GLOSSARY OF ASSESSMENT TERMS

STATE COLLABORATIVE ON ASSESSMENT AND STUDENT STANDARDS
ARTS EDUCATION ASSESSMENT CONSORTIUM (SCASS/ARTS)
COUNCIL OF CHIEF STATE SCHOOL OFFICERS

SCASS/ARTS WEBSITE ITEM POOL PROJECT

This glossary was developed by the members of SCASS/Arts Education Assessment Consortium. The terms below are words/phrases in general use in the field of assessment.

Accommodations

approved/standardized administrative or scoring adjustments (e.g., large print or Braille test booklets, individual or small group administrations, reading the test to the student) made for special populations taking standardized assessments

Accountability testing

Using student achievement tests to measure the effectiveness of an educational program. Usually summative in nature and in the form of state or other large-scale test designed to conform to psychometric standards, an accountability test purports to assign responsibility for the success or failure of an educational program or system by demanding that schools demonstrate the impact and effectiveness of educational programs in order to justify the money invested in education. Accountability testing is designed to provide achievement data that is used to evaluate and presumably improve the system.

Achievement test

a test designed to measure students' "school taught" learning, as opposed to their initial aptitude or intelligence.

Alternative assessment

assessments other than traditional multiple-choice tests; most often used to describe performance assessments or other assessments that provide more feedback about student learning than whether the answer is correct or incorrect. (Also see Accommodations.)

Analytic scoring

A method of scoring performance assessments that yields multiple scores for the same task/performance. Performance is separated into major components, traits, or dimensions and each is independently scored. (e.g., a particular sample of a student's writing may be assessed as grammatically correct at the same time it is assessed as poorly organized.) Analytic scoring is especially effective as a diagnostic tool.

Anchor

(also called exemplars or benchmarks); a sample of student work (product or performance) used to illustrate each level of a scoring rubric; critical for training scorers of performances since it serves as a standard against which other student work is compared.

Aptitude test

a test which uses past learning and ability to predict what a person can do in the future; aptitude tests depend heavily on out-of-school experiences rather than in-school learning (Also see intelligence test.)

Assessment

The ***process*** of collecting and analyzing data for the purpose of evaluation. The assessment of student learning involves describing, collecting, recording, scoring, and interpreting information about performance. A complete assessment of student learning should include measures with a variety of formats as developmentally appropriate. Assessments and the tests they use are usually classified by how the data are used; either formative, benchmark or interim, and summative.

Authentic assessments

assessments that emulate the performance that would be required of the student in real-life situations.

Benchmarks

identifiable points on a continuum toward a goal or standard. The term may be used to describe content standards when interim targets (benchmarks) have been set by age, grade, or developmental level; the term is also used interchangeably with "anchor" papers or performances which illustrate points of progress on an assessment scale (i.e., student works which exemplify the different levels of a scoring rubric).

CIA

acronym for curriculum, instruction, and assessment

Cohort

a group of students whose progress is followed and measured at different points in time.

Competency test

a test intended to verify that a student has met standards (usually minimal) of skills and knowledge and therefore should be promoted, graduated, or perhaps deemed competent.

Constructed-response assessment

a form of assessment that calls for the student to generate the entire response to a question, rather than choosing an answer from a list (e.g., paper-and-pencil responses on essay or short answer tests or performances which may be drawn, danced, acted out, performed musically, or provided in any other way to exhibit particular skills or knowledge. (Also referred to as open-response and open-ended assessments.)

Context

the surrounding circumstances or environment in which an assessment takes place (e.g., embedded in the instruction or under standardized conditions [e.g., part of a large scale assessment])

Cornerstone assessment tasks

are curriculum-embedded assessment tasks that are intended to engage students in applying their knowledge and skills in an authentic context. These tasks are described by their originator Jay McTighe as:

- *curriculum embedded* (as opposed to externally imposed);
- *recurring across the grades*, becoming increasingly sophisticated over time;
- establishing *authentic contexts* for performance;
- calling for *understanding* and *transfer* via genuine performance;
- used as rich learning activities *or* assessments;
- *integrating 21st century skills* (e.g., critical thinking, technology use, teamwork) with subject area content;
- evaluating performance with established *rubrics*;
- engaging students in *meaningful learning* while encouraging the best teaching;
- providing content for student portfolios so that students graduate with a *resume* of demonstrated accomplishments rather than simply a transcript of courses taken.

Criteria

(sometimes used as synonym for traits or attributes); the rules or guidelines used for categorizing or judging; in arts assessment, the rules or guidelines used to judge the quality of a student's performance. (Also see rubric, scoring guide, and scoring criteria.)

Criterion-referenced assessment

an assessment designed to measure performance against a set of clearly defined criteria. Such assessments are used to identify student strengths and weaknesses with regard to specified knowledge and skills (which are the goals or standards of the instruction). Synonyms include: standard-based or -referenced, objective-referenced, content-referenced, domain-referenced, or universe-referenced.

Curricular alignment

the degree to which a curriculum's scope, sequence, and content match standards, instruction, assessment, or instructional resources.

Cut score

(also called performance standard) performance level or numerical score established by the assessment system to describe how well the student performed. The cut score can be manipulated to increase or decrease the number "passing" or "failing" a test. (Also see standard-setting.)

Descriptors

explanations that define the levels of scoring scales (Also see criteria.)

Dimension

specific traits, characteristics, or aspects of performance which are fairly independent of each other and can be scored separately (e.g., rhythm and melody can be scored separately for the same musical performance).

Disaggregate

(as in disaggregated data); pulling information apart (e.g., looking at the performance of various sub-groups instead of only the performance of the large group).

Educational outcome

an educational goal, expectation, or result that occurs at the end of an educational program or event (usually a culminating activity, product, or other measurable performance).

Enhanced/extended multiple-choice assessments

selected-response assessments with additional parts (for more points); this additional part often requires the students to justify their answers, show their work, or explain why they marked a particular option.

Essay test

a paper-and-pencil test that requires students to construct their entire brief or extensive responses to the question(s); should be limited to measuring higher levels of learning.

Extended-response assessments

an essay question or performance assessment, which requires an elaborated or graphic response that expresses ideas and their interrelationships in a literate and organized manner

Evaluation

a judgment about the worth or quality of something. In education, data from tests, tasks, or performances are used to make judgments about the success of the student or program.

Formative Assessment

(Sometimes referred to as Assessment for Learning) 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. Short-interval and usually classroom-based assessments that have immediate information for teachers and students to inform the instructional process and determine what comes next in the learning process.

Generalizability

the degree to which the performances measured by a set of assessment items/tasks are representative of the entire domain being assessed (E.g., is one performance assessment sufficient for drawing conclusions about a student's ability to critique works of art?); may also be an issue in drawing a sample of students from a population (i.e. the degree to which a sample of students is representative of the population from which it is drawn)

Grade equivalent

a score, available from some standardized tests, which describes the performance of students according to how it resembles the performance of students in various grades. A GE of 5.5 means that the student is performing like a student in the fifth month of the fifth grade.

Grading

a rating system for evaluating student work; grades are usually letters or numbers and their meaning varies widely across teachers, subjects, and systems.

High-stakes testing

any testing program for which the results have highly significant consequences for students, teachers, schools, and/or districts. These summative tests are frequently used as accountability devices to determine effectiveness or success.

Holistic method

a scoring method which assigns a single score based on an overall appraisal or impression of performance rather than analyzing the various dimensions separately. A holistic scoring rubric can be specifically linked to focused (written) or implied (general impression) criteria. Some forms of holistic assessment do not use written criteria at all but rely solely on anchor papers for training and scoring.

Intelligence tests

tests designed to measure general cognitive functioning; group or individually administered tests used to determine mental age as compared to chronological age (MA/CA x 100 = IQ [intelligence quotient]); i.e., the "average" IQ of the population is 100. Some intelligence tests do not calculate mental age but compare an individual's performance to the performance of a norm group at various developmental levels, generating verbal and performance scores with a mean or "average" score of 100.

Item analysis

a statistical analysis of the items on a selected-response test to determine the relationship of the item to the test's validity and reliability as a whole. The number and nature of the students selecting each option are analyzed.

Matrix sampling

a process used to estimate the performance of large groups through testing a representative sample of the students. Each student in the sample may be given only a small segment of the total assessment.

Mean

the arithmetic average of a group of scores; one of three measures of central tendency, a way to describe a group of scores with a single number.

Median

a measure of central tendency, which identifies the point on the scale that separates a group of scores so that there is an equal number of scores above and below it.

Metacognition

the ability to think about one's own thinking; the knowledge that individuals have of their own thinking processes and strategies and their ability to monitor and regulate those processes.

Multiple-choice test

a test consisting of items (questions or incomplete statements) followed by a list of choices from which students have to select the correct or best response.

Multiple measures

the use of a variety of assessments to evaluate performance in a subject area (e.g., using multiple-choice items, short answer questions, and performance tasks to assess student achievement in a subject); the use of multiple measures is advocated to obtain a fair and comprehensive measurement of performance

Mode

a measure of central tendency which identifies the most frequent score in a group of scores (e.g., in the group of scores: 1, 2, 8, 9,9,10, the mode is 9).

Norm

the midpoint or "average" score for the group of students to which a norm-referenced test was initially administered (the norm group). By design, 50% of the students score below and 50% above this score.

Norm group

a group of students that is first administered a standardized norm-referenced test by its developers in order to establish scores for interpreting the performance of future test-takers.

Norm-referenced test

a standardized test which compares the performance of students to an original group that took the test (the norm group); results usually reported in terms of percentile scores (e.g., a score of 90 means that the student did better than 90% of the norm group).

Normal curve equivalent (NCE)

a normalized standard score used to compare scores across tests with different scales and/or between students on the same test (since arithmetic manipulations should not use percentiles); it has a mean of 50, a standard deviation of 21.06 and is often required for reporting by federal funding agencies such as Title I.

Open-ended assessments

constructed assessments (frequently tasks or problems) that require students to generate a solution to a problem for which there is no single correct answer (e.g., create a drawing that uses symbols of the Renaissance)

Open-response assessments

constructed-response assessments (ones for which students must construct the entire answer and show their work) that have a single correct answer but multiple methods of solution possible.

Percentile

a statistic provided by standardized norm-referenced tests which describes the performance of a student as compared to that of the norm group. The range is 1to 99 with 50 denoting average performance. A student scoring at the 65th percentile performed better than, or as well as, 65% of the norm group.

Performance assessment

a task/event/performance designed to measure a student's ability to directly demonstrate particular knowledge and skills. E.g., a student may be asked to demonstrate some physical or artistic achievement: play a musical instrument, create or critique a work of art, or improvise a dance or a scene. These kinds of assessments (e.g., tasks, projects, portfolios, etc.) are scored using **rubrics**: established criteria for acceptable performance.

Portfolio

a purposeful collection of student work across time which exhibits a student's efforts, progress, or level of proficiency. Examples of types of portfolios include: showcase (best work), instructional, assessment (used to evaluate the student, and process or project (shows all phases in the development of a product or performance).

Primary trait scoring

A type of rubric scoring constructed to assess a specific trait, skill or format or the impact on a designated audience. (Also see analytic scoring.)

Project

a type of performance assessment which is complex, usually requiring more than one type of activity, process, or product for completion.

Quartile

a way of describing the position of a score on a norm-referenced test, e.g., the score falls in one of four groups: 0-25th percentile, 26-50th percentile, etc.

Quintile

a way of describing the position of a score on a norm-referenced test, e.g., the score falls in one of five groups: 0-20th percentile, 21-40th percentile, etc.

Range

the most rudimentary method of describing how much a group of scores vary; range is determined by subtracting the lowest from the highest score in the group

Rating scale

a scale used to evaluate student learning using a gradation of numbers or labels; a Likert rating scale is frequently used to measure attitudes or perceptions

Reliability

a measure of the consistency of an assessment across time, judges and subparts of the assessment (assuming no real change in what is being measured).

Rating scale

a scale used to evaluate student learning using numbers or labels (e.g, a Likert scale).

Rubric

(sometime referred to as a scoring guide or scoring criteria) an established, ordered set of criteria for judging student performance/products; it includes performance descriptors of student work at various levels of achievement.

Sampling

a way to get information about a large group by examining a smaller representative number of the group (the sample).

Scale score

a score indicating an individual's performance on a standardized test, which allows comparisons across sub-groups and time. (E.g., one could use scale scores to compare test results among classes, schools, and districts; or across grades from year to year.)

Scaffolded assessments

a set of context-dependent assessments, which are sequenced to measure ascending levels of learning; this set usually contains a variety of item formats (from multiple-choice to performance tasks) about a single stimulus (e.g., a specific set of materials: a particular situation, scenario, problem, or event). Since these kinds of assessments can measure a variety of kinds of learning, they provide the opportunity for diagnosis of instruction and identification of student strengths and weaknesses.

Scoring criteria

the rules or guidelines used to assign a score (a number or a label) indicating the quality of a performance; in the analytic scoring of a performance, different rules may be applied to different dimensions or traits of the performance.

Scoring guide

directions for scoring and/or interpreting scores; the guide may include general instructions for raters, training notes, rating scales, rubric, and student work.

Selected-response items

a kind of test item for which students have to select the best or correct answer from a list of options (multiple-choice, etc.) or indicate the truth or falsity of a statement.

Self-assessment

collecting data about one's own performance for the purpose of evaluating it. Self-evaluation may include the comparison of one's own performance against established criteria, change in performance over time, and/or a description of current performance. Three types of educational standards are frequently used in education today:

Standard deviation

a measure of the variability of a group of scores. When the standard deviation is high, students are performing very differently from each other; if it is low, students are performing similarly to one another.

Standard error of measurement

a statistic used to indicate the consistency and reliability of a measurement instrument; a large standard error of measurement indicates that we have less confidence in the obtained score

Standards-based instruction

instruction designed, taught, and assessed using Standards

1. ***Content standards*** specify what students should know and be able to do in a specific content area---the essential knowledge, skills, processes, and procedures students must learn and be able to demonstrate. They answer the question: "What should be learned in this subject?" Student standards have been developed for periods of time ranging from individual grade levels to lifelong learning.
2. ***Performance standards*** specify the degree or quality of learning students are expected to demonstrate in the subject. They answer the question: "How good is good enough?" The national standards for the arts use the term "achievement standards" to avoid confusion between arts performance and performance assessment. (Some states refer to established levels of proficiency instead of performance standards.)
3. ***Opportunity-to-learn standards*** specify what schools must provide to enable students to meet content and performance standards.

student standards (achievement targets)

Stanine

A standard 9-point scale used to report the results of norm-referenced tests in order to allow comparison of scores across students, schools, districts, tests, grades, etc. The mean is 5 and the standard deviation approximately 2. Stanines of 1-3 are considered below average; 4-6 average; and 7-9 above average.

Standardized test

A test administered to a group of persons under the same specific conditions so student results can be fairly compared.

Summative Assessment

The effort to summarize student learning at a particular point in time such as the end of a chapter, unit, grading period, semester, year, or end of course.

Test

A sample of behavior or performance administered in order to provide a basis for inferences about a larger subject area or domain of study. E.g., a teacher may administer a 30-minute test to provide evidence of the student's learning for the last two weeks or for a particular unit of instruction. The test may be norm- or criterion-referenced, traditional (e.g., multiple-choice, short answer, essay, etc.), or performance-based. A *teacher-made test* is one prepared and administered by the teacher, usually for use in the classroom.

Validity

A characteristic of a measure which refers to its ability to measure what it is intended to measure AND do so **reliably** (i.e., measures consistently across time, judges, and sub-parts). A valid measure is both accurate and consistent; e.g., a bathroom scale may record 100 pounds every time a person gets on it, but if he or she actually weighs 120, the scale is ***reliable*** but not ***valid***. Types of validity include:

Content validity—The assessment has content validity if it measures the content or area it intends to measure.

Concurrent validity—The assessment has concurrent validity if it is correlated with other measures of that particular content or area.

Predictive validity—The assessment has predictive validity if it predicts later actual performance of the individual in that subject or area. Predictive validity is related to generalizability.