

Assessment Literacy and You

Glossary of
Assessment terms

and

Terms related to assessment

Compiled by the Assessment Branch at DoDEA HQ

Assessment Literacy and You

The next few pages will provide information on assessment types, assessment terms, purpose of assessments, and a chart showing all the assessments currently used system-wide by DoDEA.

Data

A collection of organized information, individual facts, statistics, or items of information. This may consist of numbers, words, or images, particularly as measurements or observations of a set of variables.

Assessment

A test or an examination (or "exam") is an assessment, often administered on paper or on the computer, intended to measure the test-takers' or respondents' (often a student) knowledge, skills, aptitudes, or classification in many other topics (e.g., beliefs). Tests are often used in education, professional certification, counseling, psychology (e.g., MMPI), the military, and many other fields. The measurement that is the goal of testing is called a test score, and is "a summary of the evidence contained in an examinee's responses to the items of a test that are related to the construct or constructs being measured."^[1]

Test scores are interpreted with regards to a norm or criterion, or occasionally both. The norm may be established independently, or by statistical analysis of a large number of subjects.

A standardized test is one that is administered and scored in a consistent matter to ensure legal defensibility. A large proportion of formal testing is standardized. A standardized test with important consequences for the individual examinee is referred to as a high stakes test.

The basic component of a test is an *item*. These are often colloquially referred to as "questions," but not every item is phrased as a question; it may be such things as a true/false statement or a task that must be performed (if a performance test).

Types of Assessments

Alternative Assessment

Many educators prefer the description "assessment alternatives" to describe alternatives to traditional, standardized, norm- or criterion-referenced traditional paper and pencil testing. An alternative assessment might require students to answer an open-ended question, work out a solution to a problem, perform a demonstration of a skill, or in some way produce work rather than select an answer from choices on a sheet of paper. Portfolios and instructor observation of students are also alternative forms of assessment.

Uses of alternative assessment data	
Appropriate	Not Appropriate
* providing rich feedback to teachers, leading to improved learning outcomes for students * use the results to evaluate their student's performance in their own classrooms and then tailoring instruction in areas of weakness	

Criterion-referenced test

A test in which the results can be used to determine a student's progress toward mastery of a content area. Performance is compared to an expected level of mastery in a content area rather than to other students' scores. Such tests usually include questions based on what the student was taught and are designed to measure the student's mastery of designated objectives of an instructional program. The "criterion" is the standard of performance established as the passing score for the test. Scores have meaning in terms of what the student knows or can do, rather than how the test-taker compares to a reference or norm group.

* A common misunderstanding regarding the term is the meaning of *criterion*. Many, if not most, criterion-referenced tests involve a cutscore, where the examinee passes if their score exceeds the cutscore and fails if it does not (often called a mastery test). The *criterion* is not the cutscore; the criterion is the domain of subject matter that the test is designed to assess. For example, the criterion may be "Students should be able to correctly add two single-digit numbers," and the cutscore may be that students should correctly answer a minimum of 80% of the questions to pass.

Uses of criterion-referenced data	
Appropriate	Not Appropriate
<ul style="list-style-type: none"> * used to assess mastery * used to rank students or schools based on their scores *To find out how much students know before instruction begins and after it has finished. *To determine whether each student has achieved specific skills or concepts. *Each individual is compared with a preset standard for acceptable achievement. The performance of other examinees is irrelevant. 	

Diagnostic assessment

provide instructors with information about student's prior knowledge and misconceptions before beginning a learning activity.

Uses of diagnostic data	
Appropriate	Not Appropriate
<ul style="list-style-type: none"> * used to assess students strengths and areas of concern * used to inform instruction * can be used for summative data 	

Norm-referenced test

A test in which a student or a group's performance is compared to that of a norm group. The student or group scores will not fall evenly on either side of the median established by the original test takers. The results are relative to the performance of an external group and are designed to be compared with the norm group providing a performance standard. Often used to measure, compare, and sort students, schools, districts, and states on the basis of norm-established scales of achievement.

Uses of norm-referenced data	
Appropriate	Not Appropriate
<ul style="list-style-type: none"> *make comparisons across large numbers of students *make decisions regarding student placement and advancement *designed to compare students (i.e., disperse average student scores along a bell curve, with some students performing very well, most performing average, and a few performing poorly). * determine individual performance in comparison to others 	<ul style="list-style-type: none"> * use to determine the effectiveness of educational programs * to provide diagnostic information for individual students *Percentiles and standard scores, for example, are not intended to measure student growth. They only show relative standing compared to others, and this relationship does not necessarily change as students progress through school

Performance-Based Assessment

Direct, systematic observation and rating of student performance of an educational objective, often an ongoing observation over a period of time, and typically involving the creation of products. The assessment may be a continuing interaction between teacher and student and should ideally be part of the learning process. The assessment should be a real-world performance with relevance to the student and learning community. Assessment of the performance is done using a rubric or analytic scoring guide to aid in objectivity. Establishment of a clear purpose for the data collected during the assessment before the assessment is undertaken. Performance-based assessment is a test of the ability to apply knowledge in a real-life setting.

Uses of performance-based data	
Appropriate	Not Appropriate
* providing rich feedback to teachers, leading to improved learning outcomes for students * use the results to evaluate their student's performance in their own classrooms and then tailoring instruction in areas of weakness	for ranking or sorting students

Formative and Summative Assessment

Formative assessment is a process which allows one to determine the degree to which students know or are able to do a given learning task, and which identifies the part of the task that the student does not know or is unable to do. Outcomes suggest future steps for teaching and learning.

"Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes."¹

Summative assessments are an evaluation at the conclusion of a unit or units of instruction or an activity or plan to determine or judge student skills and knowledge or effectiveness of a plan or activity. Outcomes are the culmination of a teaching/learning process for a unit, subject, or year's study.

Informal and formal assessments

Assessment can be either formal or informal.

Formal assessment usually implicates a written document, such as a test, quiz, or paper. A formal assessment is given a numerical score or grade based on student performance, whereas an informal assessment does not contribute to a student's final grade.

An **informal assessment** usually occurs in a more casual manner and may include observation, inventories, checklists, rating scales, rubrics, performance and portfolio assessments, participation, peer and self evaluation, and discussion.

Average

A general term applied to the various measures of central tendency. The three most widely used averages are the arithmetic mean (mean), the median, and the mode. When the term "average" is used without designation as to type, the most likely assumption is that it is the arithmetic mean.

Cut Score

A specified point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point.

Cohort

A group whose progress is followed by means of measurements at different points in time.

Distribution (Frequency Distribution)

A tabulation of the scores (or other attributes) of a group of individuals to show the number (frequency) of each score, or of those within the range of each interval.

Grade Equivalent

A score that describes student performance in terms of the statistical performance of an average student at a given grade level. A grade equivalent score of 5.5, for example, might indicate that the student's score is what could be expected of an average student doing average work in the fifth month of the fifth grade. This score allows for a theoretical or approximate comparison across grades. It ranges from September of the kindergarten year (K. 0.) to June of the senior year in high school (12.9) Useful as a ranking score, grade equivalents are only a theoretical or approximate comparison across grades.

Item Analysis

Analyzing each item on a test to determine the proportions of students selecting each answer. Can be used to evaluate student strengths and weaknesses; may point to problems with the test's validity and to possible bias.

Mean

One of several ways of representing a group with a single, typical score. It is figured by adding up all the individual scores in a group and dividing them by the number of people in the group. Can be affected by extremely low or high scores.

Median

The point on a scale that divides a group into two equal subgroups. Another way to represent a group's scores with a single, typical score. The median is not affected by low or high scores as is the mean. (See Norm.)

Mode

The score or value that occurs most frequently in a distribution.

Norm

A distribution of scores obtained from a norm group. The norm is the midpoint (or median) of scores or performance of the students in that group. Fifty percent will score above and fifty percent below the norm.

Normal Curve Equivalent

A score that ranges from 1-99, often used by testers to manipulate data arithmetically. Used to compare different tests for the same student or group of students and between different students on the same test. An NCE is a normalized test score with a mean of 50 and a standard deviation of 21.06. NCEs should be used instead of percentiles for comparative purposes.

Percentile

A ranking scale ranging from a low of 1 to a high of 99 with 50 as the median score. A percentile rank indicates the percentage of a reference or norm group obtaining scores equal to or less than the test-taker's score. A percentile score does not refer to the percentage of questions answered correctly, it indicates the test-taker's standing relative to the norm group standard.

Percentile Rank

Most commonly, the percentage of scores in a specified distribution that fall below the point at which a given score lies. Sometimes the percentage is defined to include scores that fall at the point; sometimes the percentage is defined to include half of the scores at the point. Percentile ranks indicate the status or relative standing of a pupil in comparison to other pupils. The percentile rank tells the percent of pupils in a particular norm group who obtain lower scores; thus, for example, if Ann earns a percentile rank of 70 on a particular test it means she scored better than 70 percent of the pupils in the norm group and 30 percent scored as well or better than she. The scale goes from 1 to 99 percent. If three points are used to divide the scale into four equal quarters the points are called quartiles; quartile one, quartile two, and quartile three.

Quartile

Quartiles are points, not areas. A score does not fall in a quartile. A score can be above, at, or below a quartile. A score can be within two quartiles. There is not fourth quartile. Many people have the misconceptions that since there are four quarters that there should also be four quartiles, but this is not the case. Quartiles are points, not areas, so there are four areas divided by the three quartiles, but there are not four quartiles.

Reliability

The measure of consistency for an assessment instrument. The instrument should yield similar results over time with similar populations in similar circumstances.

Sampling

A way to obtain information about a large group by examining a smaller, randomly chosen selection (the sample) of group members. If the sampling is conducted correctly, the results will be representative of the group as a whole. Sampling may also refer to the choice of smaller tasks or processes that will be valid for making inferences about the student's performance in a larger domain. "Matrix sampling" asks different groups to take small segments of a test; the results will reflect the ability of the larger group on a complete range of tasks.

Scale Scores

Scores based on a scale ranging from 001 to 999. Scale scores are useful in comparing performance in one subject area across classes, schools, districts, and other large populations, especially in monitoring change over time. A score to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores); units of a single, equal-interval scale that are applied across levels of a test.

Standard Deviation

A measure of the variability or dispersion of a distribution of scores. The most widely used measure of dispersion of a frequency distribution. It is equal to the positive square root of the population variance. The more the scores cluster around the mean, the smaller the standard deviation. For a normal distribution, approximately two thirds (68.3 percent) of the scores are within the range from one S.D. below the mean to one S.D. above the mean. Computation of the S.D. is based upon the square of the deviation of each score from the mean. The S.D. is sometimes called "sigma" and is represented by the symbol (σ).

Standard Error of Measurement

The standard deviation of an individual's observed scores from repeated administrations of a test (or parallel forms of a test) under identical conditions. Because such data cannot generally be collected, the standard error of measurement is usually estimated from group data.

Stanine Scores

Stanine scores are normalized standard scores with a range of 1 to 9, a mean of five, and a standard deviation of two. Like percentile ranks they are status scores within a particular norm group. The first stanine is the lowest scoring group and the 9th stanine is the highest scoring group. Advocates of stanine reporting cite the fact that the single digit scale is simple and convenient to use and that its use minimizes the apparent importance of small score differences. On the other hand, the stanine scale may be regarded as unnecessarily coarse particularly for relatively reliable tests. For example, all pupils scoring between the 40th and 60th percentiles are assigned a stanine of 5. However, a pupil scoring at the 59th percentile, which is in stanine 5, is probably much more similar in achievement level to a pupil scoring at the 61st percentile, stanine 6, than to one at the 41st, stanine 5. In some instances the width of the stanine band exceeds the standard error of measurement. Another reservation about the use of stanine scores is that there is evidence that skills development in the elementary schools is more variable in subjects such as reading in which the pupils have many opportunities for advancing "on their own" than they are in subjects such as mathematics in which pupil progress is more rigidly controlled through placement of concepts and processes in the curriculum.

T-Score

A derived score on a scale having a mean score of 50 units and a standard deviation of 10 units.

Triangulation

Triangulation is a shorthand term for synthesis and integrated analysis of data from multiple sources for program decision-making. It is a powerful tool that can be used to demonstrate program impact; identify areas for improvement; direct new programs and enhance existing programs; and help direct policy changes.

Process:

- Gather data/reports
- Make observations from individual data sets
- Note trends across data sets and develop hypotheses to answer key questions
- Check hypotheses and identify additional data to be captured
 - Note evidence that supports/refutes
 - Identify and gather more data, as needed
 - Select most promising hypotheses

Validity

The test measures the desired performance and appropriate inferences can be drawn from the results. The assessment accurately reflects the learning it was designed to measure.

z-score (lowercase z)

The most commonly used standard score. It is a measure of relative location in a distribution; it gives, in standard deviation units, the distance from the mean of a particular score. In z-score notation, the mean is 0 and a standard deviation is 1.

DRAFT

Other Terms Associated with Assessments

Adequate yearly progress (AYP)

A provision of the federal No Child Left Behind (NCLB, 2001) legislation requiring schools, districts, and states to demonstrate on the basis of test scores that students are making academic progress based on test scores. Each state was required by NCLB to submit by January 31, 2003 a specific plan for monitoring AYP.

Aggregation

The total or combined performance of all students for reporting purposes.

Assistive technology

A device or service that is used to increase, maintain, or improve the functional capabilities of a student with a disability.

Baseline data

The initial measures of performance against which future measures will be compared.

Constructed Response

An exercise for which examinees must create their own responses or products rather than choose a response from an enumerated set. Also considered subjective questions. Subjective assessment is a form of questioning which may have more than one correct answer (or more than one way of expressing the correct answer). Subjective questions include extended-response questions and essays.

Constructive Alignment

A principle used for devising teaching and learning activities, and assessment tasks, that directly address the learning outcomes intended in a way not typically achieved in traditional lectures, tutorial classes and examinations (Biggs and Tang, 2007). Constructive alignment was devised by Professor John B. Biggs, and represents a marriage between a constructivist understanding of the nature of learning, and an aligned design for outcomes-based teaching education.

Constructive alignment is the underpinning concept behind the current requirements for programme specification, declarations of Intended Learning Outcomes (ILOs) and assessment criteria, and the use of criterion based assessment. There are two basic concepts behind constructive alignment: Learners construct meaning from what they do to learn. This concept derives from cognitive psychology and constructivist theory, and recognizes the importance of linking new material to concepts and experiences in the learner's

memory, and extrapolation to possible future scenarios via the abstraction of basic principles through reflection.

The teacher makes a deliberate alignment between the planned learning activities and the learning outcomes. This is a conscious effort to provide the learner with a clearly specified goal, a well designed learning activity or activities that are appropriate for the task, and well designed assessment criteria for giving feedback to the learner.

Disaggregation

The collection and reporting of student achievement results by particular subgroups to ascertain the subgroup's academic progress. Disaggregation makes it possible to compare subgroups or cohorts.

Lexile

A Lexile is a standard score that matches a student's reading ability with difficulty of text material. A Lexile can be interpreted as the level of book that a student can read with 75% comprehension. Experts have identified 75% comprehension level as offering the reader a certain amount of comfort and yet still offering a challenge.

Rubric

In general a rubric is a scoring guide used in subjective assessments. A rubric implies that a rule defining the criteria of an assessment system is followed in evaluation. A rubric can be an explicit description of performance characteristics corresponding to a point on a rating scale. A scoring rubric makes explicit expected qualities of performance on a rating scale or the definition of a single scoring point on a scale.

Selected-Response (Multiple Choice)

A question or an incomplete sentence or idea. The students are expected to choose the correct or best answer/completion from a menu of alternatives. Also known as objective questioning. Objective assessment is a form of questioning which has a single correct answer. There are various types of objective and subjective questions. Objective question types include true/false answers, multiple choice, multiple-response, and matching questions.

References

The definitions in this glossary were derived from several sources, including:

1998. Understanding Test Scores: A Handout for Parents. Retrieved October 2, 2008 from <http://www.teachersandfamilies.com>
2002. New Horizons for Learning - Assessment Terminology: A Glossary of Useful Terms Retrieved October 2, 2008 from <http://www.newhorizons.org>
2003. Glossary of Assessment Terms and Acronyms. Retrieved October 2, 2008 from <http://www.ccsso.org>
2006. ERIC Digest, "A Glossary of Measurement Terms." Retrieved October 2, 2008 from eric.ed.gov/ERICDocs/data/
2006. Glossary of Assessment Terms. Retrieved October 2, 2008 from <http://ags.pearsonassessments.com>
2006. Glossary of Assessment Terms. Retrieved October 2, 2008 from <http://www.postdam.edu>
2008. Assessment Terms. Retrieved October 2, 2008 from <http://www.niu.edu>
2008. CTB McGraw-Hill. Glossary of Assessment Terms. Retrieved October 2, 2008 from <http://www.ctb.com>
2008. Glossary of Useful Terms. Retrieved October 2, 2008 from <http://sabes.org/assessment/glossary.htm>
- 1 Popham, W. James. (2008). Transformative Assessment. Association for Supervision and Curriculum Development, 5.

Department of Defense Education Activity

VISION: Communities committed to success for ALL students!

MISSION: To Provide and Exemplary Education that Inspires and Prepares All DoDEA Students for Success in a Dynamic, Global Environment.

DoDEA External (System-wide) Assessments For School Year 2008 - 2009

Assessment	Pre-K	K	1	2	3	4	5	6	7	8	9	10	11	12
Norm-Referenced (TerraNova)					X	X	X	X	X	X	X	X	X	
Reading Proficiency					DRA X			SRI X			SRI X			
U.S. History EOC											All Enrolled Students			
***Foreign Language Proficiency Test Levels II & IV											All Enrolled Students			
PSAT Preliminary SAT/MNSQT												X	X	
NAEP (Odd years only) National Assessment of Educational Progress						X				X				
ACT (College entrance)												Voluntary		
SAT (College entrance)												Voluntary		

****Language Groups: Chinese, French, Italian, German, Japanese, and Spanish (only)

Department of Defense Education Activity