

Important Assessment Vocabulary

Buros Institute on Mental Measurements
<http://www.unl.edu/buros/bimm/html/lesson01.html>

ABSOLUTE MEANING: Interpreting the meaning of an examinee's responses on an instrument by studying the content of the items without regard to what other examinees have scored on the instrument.

CONSTRUCTED RESPONSE FORMAT: A test question in which the examinee is expected to respond to an item using his or her own ideas and words rather than choosing an answer from among two or more options.

CONCURRENT VALIDITY: The extent to which individuals' current status on a criterion can be estimated from their current performance on an assessment instrument. For example, we can sample students already in college, give them a special aptitude test, and collect their current grade point averages. The relationship between the grades and the test is concurrent validity evidence, because the two measures were collected at about the same time.

CONSTRUCT: The theoretical concept or idea that explains the meaning of an examinee's response to the items in a test.

DIFFERENTIAL ITEM FUNCTIONING: An approach to studying test fairness at the level of individual test items rather than looking simply at average difference in an item's performance. The approach studies whether persons of the same ability performed differently on the item. For example, you would study how boys of low ability compared to girls of low ability, boys of average ability compared to girls of average ability, and boys of high ability compared to girls of high ability. If these comparisons show that students in the two groups who are of the same ability perform differently on a task, this may indicate that the task is biased. However, although test items may function differently in two groups, item differential functioning does not prove bias.

DIFFERENTIAL VALIDITY: The term used when the scores of a test have significantly different degrees of validity for different groups of people (e.g., males vs. females, Black persons vs. White persons).

EMPIRICAL INFORMATION: Refers to the research data used to support the development of the test, the selection of items, claims to reliability, and claims to validity.

INTERNAL CONSISTENCY RELIABILITY: A procedure for studying reliability when the focus of the investigation is on the consistency of scores on the same occasion and on similar content, but when conducting repeated testing or alternate forms testing is not possible. The procedure uses information about how consistent the examinees' scores are from one item (or one part of the test) to the next to estimate the consistency of examinees' scores on the entire test.

ITEMS: The questions, exercises, and activities appearing on an assessment procedure. Typically, the term "item" is used to refer to paper-and-pencil test exercises.

MULTIPLE-CHOICE: This format consists of a stem that poses a question or sets a problem and a set of two or more response choices for answering the question or solving the problem. Only one of the response choices is the correct or best answer.

NORM SAMPLE: A well-defined group of examinees who have been given the same assessment and under the same conditions (same time limits, same directions, same equipment and materials, etc.). The local norm group would include examinees in the local agency or school district only; the national norm group would include examinees in the publisher's sample chosen to represent all examinees in the nation of the type that would be administered the test.

NORMATIVE DATA: The statistical data resulting from the application of the test to a national representative sample of examinees.

NORM-REFERENCING: A framework for interpreting an examinee's score by comparing the examinee's test performance with the performance of other examinees in a well-defined group who took the same test. Norm-referencing answers the question, "How did this examinee do compared to other examinees?"

PERCENTILE RANKS: A norm-referenced score that tells the percentage of persons in a norm group scoring lower than a particular raw score.

PILOT TESTING: The process of trying out the test or its items with a small sample of examinees to obtain preliminary information about how well the test is functioning. The results of a piloting exercise are used to make adjustments that improve the test or its items.

PROFILE OF DIFFERENT DIMENSIONS: An assessment approach that identifies an examinee's pattern of scores in several different but related areas. The approach is useful when you know little about an examinee and want to get a rough idea of the examinee's needs. The rough profile of the examinees provides only general guidance and must be followed up with more detailed diagnosis.

RELIABILITY: The consistency of examinees' assessment results (scores) when the test procedure is repeated for a group of examinees. Reliability refers to the consistency of scores rather than to the assessment instrument itself. A reliability coefficient is any one of several statistical indices that quantify the amount of consistency in assessment scores.

STANDARDIZATION: A process in which the procedures, administration, materials, and scoring rules are fixed so that as far as possible the assessment is the same at different times and places. Sometimes the term is restricted to mean the process by which the final version of a test is administered to a nationally representative sample of examinees for purposes of developing various types of norm-referenced scores.

TECHNICAL MANUAL: A publication prepared by a test developer that explains the technical details of how the test was developed, how the norms were created, the procedures for selecting test items, the procedures for equating different forms of the test, and the reliability and validity studies that have been completed for the test.

TEST-RETEST RELIABILITY: A procedure for estimating reliability when the focus of the study is the consistency of the examinees' scores from one occasion to the next and when they are administered the same items on both occasions. The purpose of such studies is to identify the amount of error in the scores than can be attributed to the occasions on which the examinees took the test when the items' content stayed the same.

TRUE-FALSE: This type of item format consists of a statement or proposition that the examinee must judge as true or false.

VALIDITY OR VALIDITY STUDIES: Evidence and theory that support the soundness of one's interpretations and uses of examinees' assessment results. To validate interpretations and uses of examinees' assessment results, one must combine evidence from a variety of sources that collectively demonstrate that the intended interpretations and uses of the test scores are appropriate and defensible. Evidence must also demonstrate that examinees would experience no serious negative consequences when their test scores are used as intended.