Glossary

accessibility The extent to which students are easily able to display their optimal performance regarding whatever is being assessed via a test, for instance, a student's possession of a cognitive skill or mastery of a body of knowledge.

accommodation Adjustments made to a test or the test-taking process that increase a test-taker's accessibility while not fundamentally changing the nature of what is being tested. Students' accommodated scores on a test should be so comparable to unaccommodated scores that they can be aggregated together.

accountability tests Those assessment devices, often administered annually under a state's auspices, that are employed to help monitor the effectiveness of local educational programs.

achievement tests Educational tests, often standardized, whose chief function is to measure a student's mastery of the knowledge and intellectual skills taught during school. "Achievement" tests are often contrasted with "aptitude" tests that focus more directly on estimating a test-taker's future performance.

affect/affective A descriptor applied to noncognitive outcomes sometimes promoted in our schools—such as students' attitudes, interests, or values.

alternate-form reliability evidence Reliability evidence indicative of the degree to which two alleged versions of the same test are functioning similarly.

aptitude tests An assessment instrument intended to measure a student's performance so that the student's current score will provide an accurate estimate of that student's performance on a future, frequently academic, criterion. In contrast to "achievement" tests that measure a student's current knowledge and skills, "aptitude" tests attempt to predict a student's future performance.

assessment bias This label indicates whether an educational test offends or unfairly penalizes a test-taker because of such personal

characteristics as the test-taker's race, gender, ethnicity, or level of familial affluence.

bias review The judgmental scrutiny of a set of test items, often carried out on under development, not yet operational items, in an attempt to identify any items that might be biased against particular groups of test-takers based on such personal characteristics as their race, gender, or socioeconomic status. Bias reviews are typically carried out by committees of educators or other adults drawn from the test-taker groups apt to be affected adversely by assessment bias.

building blocks These are the subskills or bodies of enabling knowledge it is assumed students must master as part of a learning progression aimed at getting students to master a target curricular aim sought during a sometimes extended instructional sequence. Learning progressions and the building blocks that constitute them are encountered most frequently during implementations of the formative-assessment process.

cognitive An adjective applicable to students' intellectual activities like becoming skilled in writing a persuasive essay. In education, cognitive variables are often contrasted with affective variables and psychomotor variables.

computer-adaptive testing A sophisticated form of computer-abetted educational testing in which a student's performance in responding to earlier test items determined the difficulty of the items presented subsequently. Because such adjustments in the difficulty of items can better match item demands with a test-taker's ability, computer-adaptive testing typically saves substantial testing time.

computer-based testing When educational tests are delivered by computers, and when students' responses are recorded by computers, this procedure is referred to as computer-based testing. Unlike computer-adaptive testing in which the items presented to an individual student hinge on the student's success with earlier items, computer-based testing refers only to test-delivery and test-takers' responses.

construct-irrelevant variance Variations in students' scores attributable to extraneous factors that distort the meaning of such scores and, as a consequence, diminish the validity of score-based inferences.

content standard A synonymous term that, along with "goal," "objective," or "learning outcome," describe the curricular aim sought for students as a consequence of instruction. When initially introduced into the educational lexicon, "content" standards were contrasted with "achievement" standards—with content standards more descriptive of curricular targets, and achievement standards more focused on the desired level of a student's performance.

correlation coefficient A widely used statistical procedure indicating the

strength and direction of a relationship between two variables, for instance, between a group of students' scores on two different tests. Correlation coefficients can range from a +1.0 to a –1.0 with coefficients near zero indicating little or no relationship between the two variables involved in the computation.

cut-scores This is a score, typically a single numerical score such as "number correct," that is used to separate groups of test-takers into two or more qualitatively different categories, such as "basic" and "advanced." The determination of a defensible cut-score on any type of important educational test typically involves reliance on a standard-setting panel that provides recommendations to those individuals officially charged with actually setting cut-scores.

decision-consistency estimates A way of representing a test's reliability, this indicator focuses on the proportion of test-based decisions that are identical. This consistency index is based on the percentages of test-takers about whom the same decisions—both positive and negative—are based on two administrations of the same test, or on two administrations of allegedly equivalent test forms.

depth of knowledge The degree of cognitive demand required of students when responding to a test item or when asked to master a specific curricular aim. Lower-level depth of knowledge (DOK) requirements might demand mere memorization, but higher DOK items or curricular aims might oblige students to employ synthesis or evaluation.

diagnostic test A test whose typical function is to help isolate a given student's strengths or weaknesses for use in subsequent instruction-related decisions by teachers or by students themselves. When a diagnostic test is focused by teachers on groups of students, such as a classful of history students, a teacher is usually seeking guidance with regard to potentially useful adjustments in ongoing instruction.

differential item functioning A bias detection statistical analysis intended to identify test items that, when used with large groups, display differential item functioning (DIF) whereby an item elicits a decisively disparate response from different groups of test-takers.

enabling knowledge A body of knowledge, such as a memorized list of terms or a collection of important guidelines, that is believed necessary for students to enable their mastery of a more significant curricular aim. Enabling knowledge is usually contrasted with cognitive subskills, also regarded as precursive to a student's mastery of a particular curricular aim.

formative assessment A planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing

instructional procedures or by students to adjust their current learning tactics.

goal An intended curricular outcome like a student's development of a cognitive mathematical skill, such as being able to solve specified categories of equation-based problems. Popularly employed by educators during the mid-twentieth century to describe an intended learning outcome for students, a goal was generally thought of as more broad than an "objective"; also a statement of a curricular intention.

grain-size The breadth of the educational entity being considered, ranging from very specific to very broad. Thus, for example, a small number of broad grain-size curricular aims might embrace the intended student learnings for a particular course, but with curricular aims fashioned at small grain-size, a very large number of such aims would be required. Grain-sizes are also of substantial concern to those reporting students' test results.

high-stakes tests Educational tests whose results are used to make important decisions either about the test takers themselves or about the educators who prepared those test takers for such tests.

inference The test-based conclusion drawn about a student's unseen knowledge and/or skills. Such inferences, also described as "interpretations," are then used for making educational decisions.

instructional sensitivity The extent to which a given educational test is capable of distinguishing between well-taught and poorly taught students. Especially important when a test is to be used evaluatively, a test's instructional sensitivity can be determined via judgmental evidence, empirical evidence, or both of these.

instructional target The curricular aim, that is, goal or objective, to be promoted by a sequence of instructional activities.

internal consistency evidence This is a type of reliability evidence focused on the degree to which a test's items appear to be measuring the same construct. Because many of this nation's early educational tests were developed to measure test-takers' mastery of a single construct, for instance, a student's "quantitative competence," considerable effort was devoted to having a test's items consistently measuring the same thing.

interpretation Synonymous with a test-based "inference," this is a widely used way of signifying—based on a test-taker's overt responses—the nature of that test-taker's unseen knowledge and skills.

learning outcome A descriptor for what is being sought of students as a consequence of instruction. This is one of many synonymous labels for what it is that educators wish their students to learn.

learning progression A planned sequence of instruction, often for an

extended period of time, in which a manageable number of (1) bodies of enabling knowledge and/or (2) subskills have been identified and placed in an optimal instructional order for promoting students' mastery of a more ultimate target curricular outcome. Near the close of each of these "building blocks," students are assessed to provide evidence needed for next steps.

Likert inventory A self-report device, developed originally by Rensis Likert, intended to gauge people's affective status. An inventory contains a series of statements regarding an affective construct of concern, and then respondents (often anonymously) indicate their degree of agreement or disagreement with each statement.

National Assessment of Educational Progress (NAEP) is a federally funded national testing program, first administered in 1969. This carefully constructed and systematically administered national test periodically supplies national estimates of students' performances at certain grade levels and in both mainline subjects as well as in a few less frequently assessed subject areas. Administered on a matrix sampling basis whereby different students in a carefully chosen student sample receive different sets of items, NAEP provides no per-student scores, only group-based reports.

p-value This is usually regarded as a test item's level of difficulty, and is calculated simply as the proportion of students who answer an item correctly. Thus, high *p*-values of, say, .92 would be thought to signify an "easy" item. However, the proportion of students answering an item correctly is influenced heavily by the manner in which a test item's content has been taught. Thus, without knowing about the instructional history linked to an item, it is imprudent to use *p*-values all by themselves to reflect item difficulties.

percentile A test-taker's percentile indicates the percent of test-takers outperformed by a referent group, such as the students in a national norm sample. To illustrate, if all of a large school district's students completed a new test in mathematics, and a particular fifth-grade student's score exceeded the scores of 87 percent of the district's fifth graders, then he would have scored at the 87th percentile on the new math test.

performance-level categories Classifications employed as a way of signifying the quality of a student's performance on a test, for example, when cut-scores are used to place a student in such performance categories as *distinguished, proficient,* or *unacceptable* based on the student's test scores.

performance task The task that a student must perform during a performance test typically calls for the student to complete a sometimes complex, particularly demanding task, such as writing a sophisticated analysis of a complicated real-world science problem.

A continuing deterrent to the use of performance tests has been the costs associated with their scoring—because of the need for human scorers. As electronic technology continues to make reduced-cost scoring procedures affordable, it is expected that we will see the increased use of performance tests.

psychometrician A specialist in the development, refinement, administration, and analysis of results for educational tests.

psychomotor Educational outcomes associated with small-muscle and large-muscle skills, such as might be encountered in an automotive class or a physical education course.

raw score Typically the number of items a student answers correctly on a test, the raw score can sometimes accommodate the compilation of items weighted differently.

reliability Consistency of measurement. See Chapter 6 for its nuances.

rubric A scoring guide employed to help evaluate the quality of students' performances on constructed-response items, such as student-generated original essays or portfolios.

scale scores Because scale scores can be statistically analyzed more readily than students' raw scores, scale scores constitute a set of converted raw scores using an arbitrarily chosen score scale. Scale scores, without some sort of interpretative assistance, may be more analyzable, but they are less readily interpretable than, say, percentiles.

score report Once students' responses to an educational test have been scored and analyzed, those students' performances are displayed to those concerned in the form of a score report. To the extent that such score reports are not readily interpretable, then a key link in the educational testing cycle typically renders the entire cycle dysfunctional.

score-spread For tests whose chief purpose is to compare test-takers' performances, the more variation that exists in students' scores, the better. Fine-grained comparisons are more likely to carry out when a set of students' scores display sufficiently large "standard deviations" or "variances," both statistical indicators of the degree to which a test's scores are dispersed. The more score-spread, the more accurately comparisons among test-takers can be made.

self-report inventory Of particular use when assessing students' affective dispositions, such inventories ask respondents to reply to a series of statements or similar stimuli by supplying their anonymous reactions to whatever is contained in one of these self-report inventories.

standard deviation A commonly used index of the degree to which a distribution of test scores is widely dispersed. The larger the standard deviation, the more score- spread that will exist in a set of test scores.

standard error of measurement A numerical estimate of the consistency represented by a test-taker's score on two different administrations of the same test or on two different forms of a test. Typically presented as a plus-or-minus potential error interval, a smaller standard error of measurement (SEM) indicates a test is more consistent than a larger SEM.

standard-setting study Nearly always used when establishing the cut-scores for high-stakes educational tests, such studies call for a panel of nonpartisan individuals to review the actual items on a test, then iteratively arrive at a consensus regarding which cut-scores to recommend. In most standard-setting studies, the panel (often consisting of about one or two dozen individuals) is heavily influenced by "impact data" indicating the likely real-world consequences if cut-scores were set at various points.

standardized tests A test that is administered, scored, and interpreted in a standard, predetermined manner. Although most standardized educational tests are developed by educational measurement organizations, standardized tests are also developed by state education agencies or large school districts.

subskill When a curricular aim calls for students to master a particularly challenging cognitive skill, it is often the case that students must first master lesser cognitive skills that contribute to the mastery of the challenging skill. These contributory skills are usually referred to as subskills. In a learning progression, such subskills are regarded as "building blocks."

summative In contrast to "formative" procedures or products focused on the improvement of yet-malleable instructional sequences, "summative" refers to final version, completed procedures or products. Educational tests used in a summative fashion are typically employed to evaluate the quality of a mature instructional program.

target curricular aim A significant educational outcome sought for students, usually after an instructional period of some duration. When a learning progression is employed as the organizing framework for implementing the formative-assessment process, the overriding, final learning outcome sought can be described as the learning progression's target curricular aim.

test-retest evidence of reliability The degree to which test-takers' scores are similar on two time-separated administrations of the same test. Such evidence is also referred to as "stability" evidence of reliability.

universal design A strategy for developing educational assessments that, from the earliest moments of the test-development process, strives to maximize the accessibility of a test for all of its intended users.

validity The degree to which evidence supports the accuracy of score-based interpretations (inferences) about students related to the purpose for which an educational test is being used.

validity argument A justification regarding the degree to which accumulated evidence and theory support the accuracy of intended inferences in relation to the specific purpose for which an educational test is being used.