Assessment Glossary

Achievement Tests	Summative assessments that measure achievement of the academic content standards. They generally cover broad ranges of content knowledge, skills and processes. Many times they are considered status assessments. The definition is available in OAC Rule 3301-13-01(A)(1) at http://onlinedocs.andersonpublishing.com/oh/lpExt.dll?f=templates&eMail=Y&fn=main-h.htm&cp=PORC/269d5/26f3a/26f68 which states, " 'Achievement tests' means tests, aligned to academic content standards, designed to measure the skill in a specific content that is expected at the end of the designated grade."
Bias	Construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees. A test item is unbiased if all individuals having the same underlying ability have equal probability of getting the item correct, regardless of subgroup membership.
Biserial	Relationship between student performance on an item and the student's performance on the whole test.
Calibration	The process of determining the parameters for an item.
Conditional Standard Error of Measurement	The error of measurement that affects the scores of examinees at a specified test score level.
Consequential-related Evidence of Validity	Evaluation of the intended or unintended social consequences of test interpretation and use. The appropriateness of the intended testing purpose and the possible occurrence of unintended outcomes and side effects are the major issues.
Construct	An individual characteristic that is assumed to exist in order to explain some aspect of behavior. Constructs are theoretical constructions that are used to explain performance on an assessment.
Construct-related Evidence of Validity	Evidence that supports a proposed construct interpretation of scores on a test.
Constructed Response Items	Questions that require an examinee to plan his/her own answer and to express the answer in his/her own words.
Content Knowledge	Set of knowledge, skills and/or abilities.
Content-related Evidence of Validity	Evidence that shows the extent to which the content of a test is appropriate relative to its intended purpose.
Convergent Evidence of Validity	Relationships between test scores and other measures intended to assess similar constructs.
Criterion (Domain) Referenced Test	A test for which the test results measure an examinee's performance against a delineated set of knowledge, skills and/or abilities. An absolute score.
Decision Consistency	The percentage of times that the same decision is made when a specified decisions rule is used across alternate testings.
Diagnostic Tests	Assessments that give detailed information about specific areas of academic strengths and weaknesses. They generally cover very narrow ranges of content or knowledge. Results from the tests are used to shape or change instruction and are helpful for educational intervention programs. The definition is available in OAC Rule 3301-13-01(A)(5) at http://onlinedocs.andersonpublishing.com/oh/lpExt.dll?f=templates&eMail=Y&fn=main-h.htm&cp=PORC/269d5/26f3a/26f68 which states " 'Diagnostic assessments' means the tests designed to measure student comprehension of academic content standards and mastery of related skills for the relevant subject area at each grade . . . ."

Differential Item Functioning (DIF)	A statistical property of a test item in which different groups of test takers who have the same total test score have different average item scores or, in some cases, different rates of choosing various item options.
Divergent/discriminate Evidence of Validity	Relationship between test scores and measures purportedly of different constructs.
Error of Measurement	The difference between an observed score and the corresponding true score.
Equating	The process of putting two or more essentially parallel tests on a common scale.
Fairness	The principle that every test taker should be assessed in an equitable way.
Field Test	Test in which the items (questions) are tested with an appropriate group of examinees and item parameters are established.
Formative Assessment	A test whose results are used to modify instruction.
Generalizability Theory	An analysis used to evaluate the ability to generalize score interpretations beyond the specific sample of items, persons, and observational conditions.
Item Difficulty	Perceived or empirically based notion of how hard an item (question) is. Generally based on the percentage of students who get it correct (empirical) but can be based on expert opinion (perceived).
Item Discrimination	An item's ability to distinguish between different levels of ability or achievement.
Item Parameters	Statistical characteristics of items (questions) represented as numbers. Parameters include item discrimination, theta values, p-values, etc.
Item Response Theory (IRT)	A mathematical model of the relationship between performance on a test item (question) and the test taker's level of performance on a scale of ability, trait or proficiency being measured, usually denoted as theta.
Norm Referenced Test	A test for which the test results indicate an examinee's position relative to some group. A relative score.
PEG	Project Essay Grade Automated Essay Scoring. It is a computer scoring system that is being used in concert with human scorers in a variety of "higher stakes" standardized testing environments such as the GRE and GMAT. It has been validated by more independent research than all other competing systems combined.
Performance Assessment	Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied.
Predictive-related Evidence of Validity	How accurately test data can predict criterion scores that are obtained at a later time.
Process	A procedure that can be applied in many settings.
Proficiency Tests	Measure proficiencies on learning outcomes. The definition is available in OAC Rule 3301-13-01(A)(15) at http://onlinedocs.andersonpublishing.com/oh/lpExt.dll?f=templates&eMail=Y&fn=main-h.htm&cp=PORC/269d5/26f3a/26f68 which states " 'Proficiency test' means all fourth-, sixth-, and ninth-, grade proficiency tests, designed to measure the skill expected at the end of the designated grade."
P-values	The percentage of students getting the item correct.
Rasch Model	A one parameter IRT model that is used by Ohio. The model is named after Georg Rasch, a Danish mathematician.
Scale Score	A score to which raw scores are converted by numerical transformations.
Scaling	The process of creating a scale or a scale score. Scaling may enhance test score interpretation by placing scores from different test forms onto a common scale or by producing scale scores designed to support criterion-referenced score interpretations.

Sensitivity	To not promote nor inquire as to individual moral or social values of beliefs; ensure diverse cultures are represented in assessments; assessment materials used are neither offensive to nor stereotypes of any student group.
Standard Error of Measurement	The error of measurement that is associated with the test scores for a specified group of test takers.
Summative Assessment	A test used at the end of instruction to determine effectiveness of instruction. A measure of maximum performance.
Test Accommodations	Changes made in the format and/or administration procedure of a test in order for test takers, who are unable to take the original test under standard test conditions, to access the information and questions in order to respond. An accommodation does not change the construct being measured.
Test Blueprint (Specifications)	A detailed description for the test that specifies the number or proportion of items that assess each content and process/skill area; the format of items, responses, and scoring rubrics and procedures; and the desired psychometric properties of the items and test.
Test Modification	Changes made in the content, format and/or administration procedure of a test in order for test takers, who are unable to take the original test under standard test conditions, to access the information and questions in order to respond. A modification changes the construct being measured.
Theta Values	Measure of the underlying ability. Theta (ability scores) can be transformed into achievement scores.
Thinking Skills	Cognitive processes that range from simple to complex. Often delineated in a taxonomy.
Universal Design	The design of assessments to be usable by all test takers, to the greatest extent possible, without the need for accommodations or modifications.
Validity	The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of the test.

There are multiple sources for these definitions including:

Standards for Educational and Psychological Testing produced by American Educational Research Association, American Psychological Association & National Council on Measurement in Education in 1985.

Standards for Educational and Psychological Testing produced by American Educational Research Association, American Psychological Association & National Council on Measurement in Education in 1999.