scaling:

Statistically transforming scores from one set of numbers (called the scale score) to another. Some types of scaling are used to make scores of different tests comparable.

Standard deviation (of test scores):

A measure in the amount of variation in the scores of a group of test takers. It is the average distance of the scores from the group mean score. The standard deviation is expressed in the same units as the scores, e.g. number of correct answers, or scaled-score points. If there are many high and low scores the standard deviation will be large. If the scores are bunched closely together the standard deviation will be small.

Standard error of measurement (SEM):

Measure of the tendency of test takers' scores to vary because of random factors, such as the particular selection of items on the form the test taker happened to take, or the particular scorers who happened to score the test taker's responses. The smaller the SEM, the smaller the influence of these factors. The SEM is expressed in the same units as the scores themselves.

Standard-Setting Study:

A formal study conducted by a testing organization to determine standards for a testing program, which are manifested as a cut score.

Subject Matter Expert (SME):

An extremely vital person in the test development process. SMEs are necessary to write items, review items, participate in standard-setting studies and job analyses, and oversee the testing program to ensure its fidelity to its true intent.

Summative assessment:

Assessing students' skills for the purpose of determining whether instruction has been effective.

Validity:

Validity is the extent to which scores on a test are appropriate for a particular purpose. The validity of the scores depends on the way they are being interpreted and used. Scores on a test can be highly valid for one purpose and much less so for another. Statistics can provide evidence for the validity of a test, but the validity of a test cannot be measured by a single statistic.