Glossary of Standardized Testing Terms

https://www.ets.org/understanding_testing/glossary/

_α_ parameter
In item response theory (IRT), the *a* parameter is a number that indicates the discrimination of a test item — how sharply the item differentiates between generally strong and generally weak test takers. If the *a* parameter for an 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
A 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 any questions that are much too easy or much too difficult for them. Adaptive testing requires special procedures for computing test takers' scores, because many different combinations of questions are possible, and some test takers get more difficult questions than others.

Alpha 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 the items. Under some assumptions that are usually reasonable, alpha also indicates 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 on tests in which the questions all measure the same general type of knowledge or skill. However, it can also be used to indicate halo effect among ratings that are intended to measure different characteristics of the people being rated.

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. (Compare with holistic scoring.)

Anchor test
For equating the scores on two 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
These terms all refer to devices or procedures for getting information about the 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 question.

Biserial correlation
A statistic used at ETS to describe the relationship between performance on a single test item and on the full test. It is an estimate of the correlation between the test score and an unobservable variable assumed to determine performance on the item and assumed to have a normal distribution (the familiar "bell curve"). Compare to correlation, point biserial correlation.

***c* parameter**
In item response theory (IRT), the *c* parameter is a number that indicates the probability that a test taker with little or no knowledge of the subject will answer the question correctly.

Calibration
The meaning of this 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.

Claims
Statements about the knowledge, skills or abilities of test takers who have attained a specified level of performance on the test. Claims communicate the meaning of the test scores. Claims may be general (for example, "The test taker can read at the second grade level.") or specific (for example, "The test taker can decode initial consonants.").

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 estimated using the standard error of measurement.

Classification error
See decision error.

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 the 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 true score), computed in such a way as to have a specified probability of including the unknown number. That specified probability is called the "confidence level" and is usually high, typically 90 or 95.

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 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 other than a raw score. One common type of converted score is a "scaled score" — a score that has been transformed onto a different set of numbers 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 can range from -1.00 to +1.00. When there is no tendency of the scores to vary together, the correlation is .00.

Criterion referencing
Making test scores meaningful without indicating the test taker's relative position in 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 defined 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. (Compare with norm referencing.)

Cutscore
A point on the test score scale used for classifying the test takers into groups on the basis of their scores. Sometimes these classifications are used only to report statistics, such as the percent of students classified as proficient in a subject. More often, the classifications have consequences for individual test takers — consequences such as being granted or denied a license to practice a profession. (See also performance level descriptor.)

Decision error
When test takers' scores are compared with a specified cut score, two kinds of decision errors are possible: (1) a test taker whose true score is above the cut can get a score below the cut; (2) a test taker whose true score is below the cut can get a score above the cut. It is possible to modify the decision rule to make one kind of decision error occur less often, but only at the cost of making the other kind of decision error occur more often. Also called "classification error."

Dichotomously scored item
An item for which there are only two possible scores, most often 1 for a correct answer and 0 for any other response. Compare with polytomously scored item.

Differential item functioning (DIF)
Differential item functioning (DIF) is the tendency of a test question to be more difficult (or easy) for certain specified groups of test takers, after controlling for the overall ability of the groups. It is possible to perform a DIF analysis for any two groups of test takers, but the groups of test takers ETS is particularly concerned about are female test takers and test takers from specified ethnic groups. ETS refers to those groups as "focal groups." For each focal group, there is a corresponding "reference group" of test takers who are not members of the focal group. A DIF analysis asks, "If we compare focal-group and reference-group test takers of the same overall ability (as indicated by their performance on the full test), are any test questions significantly harder for one group than for the other?"

Discrimination
Outside the testing context, this term usually means treating people differently because they are members of particular groups, e.g., male and female. In the testing context, discrimination means something quite different. It refers to the power of a test or (more often) a test question to separate high-ability test takers from low-ability test takers.

Distracters (or distractors)
In a multiple-choice test item, the distracters are the wrong answers presented to the test taker along with the correct answer. Writers of test questions often use distracters that represent common mistakes or misinformation.

Equating
Statistically adjusting scores on different forms of the same test to compensate for differences in difficulty (usually, fairly small differences). Equating makes it possible to report scaled scores that are comparable across different forms of the test.

Evidence-centered design
An approach to constructing educational assessments that uses evidentiary arguments to reveal the reasoning underlying the design of the test. The test designers begin with an analysis of the types of evidence necessary to make valid claims about what test takers know or can do.

Formative assessment
Assessing students' skills for the purpose of planning instruction for those students. Formative assessment is done before instruction begins and/or while it is taking place. (Compare with summative assessment.)

Formula scoring
A scoring rule in which each wrong answer reduces the test-taker's total score by a fraction of a point. That fraction is chosen to make the test-taker's expected gain from random guessing equal to zero. Compare with number-correct scoring.

Grade-equivalent score
A type of norm-referenced score expressed in terms of the performance typical of students at a particular grade level, at a particular point in the school year. For example, a grade-equivalent score of 4.2 implies that the test taker's performance on the test would be typical for students in the second month of their fourth-grade year. (See norm referencing.)

Halo effect
When raters are being asked to rate people on several different qualities, they sometimes tend to rate each person similarly on all those qualities, without recognizing that some people are high on some qualities and low on others. The tendency of raters to ignore these kinds of differences is called "halo effect."

Holistic scoring
A procedure for scoring responses on a constructed-response test, in which the scorer makes a single judgment of the overall quality of the response, instead of awarding points separately for different features of the response. (Compare with analytic scoring.)

Item
A test question, including the question itself, any stimulus material provided with the question, and the answer choices (for a multiple-choice item) or the scoring rules (for a constructed-response item).

Item analysis
Statistical analyses of test takers' responses to test questions, done for the purpose of gaining information about the quality of the test questions.

Item banking
Creating and maintaining a data base of test questions. The record for each question includes the text of the question and statistical information computed from the responses of test takers who have taken it.

Item response theory (IRT)
A statistical theory and a set of related methods in which the likelihood of achieving each possible score on a test question depends on one characteristic of the test taker (called "ability") and a small number (usually three or fewer) of characteristics of the test question. These characteristics of the test question are indicated by numbers called "parameters." They always include the difficulty of the question and often include its discrimination (the sharpness with which it separates stronger from weaker test takers). Some ETS testing programs use IRT for item analysis, item banking and score equating.

Mean (of test scores)
The average, computed by summing the test scores of a group of test takers and 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 a group of test takers from the lower half. The median has a percentile rank of 50.

Multiple-choice item
A test question that requires the test taker to choose the correct answer from a limited number of possibilities, usually four or five. (Compare with constructed-response item.)

Noncognitive assessment
Attempts to measure traits and behaviors other than the kinds of knowledge and skills measured by traditional academic tests — traits such as "perseverance, self-confidence, self-discipline, punctuality, communication skills, social responsibility and the ability to work with others and resolve conflicts" (R. Rothstein, The School Administrator, December, 2004; www.aasa.org/publications).

Norm referencing
Making test scores meaningful by providing information about the performance of one or more groups of test takers (called "norm groups"). A norm-referenced score typically indicates the test taker's relative position in the norm group. One common type of norm-referenced score is a percentile score. Another type is a "standard score," which indicates the test taker's relative position in terms of the mean (average score) and standard deviation of the scores of the group. (Compare with criterion referencing.)

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.

Normal distribution
The symmetrical, bell-shaped distribution commonly used in many statistical and measurement applications, especially in computing confidence intervals including score bands.

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

Number-correct scoring
Computing the total score by counting the number of correct answers, with no penalty for incorrect answers. Also referred to as “number-right scoring” or “rights scoring.” Compare with formula scoring.

Objective scoring
A scoring system in which a response will receive the same score, no matter who does the scoring. No judgment is required to apply the scoring rule. Compare with subjective scoring. Also see analytic scoring and holistic scoring.

Percentile score (percentile rank)
A test score that indicates the test taker's relative position in a specified group. A test taker's percentile score (also called "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 percentage of the group with lower scores, plus half the percentage with exactly the same score as the test taker. (Sometimes none of the test takers with exactly that score are included; sometimes all of them are.) Percentile scores are easy for most people to understand. However, many people do not realize that averages or differences of percentile scores can be very misleading. For example, the difference between percentile scores of 90 and 95 nearly always represents a larger difference in performance than the difference between percentile scores of 45 and 55. Comparisons of percentile scores are

meaningful only if those percentile scores refer to the same group of test takers tested on the same test.

Performance assessment
A test in which the test taker actually demonstrates the skills the test is intended to measure by doing real-world tasks that require those skills, rather than by answering questions asking how to do them. Typically, those tasks involve actions other than marking a space on an answer sheet or clicking a button on a computer screen. A pencil-and-paper test can be a performance assessment, but only if the skills to be measured can be exhibited, in a real-world context, with a pencil and paper. (Compare with constructed-response test.)

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
The actual correlation between a dichotomous variable (a variable with only two possible values) and a variable with many possible values. Compare to correlation, biserial correlation.

Polytomously scored item
An item for which there are more than two possible scores (for example, an item with possible scores of 0, 1, 2 or 3). Compare with dichotomously scored item.

Portfolio
A systematic collection of materials selected to demonstrate a person's level of knowledge, skill or ability in a particular area. Portfolios can include written documents (written by the person being evaluated or by others), photos, drawings, audio or video recordings, and other media. Often the types of documents and other media to be provided are specified in detail.

Psychometrician
An expert in the statistical operations associated with tests of psychological characteristics, mental abilities, or educational or occupational knowledge and skills.

Rasch model
A type of item response theory that assumes that a test-taker's probability of answering a test question correctly depends on only one characteristic of the test question, its difficulty. Compare to item response theory.

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 scorers to a test taker's responses. (Compare with converted score.)

Reliability
The tendency of test scores to be consistent on two or more occasions of testing, if there is no real change in the test takers' 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. The type of reliability ETS is most often concerned about is consistency across different forms of a test. For a constructed-response test, ETS is also concerned about the consistency of the scores assigned by different scorers (called "scoring reliability" or "inter-rater reliability").

Reliability coefficient
A statistic that indicates the reliability of test scores; it is an estimate of the correlation between the scores of the same test takers on two occasions of testing with the same test (typically with different forms of the test).

Rights scoring
See number-correct scoring.

Rubric
A set of rules for scoring the responses on a constructed-response item. Sometimes called a "scoring guide."

Scaling
Statistically transforming scores from one set of numbers (called a "score scale") to another. Some types of scaling are used to make scores on different tests comparable in some way. The most common application of scaling is to make scores on different editions ("forms") of the same test comparable. Sometimes tests in different subjects are scaled to be comparable for a particular group of test takers. Sometimes tests at different difficulty levels in the same subject are scaled so that scaled scores on the tests at any two adjacent levels (e.g., grade levels) will reflect the same degree of proficiency in the subject; this type of scaling is called "vertical scaling."

Score band
An interval around a test taker's score, intended to convey the idea that an individual's score on a test is influenced by random factors. Often, the boundaries of the score band are one standard error of measurement above and below the test taker's actual score. (A score band determined in this way is a confidence interval with a confidence level, assuming a normal distribution, of 68 percent.) Score bands illustrate the limited precision of the test score as a measure of anything beyond the test taker's performance on one occasion of testing. However, score bands can be misleading in two ways. They imply that the test taker's true score cannot lie outside the band, and they imply that all values within the band are equally likely values for the test taker's true score. Neither of these implications is correct.

Selected-response item
Any type of test item in which the test-taker's task is to select the correct answer from a set of choices. Multiple-choice items, true-false items and matching items are all selected-response items. Compare with constructed-response item.

Standard deviation (of test scores)
A measure of 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 (but with the average distance computed by a procedure called "root-mean-square," which is a bit more complicated than the usual procedure). 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)
A 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 a 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
The process of choosing cutscores on a test.

Standardized test
A test in which the content and format of the test and the conditions of testing (such as timing, directions, use of calculators) are controlled to make them the same for all test takers. (Exceptions may be made for test takers with disabilities.)

Stanine score
A type of norm-referenced score, in which the only possible scores are the whole numbers from 1 to 9. The score scale is defined so that each score level will include a specified percentage of the norm group: small percentages for the highest and lowest levels; large percentages for the middle levels. (See norm referencing.)

Subjective scoring
Any scoring system that requires judgment on the part of the scorer. With subjective scoring, different scorers could possibly assign different scores to the same response. Compare with objective scoring. Also see analytic scoring and holistic scoring.

Summative assessment
Assessing students' skills for the purpose of determining whether instruction has been effective. Summative assessment is done after the instruction has been completed. (Compare with formative assessment.)

True score

In classical test theory, a test taker's true score on a test is defined as the average of the scores the test taker would get, averaging over some very large set of theoretically possible conditions of testing — for example, all possible forms of the test, or all possible scorers that might score the responses. It is not possible to know an individual test taker's true score, but it is possible to estimate the true scores of a large group of test takers.

Validity

Validity is the extent to which the 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 valid 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. Evidence for validity can include:

- statistical relationships of test scores with other information (e.g., scores on other tests of the same or related abilities, school grades, ratings of job performance)
- statistical relationships between parts of the test
- statistical indicators of the quality and fairness of the test questions
- the qualifications of the test designers, question writers and reviewers
- the process used to develop the test
- experts' judgments of the extent to which the content of the test matches a curriculum or the requirements of a job