Congressional Research Service
Informing the legislative debate since 1914

Basic Concepts and Technical Considerations in Educational Assessment: A Primer

Rebecca R. Skinner
Specialist in Education Policy

December 19, 2017

Congressional Research Service
7-5700
www.crs.gov
R45048

CRS REPORT
Prepared for Members and
Committees of Congress

Summary

Federal education legislation continues to emphasize the role of assessment in elementary and secondary schools. Perhaps most prominently, the Elementary and Secondary Education Act (ESEA), as amended by the Every Student Succeeds Act (ESSA; P.L. 114-95), requires the use of test-based educational accountability systems in states and specifies the requirements for the assessments that states must incorporate into state-designed educational accountability systems. These requirements are applicable to states that receive funding under Title I-A of the ESEA. More specifically, to receive Title I-A funds, states must agree to assess all students annually in grades 3 through 8 and once in high school in the areas of reading and mathematics. Students are also required to be assessed in science at least once within each of three specified grade spans (grades 3-5, 6-9, and 10-12). The results of these assessments are used as part of a state-designed educational accountability system that determines which schools will be identified for support and improvement based on their performance. The results are also used to make information about the academic performance of students in schools and school systems available to parents and other community stakeholders.

As student assessments continue to be used for accountability purposes under the ESEA as well as in many other capacities related to federal programs (e.g., for identifying students eligible to receive extra services supported through federal programs), this report provides Congress with a general overview of assessments and related issues. It discusses different types of educational assessments and uses of assessment in support of the aims of federal policies. The report aims to explain basic concepts related to assessment in accessible language, and it identifies commonly discussed considerations related to the use of assessments. The report provides background information that can be helpful to readers as they consider the uses of educational assessment in conjunction with policies and programs. This report accompanies CRS Report R45049, *Educational Assessment and the Elementary and Secondary Education Act*, by Rebecca R. Skinner, which provides a more detailed examination of the assessment requirements under the ESEA.

The following topics are addressed in this report:

Purposes of Assessment: Assessments are developed and administered for different purposes: instructional, diagnostic, predictive, and evaluative. Increasingly, states are attempting to use assessments for these purposes within a balanced assessment system. A balanced assessment system often incorporates various assessment types, such as formative and summative assessments. Formative assessments are used to monitor progress toward a goal and summative assessments are used to evaluate the extent to which a goal has been achieved.

Types of Tests: Educational assessments can be either norm-referenced tests (NRTs) or criterion-referenced tests (CRTs). An NRT is a standardized test that compares the performance of an individual student to the performance of a large group of students. A CRT compares the performance of an individual student to a predetermined standard or criterion. The majority of tests used in schools are CRTs. The results of CRTs, such as state assessments required by Title I-A of the ESEA, are usually reported as scaled scores or performance standards. A scaled score is a standardized score that exists along a common scale that can be used to make comparisons across students, across subgroups of students, and over time. A performance standard is a generally agreed upon definition of a certain level of performance in a content area that is expressed in terms of a cut score (e.g., basic, proficient, advanced).

Technical Considerations in Assessment: The technical qualities of assessments, such as validity, reliability, and fairness, are considered before drawing conclusions about assessment results.

Validity is the degree to which an assessment measures what it is supposed to measure. Reliability is a measure of the consistency of assessment results. The concept of fairness is a consideration of whether there is equity in the assessment process. Fairness is examined so that all participants in an assessment are provided the opportunity to demonstrate what they know and can do.

Using Assessment Results Appropriately: Assessment is a critical component of accountability systems, such as those required under Title I-A of the ESEA, and can be the basis of many educational decisions. An assessment can be considered low-stakes or high-stakes, depending on the type of educational decisions made based on its result. For example, a low-stakes assessment may be a formative assessment that measures whether students are on-track to meet proficiency goals. On the other hand, a state high school exit exam is a high-stakes assessment if it determines whether a student will receive a diploma. When the results of assessments are used to make high-stakes decisions that affect students, teachers, districts, and states, it is especially important to have strong evidence of validity, reliability, and fairness. It is therefore important to understand the purpose of educational assessments, and the alignment between the purpose and their use, and to give consideration to the appropriateness of inferences based on assessment results.

A glossary containing definitions of commonly used assessment and measurement terms is provided at the end of this report. The glossary provides additional technical information that may not be addressed within the text of the report.

Contents

Appendixes

Contacts

Overview

Federal education legislation continues to emphasize the role of assessment in elementary and secondary schools. Perhaps most prominently, the Elementary and Secondary Education Act (ESEA), as amended by the Every Student Succeeds Act (ESSA; P.L. 114-95), requires the use of test-based educational accountability systems in states and specifies the requirements for the assessments that states must incorporate into state-designed educational accountability system. These requirements are applicable to states that receive funding under Title I-A of the ESEA, which authorizes aid to local educational agencies (LEAs) for the education of disadvantaged children. Title I-A grants provide supplementary educational and related services to low-achieving and other students attending elementary and secondary schools with relatively high concentrations of students from low-income families. All states currently accept Title I-A funds. For FY2017, the program was funded at $15.5 billion.

More specifically, to receive Title I-A funds, states must agree to assess all students annually in grades 3 through 8 and once in high school in the areas of reading and mathematics. Students are also required to be assessed in science at least once within each of three specified grade spans (grades 3-5, 6-9, and 10-12). The results of these assessments are used as part of a state-designed educational accountability system that determines which schools will be identified for support and improvement based on their performance. The results are also used to make information about the academic performance of students in schools and school systems available to parents and other community stakeholders.

These requirements have been implemented within a crowded landscape of state, local, and classroom uses of educational assessments, ranging from small-scale classroom assessments to high school exit exams. The emphasis on educational assessment within federal education policies, which has coincided with expanded assessment use in many states and localities, has led to considerable debate about the amount of time being spent taking tests and preparing for tests in schools, the fit between various types of assessments and intended uses, and optimal ways to increase the usefulness of assessments.[1]

As student assessments continue to be used for accountability purposes under the ESEA as well as in many other capacities related to federal programs (e.g., for identifying students eligible to receive extra services supported through federal programs), this report provides Congress with a general overview of assessments and related issues. It discusses different types of educational assessments and uses of assessment in support of the aims of federal policies. As congressional audiences sometimes seek clarification on how the assessments required under federal programs fit into the broader landscape of educational assessments, the report situates the types of assessment undertaken in conjunction with federal programs within the broader context of assessments used for varied purposes within schools. The report explains basic concepts related to assessment in accessible language, and it identifies commonly discussed considerations related to the use of assessments. The report provides background information that can be helpful to readers as they consider the uses of educational assessment in conjunction with policies and programs. This report accompanies CRS Report R45049, *Educational Assessment and the Elementary and Secondary Education Act*, by Rebecca R. Skinner, which provides a more detailed examination of the assessment requirements under the ESEA.

[1] For example, see https://www.washingtonpost.com/local/education/study-says-standardized-testing-is-overwhelming-nations-public-schools/2015/10/24/8a22092c-79ae-11e5-a958-d889faf561dc_story.html?utm_term=.b248615b3e7a; http://www.politico.com/story/2015/10/education-department-too-much-testing-215131; and https://www.theatlantic.com/education/archive/2016/06/how-much-testing-is-too-much/485633/.

This report begins by briefly discussing the current types of assessments used in elementary and secondary education. It then provides a framework for understanding various types of assessments that are administered in elementary and secondary schools. It broadly discusses several purposes of educational assessment and describes the concept of balanced assessment systems. The report also provides a description of technical considerations in assessments, including validity, reliability, and fairness, and discusses how to use these technical considerations to draw appropriate conclusions based on assessment results.

This report does not comprehensively or exclusively discuss specific assessments required by federal legislation. The information herein can be applied broadly to all assessments used in elementary and secondary schools, including those required by federal legislation. Examples from federal legislation, such as the ESEA and the Individuals with Disabilities Education Act (IDEA; P.L. 108-446), are used to highlight assessment concepts. The examples provided are not exhaustive but rather serve to demonstrate the application of assessment concepts to actual assessments administered in schools.

Assessments in Elementary and Secondary Education

Students in elementary and secondary education participate in a wide range of assessments, from small-scale classroom assessments to large-scale international assessments. Some assessments are required, and some are voluntary. Some assessment results are reported on an individual level, and some are reported at a group level. Some assessments have high-stakes consequences, and some do not. The most common type of assessment used in educational settings is achievement testing. Although educational assessment involves more than testing, this report uses "assessment" and "test" interchangeably.[2]

Examples of Types of Assessments

- State reading, mathematics, and science assessments required by Title I-A of the ESEA
- High school exit exams
- National Assessment of Educational Progress (NAEP)
- International assessments:
  - Programme for International Student Achievement (PISA)
  - Progress in International Reading Literacy Study (PIRLS)
  - Trends in International Mathematics and Science Study (TIMSS)
- Assessments to identify children for special services (e.g., special education, English learners)

Among the assessments discussed, state assessments required by the ESEA[3] receive considerable attention in this report. These assessments are administered annually in reading and mathematics to all students in grades 3 through 8 and once in high school. In addition, science assessments are administered once in each of three grade spans (grades 3-5, 6-9, and 10-12). The results of reading and mathematics assessments are used as indicators in the state accountability systems required by Title I-A.[4] Results are aggregated and reported for various groups of students.[5]

[2] Other types of educational assessment may include student or parent interviews, rating scales, performance tasks, etc.

[3] ESEA, Section 1111(b) describes academic standards and assessments.

[4] The results of science assessments are not required to be used in the state accountability system.

Though they are not required to do so by federal law, states may require students to pass exit exams to graduate from high school.[6] A state "exit exam" typically refers to one or more tests in different subject areas, such as English, mathematics, science, and social studies. Exit exams can take several forms, including minimum competency exams,[7] comprehensive exams, end-of course exams, or some combination of the three.

Students may also participate in national assessments. The National Assessment of Educational Progress (NAEP) is a series of assessments that have been used since 1969. The NAEP tests are administered to students in grades 4, 8, and 12. They cover a variety of content areas, including reading, mathematics, science, writing, geography, history, civics, social studies, and the arts. The NAEP is a voluntary assessment for students; however, states that receive funding under Title I-A of the ESEA are required to participate in the reading and mathematics assessment for grades 4 and 8. A sample of students in each state is selected to participate in the NAEP.[8]

Some students are selected to participate in international assessments. There are currently three international assessments that are periodically administered: (1) the Programme for International Student Achievement (PISA),[9] (2) the Progress in International Reading Literacy Study (PIRLS),[10] and (3) the Trends in International Mathematics and Science Study (TIMSS).[11] Participation in international assessments is voluntary, and the countries that choose to participate can vary from one administration to the next. As with the NAEP, a sample of students from a participating country is selected to take the assessment.

Certain students also take assessments to qualify for special services. States are required by the federal government to provide special services to students with disabilities and English learners (ELs). To receive special services, a student must be found eligible for services based on a variety of assessments. States are required to designate the specific assessments that determine eligibility. In addition, states are required to assess ELs in English language proficiency, which includes the domains of listening, speaking, reading, and writing.

On the surface, it may be difficult to understand why students participate in so many assessments. Each assessment has a specific purpose and reports a specific kind of score. Teaching and learning can benefit from educational assessment, but there is a balance between the time spent on educational assessment and the time spent on teaching and learning. Determining the number and type of assessments to administer in elementary and secondary education is important, and

(...continued)

[5] For reporting purposes, results must be disaggregated within each state, local educational agency, and school by (1) each major racial and ethnic group, (2) economically disadvantaged students as compared to students who are not economically disadvantaged, (3) children with disabilities as compared to children without disabilities, (4) English proficiency status, (5) gender, and (6) migrant status. For some data elements, data must also be reported by additional subgroups (e.g., military). For accountability purposes, data are only disaggregated for the first four of the aforementioned subgroups.

[6] According to the Education Commission of the States, 15 states require students to pass exit exams to graduate from high school (see https://www.ecs.org/ec-content/uploads/Exit-Exam-Requirements-for-Class-of-2017_07.26.16.pdf). In addition, the Center on Education Policy (CEP) tracks state policies regarding exit exams. For the most recent report, see Shelby McIntosh, "State High School Exit Exams: A Policy in Transition," CEP, September 2012, https://www.cep-dc.org/displayDocument.cfm?DocumentID=408.

[7] Historically, minimal competency exams have referred to achievement in basic reading, writing, and math skills.

[8] For more information about NAEP, see https://nces.ed.gov/nationsreportcard/.

[9] See http://www.oecd.org/pisa/.

[10] See https://nces.ed.gov/surveys/pirls/.

[11] See https://nces.ed.gov/timss/.

the information in this report is intended to help policymakers as they contribute to these decisions.

Assessment Framework

Educational assessment is a complex task involving gathering and analyzing data to support decision-making about students and the evaluation of academic programs and policies. There are many ways to classify assessments in frameworks. The framework offered below is meant to provide a context for the remainder of the report and present an easily accessible vocabulary for discussing assessments. This framework addresses the various purposes of assessment, the concept of balanced assessment systems, and the scoring of assessments. After outlining a general assessment framework, this report discusses technical considerations in assessment and how to draw appropriate conclusions based on assessment results.

A glossary containing definitions of commonly used assessment and measurement terms is provided at the end of this report. The glossary provides additional technical information that may not be addressed within the text of the report.

Purposes of Educational Assessment

Educational assessments are designed with a specific purpose in mind, and the results should be used for the intended purpose. Although it is possible that a test was designed for multiple purposes and results could be interpreted and used in multiple ways, it is often the case that test results are used for multiple purposes when the test itself was designed for only one. This "over-purposing" of tests is an issue of concern in education and can undermine test validity.[12] In the sections below, four general purposes of assessment are discussed: instructional, diagnostic (identification), predictive, and evaluative.

Four General Purposes of Assessment

1. **Instructional assessments**: Assessment used to modify and adapt instruction to meet students' needs. It can be an informal or formal assessment and usually takes place within the context of a classroom.

   Example: Quiz on reading assignment

2. **Diagnostic assessments**: Assessment used to determine a student's academic, cognitive, or behavioral strengths and weaknesses.

   Example: Assessment to identify a student for special education or English language services

3. **Predictive assessments**: Assessment used to determine the likelihood that a student or school will meet a particular predetermined goal.

   Example: Interim reading assessment to determine whether a student is on-track to pass the state reading assessment required under Title I-A

4. **Evaluative assessments**: Assessment used to determine the outcome of a particular curriculum, program, or policy. The results are often compared to a predetermined goal or objective.

   Example: State reading, mathematics, and science assessments required by ESEA Title I-A

[12] See, for example, W. James Popham, "The Fatal Flaw of Educational Assessment," *Education Week*, March, 22, 2016, http://www.edweek.org/ew/articles/2016/03/23/the-fatal-flaw-of-educational-assessment.html; W. James Popham, "The Right Test for the Wrong Reasons," *Phi Delta Kappan* (vol. 96, Issue 1), September 2014.

Instructional

Instructional assessments are used to modify and adapt instruction to meet students' needs. These assessments can be informal or formal and usually take place within the context of a classroom. Informal instructional assessments can include teacher questioning strategies or reviewing classroom work. A more formal instructional assessment could be a written pretest in which a teacher uses the results to analyze what the students already know before determining what to teach. Another common type of instructional assessment is progress monitoring.[13] Progress monitoring consists of short assessments throughout an academic unit that can assess whether students are learning the content that is being taught. The results of progress monitoring can help teachers determine if they need to repeat a certain concept, change the pace of their instruction, or comprehensively change their lesson plans.

Commercially available standardized tests are often not appropriate to use as instructional assessments. It may be difficult for teachers to access assessments that are closely aligned with the content they are teaching. Even when a commercially available assessment is well aligned with classroom instruction, teachers may not receive results in a timely manner so that they can adapt instruction.

Diagnostic (Identification)

Diagnostic assessments are used to determine a student's academic, cognitive, or behavioral strengths and weaknesses. These assessments provide a comprehensive picture of a student's overall functioning and go beyond exclusively focusing on academic achievement. Some diagnostic assessments are used to identify students as being eligible for additional school services like special education or English language services. Diagnostic assessments to identify students for additional school services can include tests of cognitive functioning, behavior, social competence, language ability, and academic achievement.

The IDEA requires diagnostic assessments for the purpose of determining whether a student is a "child with a disability"[14] who is eligible to receive special education and related services. States develop criteria to determine eligibility for special education and select assessments that are consistent with the criteria for all areas of suspected disability.[15] For example, if it is suspected that a student has an "intellectual disability," a state may administer a test of cognitive functioning, such as the Wechsler Intelligence Scale for Children (WISC).[16] If it is suspected that the same student may have a speech-language impairment, a state may require hearing and vision screenings, followed by a comprehensive evaluation. If it is suspected that a student has "serious emotional disturbance," a state may administer a series of rating scales and questionnaires, such as the Behavioral and Emotional Rating Scale or the Scales for Assessing Emotional Disturbance. Assessments for special education eligibility may also involve more informal measures such as parent interviews and classroom observations.

Title I-A of the ESEA requires diagnostic assessments for the purpose of determining whether a student has limited English proficiency. States must ensure that local educational agencies (LEAs) annually assess ELs to determine their level of English language proficiency. The

[13] See, for example, Stanley L. Deno, "Curriculum-based Measures: Development and Perspectives," Research Institute on Progress Monitoring, at http://progressmonitoring.net/CBM_Article_Deno.pdf.

[14] IDEA, Section 614(b) describes evaluation procedures for students referred for special education eligibility.

[15] IDEA, Section 614(b)(3).

[16] Tests of "cognitive functioning" are sometimes referred to as intelligence tests.

assessment must be aligned to state English language proficiency standards within the domains of speaking, listening, reading, and writing.[17] Most states currently participate in the WIDA consortium,[18] which serves linguistically diverse students.[19] The consortium provides for the development and administration of ACCESS 2.0, which is currently the most commonly used test of English proficiency.[20]

Predictive

Predictive assessments are used to determine the likelihood that a student or school will meet a particular predetermined goal. One common type of predictive assessment used by schools and districts is a benchmark (or interim) assessment, which is designed primarily to determine which students are on-track for meeting end-of-year achievement goals. Students who are not on-track to meet these goals can be offered more intensive instruction or special services to increase the likelihood that they will meet their goals. Similarly, entire schools or districts that are not on-track can undertake larger, programmatic changes to improve the likelihood of achieving the end goals.

Some states are now using a common assessment that is aligned with the Common Core State Standards (CCSS)[21] to meet federal assessment requirements under the ESEA. There are two common assessments currently in place: the Partnership for Assessment of Readiness for College and Career (PARCC)[22] and the Smarter Balanced Assessment Consortium (SBAC).[23] Both PARCC and SBAC administer interim assessments that are intended to be predictive of end-of-year performance.

Evaluative

Evaluative assessments are used to determine the outcome of a particular curriculum, program, or policy. Results from evaluative assessments are often compared to a predetermined goal or objective. These assessments, unlike instructional, diagnostic, or predictive assessments, are not necessarily designed to provide actionable information on students, schools, or LEAs. For example, if a teacher gives an evaluative assessment at the end of a science unit, the purpose is to determine what a student learned rather than to plan instruction, diagnose strengths and weaknesses, or predict future achievement.

Assessments in state accountability systems are typically conducted for an evaluative purpose. These assessments are administered to determine the outcome of a particular policy objective (e.g., determining the percentage of students who are proficient in reading). For example, under the ESEA, states must conduct annual assessments in reading and mathematics for all students in grades 3 through 8 and once in high school. Results from these assessments are then used in the state accountability system to differentiate schools based, in part, on student performance.[24] Some

[17] ESEA, Section 1111(b)(2)(G).

[18] WIDA is a historical acronym referring to three states receiving an initial grant for the organization: Wisconsin (WI), Delaware (D), and Arkansas (A), hence WIDA. When Arkansas dropped out, the acronym referred to World-class Instructional Design and Assessment. This descriptor, however, no longer fits the mission of the group, and it has since come to be known simply as WIDA.

[19] To view a national map of WIDA participation, see https://www.wida.us/membership/states/.

[20] For more information on ACCESS 2.0, see https://www.wida.us/assessment/ACCESS20.aspx.

[21] For more information, see http://www.corestandards.org/.

[22] For more information, see http://parc-assessment.org/.

[23] For more information, see http://www.smarterbalanced.org/.

[24] For more information, see CRS In Focus IF10556, *Elementary and Secondary Education Act: Overview of Title I-A* (continued...)

states currently use common assessments (PARCC and SBAC) to meet these federal requirements; other states have opted to use state-specific assessments.

The assessment indicators required by the accountability system in the ESEA are based primarily on the result of evaluative assessments. Because these indicators are often reported following the end of an academic year, it would be difficult to use them for instructional or predictive purposes. It would be unlikely to use the results of these assessments to guide instruction for individual students.

Balanced Assessment System: Formative and Summative Assessments

One assessment cannot serve all the purposes discussed above. A balanced assessment system is necessary to cover all the purposes of educational assessment. A balanced assessment system would likely include assessments for each aforementioned purpose. Federal requirements under the ESEA call for evaluative assessments to be used in the accountability system. States and LEAs, however, conduct additional assessments to serve other purposes in creating a more balanced assessment system. The addition of instructional, diagnostic, and predictive assessments at the state and local levels may contribute to the perception that there are "too many assessments." And while assessments may occasionally intrude on instructional time, some of these are conducted to guide and improve instruction (i.e., instructional and diagnostic assessments).

Balanced Assessment System

Balanced assessment system: An assessment system that may include assessments for instructional, diagnostic (identification), predictive, and evaluative purposes. It may include both formative and summative assessments.

Formative assessment: A type of assessment that is used during the learning process in order to improve curricula and instruction. It is a process of assessment that teachers use within the classroom to determine gaps in a student's knowledge and to adjust instruction accordingly. Formative assessment takes place within a relatively short time frame and is mainly used to inform the teaching process.

Example: Small-scale, classroom-based assessments, interim assessments

Summative assessment: A type of assessment that is generally given at the end of a lesson, semester, or school year to "sum up" what the student knows and has learned.

Example: State reading, mathematics, and science assessments required by ESEA Title I-A

One type of balanced assessment system uses a combination of formative and summative assessments. This type can be seen as overlapping with the purposes of assessment discussed above. That is, the purposes of assessment are embedded within "formative" and "summative" assessments.

Generally speaking, formative assessments are those that are used during the learning process in order to improve instruction, and summative assessments are those that are used at the end of the learning process to "sum up" what students have learned. In reality, the line between a formative assessment and a summative assessment is less clear. Depending on how the results of an assessment are used, it is possible that one assessment could be designed to serve both formative and summative functions. The distinction, therefore, between formative and summative assessments often is the manner in which the results are used. If an assessment has been designed

(...continued)

Academic Accountability Provisions, by Rebecca R. Skinner.

so that results can inform future decision-making processes in curriculum, instruction, or policy, the assessment is being used in a formative manner (i.e., for instructional, diagnostic, and predictive purposes). If an assessment has been designed to evaluate the effects or the outcome of curricula, instruction, or policy, the assessment is being used in a summative manner (i.e., for diagnostic or evaluative purposes). For example, a teacher may give a pretest to determine what students know prior to deciding what and how to teach. The results of the pretest may be used to plan instruction; therefore, the pretest is a formative assessment. When the teacher has finished teaching a certain concept or topic, however, the same test could be administered as a posttest. The results of the posttest may be used as the student's grade; therefore, the posttest is a summative assessment.

In a balanced assessment system, a state must consider its and the LEAs' needs for various types of information and choose formative and summative assessments consistent with those needs.

Formative Assessment

While this topic has received a lot of attention in recent years, there is no universal agreement on what constitutes a formative assessment. Teachers, administrators, policymakers, and test publishers use the term "formative assessment" to cover a broad range of assessments, from small-scale, classroom-based assessments that track the learning of individual students to large-scale interim assessments that track the progress of a whole school or district to determine if students will meet certain policy goals. The confusion over exactly what a formative assessment is has led some in the testing industry to avoid the term altogether and others to offer alternative names for certain types of formative assessment.[25] In this section, various types of assessments that have been described as formative will be discussed, including classroom-based and interim assessments.

Formative assessments are often used in the classroom. They can be as informal as teacher questioning strategies and observations or as formal as standardized examinations. Teachers use formative assessments for both instructional and predictive purposes. The results of formative assessment can be used to determine holes in a student's knowledge and to adjust instruction accordingly. Teachers may adjust their instruction by changing the pace of instruction, changing the method of delivery, or repeating previously taught content. After these adjustments, teachers may administer another assessment to determine if students are learning as expected. The process of administering assessments, providing feedback to the student, adjusting instruction, and re-administering assessments is what makes the assessment formative.

To supplement classroom-based formative assessments, test publishers began promoting commercial formative assessment products in the form of interim assessments. Some testing experts believe that referring to interim assessments as "formative" is inaccurate because the results are not likely to generate information in a manner timely enough to guide instruction. Others believe that these assessments can be used in a formative way to determine how school or LEA practices need to change to meet policy goals. The latter position considers the use of interim assessments as formative assessment at the school or LEA level as opposed to the classroom level. Instead of adjusting teaching practices to increase student learning, this type of

[25] Scott J. Cech, "Test Industry Split Over 'Formative' Assessment," *Education Week*, September 17, 2008, at http://edweek.org/ew/articles/2008/09/17/04formative_ep.h28.html. Marianne Perie, Scott Marion, Brian Gong, and Judy Wurtzel, "The Role of Interim Assessments in a Comprehensive Assessment System: A Policy Brief," Achieve, Inc., The Aspen Institute, and the National Center for the Improvement of Educational Assessments, Inc., November 2007.

formative assessment would require adjusting school or district practices to increase student achievement across the board. Interim assessments can track the progress of students, schools, and LEAs toward meeting predetermined policy goals. For example, as discussed above, PARCC and SBAC provide interim assessments as part of their state assessment systems for reading and mathematics. The results of the interim assessments can be used in a formative way to adjust school or district policies and practices that affect student achievement.

While both classroom-based assessments and interim assessments can be considered formative, they are not interchangeable. In classroom-based formative assessment, results are often immediate and instructionally relevant, allowing teachers to adjust instruction. On the other hand, this type of formative assessment may not provide any information on whether a student or school is on-track to be proficient on a future summative assessment. In the case of interim assessments, the content and timing of the assessment is usually determined by the state, not the teacher, making it a less flexible classroom tool. In addition, interim assessments are less likely to be used to guide classroom instruction because the results of the assessments may not be reported quickly enough to be useful to a classroom teacher. Interim assessments can, however, be used to predict whether a school or district is likely to meet predetermined goals on a later summative assessment and to identify areas requiring additional support.

Summative Assessment

Summative assessments are tests given at the end of a lesson, course, or school year to determine what has been learned. Summative assessments are used for diagnostic or evaluative purposes. Most test results that are reported by the school, LEA, state, or media are based on summative assessments. State assessments required by the ESEA, the NAEP, international assessments, and state exit exams are all summative assessments. Some forms of summative assessment are considered high-stakes assessments because they have rewards and consequences attached to performance. For example, some states require students to pass high-stakes high school exit exams or end-of-course exams to graduate. Under the ESEA, states must use assessments in reading and mathematics to differentiate schools based, in part, on student performance in their accountability systems.

Not all summative assessments have high-stakes school or district consequences attached to the results. An end-of-unit mathematics test, for example, is a summative assessment used to determine a student's grade, but there are no school- or LEA-level consequences attached. On a larger scale, NAEP and international assessments are used to provide an overall picture of national and international achievement, but there are no major consequences associated with the results.

Types of Tests and How Results Are Reported

Test scores are reported in a variety of ways. Sometimes scores may compare an individual to a group of peers in the form of standard scores or percentiles. Other times, scores may indicate a student is "proficient" or has "met expectations" in a certain subject. Misinterpreting test scores or misunderstanding a reported score can lead to inaccurate conclusions regarding the academic performance of students, schools, districts, and states, so it is essential to understand what the reported score actually means. The following sections describe common methods of score reporting in educational assessment, including using scores from norm-referenced tests (NRTs), scores from criterion-referenced tests (CRTs), scaled scores, and performance standards. A brief discussion of the advantages and disadvantages of each type of score is provided.

Norm-referenced and Criterion Referenced Tests

Norm-referenced test (NRT): A standardized test in which results compare the performance of an individual student to the performance of a large group of students.

Example: SAT

Criterion-referenced test (CRT): An assessment that compares the performance of an individual to a predetermined standard or criterion.

Example: State reading, mathematics, and science assessments required by ESEA Title I-A

Norm-referenced Tests

An NRT is a standardized test in which results compare the performance of an individual student to the performance of a large group of students. NRTs are sometimes referred to as scores of "relative standing." NRTs compare individual scores to a normative sample, which is a group of students with known demographic characteristics (e.g., age, gender, ethnicity, or grade in school). Comparisons are made using two statistical properties of the normative sample: the mean and the standard deviation.[26]

NRTs produce raw scores that are transformed into standard scores using calculations involving the mean and standard deviation. The standard score is used to report how a student performed relative to peers. Standard scores are often reported as percentiles because they are relatively easy for parents and educators to interpret, but there are many other types of standard scores that may be reported.[27]

Commercially available cognitive and achievement tests are often norm-referenced. For example, the SAT, the Graduate Record Examination (GRE), and the WISC are norm-referenced tests. Language proficiency tests used to identify students who are ELs, such as ACCESS 2.0, are also NRTs. Generally speaking, any test that can report results as a percentile is norm-referenced because it is comparing an individual score against a normative sample.

NRTs are particularly useful due to their ease of administration and scoring. Commercially available NRTs usually require no further development or validation procedures, so they are relatively cost-effective and time-efficient. NRTs can often be administered to large groups of students at the same time and are useful for making comparisons across schools, districts, or states.

On the other hand, NRTs have been criticized for several reasons. Some fault NRTs for measuring only superficial learning through multiple choice and short-answer formats instead of measuring higher-level skills such as problem solving, reasoning, critical thinking, and comprehension. Others have criticized NRTs for lacking instructional utility because they sample a wide range of general skills within a content area, but NRTs are rarely linked to the standards or curriculum.[28] In

[26] The mean is the arithmetic average of scores in the normative sample. The standard deviation is a measure of the degree of dispersion or variability within the normative sample. In simple terms, the mean is the average score and the standard deviation is a measure of how spread out students' scores are from the average score.

[27] The most commonly reported standard scores are z-scores and T-scores. For more information, see American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), "Standards for Educational and Psychological Testing" (Washington, DC: American Psychological Association, 2014).

[28] See, for example, http://www.fairtest.org/sites/default/files/norm%20refrenced%20tests.pdf; http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/A-guide-to-standardized-testing-The-nature-of-assessment.

addition, results from NRTs can be difficult for educators to interpret because there is no designation of what score denotes mastery or proficiency.

Criterion-Referenced Tests

A CRT compares the performance of an individual to a predetermined standard or criterion. Like NRTs, CRTs are often standardized. They do not, however, report scores of "relative standing" against a normative sample. CRTs report scores of "absolute standing" against a predetermined criterion. CRTs are designed to determine the extent to which a student has mastered specific curriculum and content skills. "Mastery" of curriculum and content skills is usually determined through a collaborative process involving policymakers, educators, and measurement professionals. Different levels of mastery are set through a combination of measurement techniques and professional judgment. Mastery can be defined in many ways. In the classroom, it may be defined as answering 80% of the items on an assessment correctly. Alternatively, it may be defined as meeting some level of proficiency within a content area based on an observation of the student performing the skills.

Unlike NRTs, CRTs are not necessarily designed to differentiate between students or compare an individual student to a normative group. CRT results may be reported as grades, grade equivalents, pass/fail, number correct, percentage correct, scaled scores, or performance standards. They may be measured by multiple choice formats, short-answer formats, rating scales, checklists, rubrics, or performance-based assessments. CRTs are flexible and can be designed to meet various educational needs.

The major advantage of CRTs is that they are versatile tests that can be used for a variety of purposes. While many CRTs, like state assessments, are summative tests used for evaluative purposes, other CRTs can be used for instructional, diagnostic, or predictive purposes. They can be directly linked to the standards and curriculum, and the results from CRTs can be used for planning, modifying, and adapting instruction. Additionally, like commercially available NRTs, commercially available CRTs are relatively cost-effective and time-efficient. A disadvantage of CRTs is that they do not typically facilitate good comparisons across schools, LEAs, and states. When using CRTs, there is no normative sample; therefore, there is no common metric for comparisons. It is possible to design CRTs so that comparisons can be made. However, to facilitate good comparisons, it would be necessary to have (1) consistent standards across schools, LEAs, and states; and (2) consistent definitions of "mastery" across schools, districts, and states.[29]

In elementary and secondary education, one way that test designers create a common metric for comparisons with CRTs is by using scaled scores and performance standards. Scaled scores and performance standards are two different ways of representing the same assessment result. A scaled score is a single score that exists along a scale ranging from limited mastery of a content

[29] Some states have adopted common standards and assessments. The Common Core State Standards Initiative (CCSSI) was a state-led effort to develop common standards in English and mathematics (see http://www.corestandards.org/). In addition, the U.S. Department of Education administered a competitive federal grant program for consortia of states to develop common assessments aligned with the common standards (see https://www2.ed.gov/programs/racetothetop-assessment/index.html). Two competitive grants were awarded, and two common assessments were developed: Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC). Both consortia chose to align their common assessments with the common core state standards. Some states choose to implement common assessments. In these cases, comparisons across states may be possible; however, several states use a blended approach, using some common assessment items and some state-specific items. As such, comparisons among states remains difficult.

area to complete mastery of it. A performance standard is a description of whether a certain level of mastery is achieved based on the grade level of the student. These types of scores are discussed in more detail below.

Scaled Scores. State assessments used for accountability often report a scaled score. A scaled score is a standardized score that exists on a common scale that can be used to make annual and longitudinal comparisons across students and subgroups of students. "Scaling" is conducted by adjusting a raw score based on the differences in form difficulty from a "reference form."[30] Just as an NRT score can be compared to a "normative group" of students, a scaled score can be compared to the "reference form." In the case of scaled scores, students and subgroups of students can be compared to each other directly even though there is no "normative group."[31]

A scaled score is usually a three- or four-digit score that exists across a scale with cut points determining various levels of mastery.[32] Sometimes, the scaled score is accompanied by a grade level. For each grade level, there is a range of scaled scores that corresponds to students achieving mastery of a specific content area.[33]

Scaled scores are particularly useful if they are "vertically scaled." A vertically scaled score can be compared across time and can be used to measure growth of students and student subgroups. A vertically scaled assessment is independent of grade level; however, the descriptors attached to the scaled score (e.g., basic, proficient, advanced) change according to the grade level.[34] For example, consider a group of third grade students and a group of fifth grade students that both scored 300 on a reading assessment. The two groups are comparable in terms of their reading ability; however, a scaled score of 300 may represent that the third-grade students "met expectations" but the fifth-grade students "did not meet expectations."

Performance Standards. A performance standard is another way to report results from a CRT. Performance standards are also sometimes referred to as achievement levels.

A performance standard is a generally agreed upon definition of a certain level of performance in a content area that is expressed in terms of a cut score. The predetermined cut score denotes a level of mastery or level of proficiency within a content area. An assessment system that uses performance standards typically establishes several cut scores that denote varying levels of mastery. For example, NAEP uses a system of performance standards with three achievement levels: basic, proficient, and advanced. Similarly, common assessments use performance standards to determine whether students met expectations. SBAC uses a four-level system (Level 1 through Level 4), which corresponds to "novice, developing, proficient, and advanced." PARCC uses a five-level system (Level 1 through Level 5), which corresponds to "did not yet

[30] A reference form is the initial base form of an assessment. When alternate forms of an assessment are developed, they can be compared back to the reference form. Usually, a reference form has been administered previously so that it can serve as a reference as new forms are developed.

[31] For more information on the process of "scaling" or "equating," see Xuan Tan and Rochelle Michel, "Why Do Standardized Testing Programs Report Scaled Scores?" *R&D Connections* (ETS, 2011); https://www.ets.org/Media/Research/pdf/RD_Connections16.pdf

[32] See, for instance, an example of PARCC score results (http://parcc-assessment.org/assessments/score-results) and SBAC score results (http://www.smarterbalanced.org/assessments/scores).

[33] For an example of how grade level can correspond to levels of mastery, see the SBAC score results example at http://www.smarterbalanced.org/assessments/scores.

[34] These descriptors are often referred to as performance standards or achievement levels and are discussed in the following section.

meet expectations, partially met expectations, approached expectations, met expectations, and exceeded expectations."

Performance standards can be aligned with state content standards and curricula, and results can be used for planning, modifying, and adapting instruction. The main difference between reporting a score as a scaled score or a performance standard is the label itself, which can attach meaning to a score and provide an appropriate context. A CRT may report that a student scored 242 on a scale of 500, but the score of 242 may be meaningless to most educators and parents unless there is some context surrounding it. Performance standards provide the context. If the cut score to meet expectations was predetermined to be 240, a score of 242 would be above the cut score, and therefore the student would be considered to have "met expectations" in the content area.

Although they can provide a meaningful context for assessment results, performance standards are criticized for being imprecise and for their inability to adequately measure student growth. While there are rigorous methods of determining cut scores that denote various levels of mastery,[35] there is rarely any meaningful difference between the abilities of a student who scores just below the cut score and a student who scores just above the cut score. Consider the example above in which a score of 240 is the cut score for "met expectations**."** One student may score 238 and not be considered to have met expectations, while another student may score 242 and be considered to have met expectations. In reality, the cut score of the performance standard may be making an inappropriate distinction between two students who have similar abilities. Another criticism of performance standards is that they are insensitive to student growth. Suppose the cut score for the "exceeded expectations" level is 300. A student in the previous example could move from a score of 242 to 299 within one year, making considerable progress; however, a score of 242 and a score of 299 are both considered to be within the same performance standard of "met expectations."

Technical Considerations in Assessment

This section will discuss technical considerations such as validity, reliability, and fairness. It is generally the responsibility of the test developer to investigate the technical characteristics of an assessment and report any relevant statistical information to test users. Usually, this information is reported in testing manuals that accompany the assessment. It is the responsibility of the test user to administer the test as intended and use the reported information concerning validity, reliability, and fairness to interpret test results appropriately.

Learning how to evaluate the validity, reliability, and fairness of an assessment allows test users to make appropriate inferences (i.e., conclusions drawn from the result of a test). Inferences may be either appropriate or inappropriate based on a number of technical and contextual factors. Following a discussion of the concepts of validity, reliability, and fairness, this report will conclude with a discussion of how to avoid making inappropriate inferences from educational assessments. It will also highlight some of the issues to consider when making inferences from high-stakes assessments versus low-stakes assessments.

[35] For more information on setting cut scores, see Michael Zieky and Marianne Perie, "A Primer on Setting Cut Scores on Tests of Educational Achievement," ETS, 2006; https:///www.ets/org/research/policy_research_reports/publications/publication/2006/dbkw.

Validity, Reliability, and Fairness

Validity: The degree to which accumulated evidence and theory support specific interpretations of test scores based on proposed uses of a test.

Reliability: The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group.

Fairness in testing: The principle that every test taker should be assessed in an equitable way.

Validity

Validity is arguably the most important concept to understand when evaluating educational assessments. When making instructional or policy decisions on the basis of an assessment, the question is often asked, "Is the test valid?" Validity, however, is not a property of the test itself; it is the degree to which a certain inference from a test is appropriate and meaningful.[36] The appropriate question to be asked is, therefore, "Is the inference being drawn from the test result valid?" The distinction between these questions may seem unimportant, but consider the following situation. Teachers, administrators, or policymakers often would like to support multiple conclusions from the same assessment. Some of these conclusions, or inferences, may be valid and others may not. For example, the SAT, a college entrance examination intended to measure critical thinking skills that are needed for success in college, is taken by many high school students. Suppose a group of high school seniors in School A scored well on the SAT and a group of high school seniors in School B scored poorly. One potentially valid inference from this result is that seniors from School A are more likely to succeed in college. However, there are many possible inferences that may be less valid. For example, one could infer that School A had a better academic curriculum than School B, or that School A had better teachers than School B. Neither of these inferences may be valid because the SAT was designed for the purpose of predicting the likelihood of success in college and not for the purposes of evaluating teachers or curriculum. The validity of an inference, therefore, is tied inextricably to the purpose for which the test was created.

When an assessment is created or a new use is proposed for an existing assessment, a process of validation is seen as necessary. Validation involves collecting evidence to support the use and interpretation of test scores based on the test construct. In testing, a construct is the concept or characteristic that a test is designed to measure. The process of validation includes, at a minimum, investigating the construct underrepresentation and construct irrelevance of the assessment instrument. Construct underrepresentation refers to the degree to which an assessment fails to capture important aspects of the construct. For example, if the construct of an assessment is addition and subtraction skills, the entire construct would include addition, addition with carrying, subtraction, subtraction with borrowing, two-digit addition, two-digit addition with carrying, and so forth. If the assessment does not measure all the skills within a defined construct, it may be susceptible to construct underrepresentation, and the inference based on an assessment score may not reflect the student's actual knowledge of the construct.

Similarly, construct irrelevance can threaten the validity of an inference. Construct irrelevance refers to the degree to which test scores are affected by the content of an assessment that is not part of the intended construct. Again, if the construct of an assessment is addition and subtraction

[36] For a thorough discussion of validity, see AERA, APA, NCME, "Standards for Educational and Psychological Testing," (Washington, DC: American Psychological Association, 2014).

skills, any test items that contain multiplication or division would create construct irrelevance, and the inference based on the assessment score may not reflect the student's actual knowledge of the construct.

Construct underrepresentation is investigated by answering the question, "Does the assessment adequately cover the full range of skills in the construct?" Construct irrelevance is investigated by answering the question, "Are any skills within the assessment outside of the realm of the construct?" These two questions are investigated using statistical procedures that examine properties of the assessment itself and how the properties of the assessment interact with characteristics of individuals taking the test. One important consideration is to determine if the degree of construct underrepresentation or construct irrelevance differentially affects the performance of various subgroups of the population. If, for example, there was a moderate degree of construct irrelevance (e.g., multiplication questions on an assessment designed to measure addition and subtraction skills), students from higher socioeconomic subgroups may be more likely to score well on a test than students from lower socioeconomic subgroups, even if both subgroups have equal knowledge of the construct itself. Students from higher socioeconomic subgroups are more likely to have learned the irrelevant material, given that they generally have more access to early education and enrichment opportunities.[37] The construct irrelevance, therefore, may lead to an invalid inference that students from higher socioeconomic subgroups outperform students from lower socioeconomic subgroups on a given construct.

There are many other types of evidence that may be collected during validation. For example, test developers might compare student scores on the assessment in question with existing measures of the same construct. Or, test developers might investigate how well the assessment in question predicts a later outcome of interest, such as pass rates on a high-stakes exam, high school graduation rates, or job attainment. Validation is not a set of scripted procedures but rather a thoughtful investigation of the construct and proposed uses of assessments.

Reliability

Reliability refers to the consistency of measurement when the testing procedure is repeated on a population of individuals or groups. It describes the precision with which assessment results are reported and is a measure of certainty that the results are accurate. The concept of reliability presumes that each student has a true score for any given assessment. The true score is the hypothetical average score resulting from multiple administrations of an assessment; it is the true representation of what the student knows and can do. For any given assessment, however, the score that is reported is not a student's true score; it is a student's observed score. The hypothetical difference between the true score and the observed score is measurement error. Measurement error includes student factors, such as general health, attention span, and motivation. It can also include environmental factors, such as the comfort or familiarity of the assessment location.[38] Reliability and measurement error are inversely related: the lower the measurement error, the higher the reliability. As reliability increases, the likelihood that a student's observed score and true score are reasonably equivalent is increased.

[37] For an overview of how socioeconomic status affects education issues, see American Psychological Association, *Education and Socioeconomic Status*, http://www.apa.org/pi/ses/resources/publications/education.aspx.

[38] For more information about measurement error, see AERA, APA, NCME, "Standards for Educational and Psychological Testing" (Washington, DC: American Psychological Association, 2014).

Reliability can be reported in multiple ways. The most common expressions of reliability in educational assessment are the reliability coefficient, range of uncertainty, and consistency of classification.

Reliability Coefficient

The reliability coefficient is a number that ranges from 0 to 1. It is useful because it is independent of the scale of the assessment and can be compared across multiple assessments. A reliability coefficient of 0 implies that a score is due completely to measurement error; a reliability coefficient of 1 implies that a score is completely consistent and free of measurement error. There is no rule of thumb for deciding how high a reliability coefficient should be; however, most commercially available assessments report reliability coefficients above 0.8, and many have reliability coefficients above 0.9.

The most common types of reliability coefficients used in educational assessment are alternate-form coefficients, test-retest coefficients, inter-scorer agreement coefficients, and internal consistency coefficients. Alternate-form coefficients measure the degree to which the scores derived from alternate forms of the same assessment are consistent. For example, the SAT has multiple forms that are administered each year. A high alternate-form reliability coefficient provides some certainty that a student's score on one form of the SAT would be reasonably equivalent to the student's score on another form of it. Test-retest coefficients measure the stability of an individual student's score over time. If a reading assessment was administered to a student today and re-administered in two weeks, one would expect that the student would have comparable scores across the two administrations. A high test-retest reliability coefficient provides a measure of certainty that a student's score today would be similar to the student's score in the near future. Inter-scorer agreement coefficients measure the degree to which two independent scorers agree when assessing a student's performance. A high inter-scorer agreement coefficient provides a measure of certainty that a student's score would not be greatly affected by the individual scoring the assessment.

Internal consistency coefficients are slightly more complicated. They are a measure of the correlation of items within the same assessment. If items within an assessment are related, a student should perform consistently well or consistently poorly on the related items. For example, a mathematics assessment may test multiplication and division skills. Suppose a student is proficient with multiplication but has not yet mastered division. Within the mathematics assessment, the student should score consistently well on the multiplication items and consistently poorly on the division items. A high internal consistency coefficient provides a measure of certainty that related items within the assessment are in fact measuring the same construct.

The decisions regarding the type of reliability coefficients to investigate and report depend on the purpose and format of the assessment. For example, many assessments do not use alternate forms, so there would be no need to report an alternate-form coefficient. As another example, consider a test that was designed to measure student growth over a short period of time. In this case, it may not make sense to report a test-retest reliability coefficient because one does not expect any stability or consistency in the student's score over time. Test developers also typically consider the format of the test. In tests with multiple-choice or fill-in-the-blank formats, inter-scorer agreement may not be of great concern because the scoring is relatively objective. However, in tests with constructed responses, such as essay tests or performance assessments, it may be important to investigate inter-scorer agreement because the scoring has an element of subjectivity.

Range of Uncertainty—Confidence Intervals

As stated above, reliability describes the precision with which assessment results are reported and is a measure of certainty that the results are accurate. Results can often be reported with greater confidence if the observed score is reported along with a range of uncertainty. In educational assessment, the range of uncertainty is usually referred to as a confidence interval. A confidence interval estimates the likelihood that a student's true score falls within a range of scores. The size of the confidence interval, or the size of the range, depends on how certain one needs to be that the true score falls within the range of uncertainty.

A confidence interval is calculated by using an estimated true score, the standard error of measurement (SEM),[39] and the desired level of confidence. The confidence interval is reported as a range of scores with a lower limit and an upper limit. In education, it is common to see 90%, 95%, or 99% confidence intervals. The following example illustrates how the size of the confidence interval (i.e., the range of scores) can change as the degree of confidence changes.

If the estimated true score of a student is assumed to be 100 and the SEM is assumed to be 10:

- A 90% confidence interval would be 84 to 116 (a range of 32). In this case, about 90% of the time the student's true score will be contained within the interval from 84 to 116. There is about a 5% chance that the student's true score is lower than 84 and about a 5% chance that the student's true score is higher than 116.
- A 95% confidence interval would be 80 to 120 (a range of 40). In this case, about 95% of the time the student's true score will be contained within the interval from 80 to 120. There is about a 2.5% chance that the student's true score is lower than 80 and about a 2.5% chance that the student's true score is higher than 120.
- A 99% confidence interval would be 74 to 126 (a range of 52). In this case, about 99% of the time the student's true score will be contained within the interval from 74 to 126. There is about a 0.5% chance that the student's true score is lower than 74 and about a 0.5% chance that a student's true score is higher than 126.

The illustration above demonstrates that the range of scores in a confidence interval increases as the desired level of confidence increases. A 90% confidence interval ranges from 84 to 116 (a range of 32) while a 99% confidence interval ranges from 74 to 126 (a range of 52).

Consistency of Classification

Consistency of classification is a type of reliability that is rarely reported but can be important to investigate, especially when high-stakes decisions are made with the results of educational assessments. When assessments are used to place students and schools into discrete categories based on performance (e.g., proficient vs. not proficient; pass/fail; Level 1 through Level 4; met expectations vs. partially met expectations), the consistency of classification is of interest. If students with similar abilities are not consistently classified into the same performance standard

[39] The SEM is the standard deviation of an individual's observed scores from repeated administrations of a test under identical conditions. Because such data cannot generally be collected, the SEM is usually estimated from group data. The SEM is used in the calculation of confidence intervals. A confidence interval is a range within which an assessment score will be included based on a specific level of certainty. A confidence interval is calculated by dividing the standard deviation by the square root of the sample size and multiplying this result by the confidence coefficient.

category, there may be a problem with the reliability of the assessment. Although students may move in and out of performance standard categories over time, students who achieve similarly should be consistently classified into the same performance standard category at any given time.

Within school settings, consistency of classification is particularly important when using performance standards to place students in achievement levels based on state assessments. For example, if the classification of students into achievement levels for accountability purposes is not consistent over short periods of time, the accountability system may become highly variable and unreliable. Another example of the importance of consistency of classification is the use of state exit exams to award high school diplomas (i.e., pass/fail). Without consistency in classification, the system that awards diplomas to high school seniors may be unreliable. Consistency of classification has not been well studied in these instances, but statistical modeling demonstrates that it is possible to have considerable fluctuations in classification depending on the reliability of the assessment and the predetermined cut score used to categorize students.[40]

Consistency of classification is also relevant for decisions that determine eligibility for services, such as the classification of students with disabilities. Students who are suspected to have a disability are assessed using a wide range of diagnostic assessments. Results of these assessments are interpreted based on state definitions of IDEA disability categories,[41] and students receive special education services if they are determined to be eligible. Over time, while it is possible that students become "declassified" and ineligible for special education services due to their improvement in academic skills or due to a change in the definition of "disability," it may be that the rate of "declassification" is also affected by the reliability of assessments used to determine their initial eligibility and the cut scores that are used in state definitions of disability.[42]

Fairness

Fairness is a term that has no technical meaning in testing procedures, but it is an issue that often arises in educational assessment and education policy generally. Educational assessments are administered to diverse populations, and fairness presumes that all members of each population are treated equally. The notion of fairness as "equal treatment" has taken several forms: (1) fairness as a lack of bias, (2) fairness as equitable treatment in the testing process, (3) fairness as equality in outcomes of testing, and (4) fairness as opportunity to learn.[43]

Fairness as a Lack of Bias

Bias is a common criticism in educational assessment; however, it is not well documented or well understood. Test bias exists if there are systematic differences in observed scores based on subgroup membership when there is no difference in the true scores between subgroups. For

[40] Daniel Koretz, "Error and Reliability: How Much We Don't Know What We're Talking About," in *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, MA; Harvard University Press, 2008), pp. 143-178.

[41] Under Part B of the IDEA, a child may qualify for special education and related services under any one of 14 different disability categories: autism, deaf-blindness, deafness, developmental delay, emotional disturbance, hearing impairment, intellectual disability, multiple disabilities, orthopedic impairment, other health impairment, specific learning disability, speech or language impairment, traumatic brain injury, and visual impairment. (34 C.F.R. 300.8).

[42] Students who receive special education services are reevaluated periodically for eligibility. If the reevaluation determines that the student is no longer eligible to receive special education services, he or she becomes "declassified." "Declassification" refers to a process by which a student who once received special education services is no longer eligible to receive such services.

[43] For a comprehensive discussion of fairness in testing, see AERA, APA, NCME, "Standards for Educational and Psychological Testing" (Washington, DC: American Psychological Association, 2014).

example, bias can arise when cultural or linguistic factors[44] influence test scores of individuals within a subgroup despite the individuals' inherent abilities. Or, bias can arise when a disability precludes a student from demonstrating his or her ability. Bias is a controversial topic and difficult to address in educational assessment. There is no professional consensus on how to mitigate bias in testing. There are statistical procedures, such as differential item functioning,[45] that may be able to detect bias in specific test items; however, such techniques cannot directly address the bias in the interpretation of assessment results. Test bias, if present, undermines the validity of the inferences based on assessment results.

A simple difference in scores between two subgroups does not necessarily imply bias. If a group of advantaged students performs higher on a reading assessment than a group of disadvantaged students, the test may or may not be biased. If the advantaged and disadvantaged students have the same reading ability (true score), and the advantaged students still score higher on the reading assessment (observed score), bias may be present. If, however, the advantaged students have higher reading ability and higher scores on the reading assessment, the test may not be biased.

Fairness as Equitable Treatment in the Testing Process

Fairness as equitable treatment in the testing process is less controversial and more straightforward than the issue of bias. There is professional consensus that all students should be afforded equity in the testing process. Equity includes ensuring that all students are given a comparable opportunity to demonstrate their knowledge of the construct being tested. It also requires that all students are given appropriate testing conditions, such as a comfortable testing environment, equal time to respond, and, where appropriate, accommodations for students with disabilities and ELs.

Equitable treatment affords each student equal opportunity to prepare for a test. This aspect of equitable treatment may be the most difficult to monitor and enforce. In some schools or LEAs, it is common practice to familiarize students with sample test questions or provide examples of actual test questions from previous assessments. In other LEAs, this type of test preparation may not be routine. Furthermore, some students receive test preparation services outside of the classroom from private companies, such as Kaplan, Inc. or Sylvan Learning. The amount of test preparation and the appropriateness of this preparation is not consistent across classrooms, schools, and LEAs and can undermine the validity of inferences drawn from assessments.

Fairness as Equality in Outcomes of Testing

There is no professional consensus that fairness should ensure equality in the outcomes of testing. Nonetheless, when results are used for high-stakes decisions, such as the use of state exit exams for high school graduation, the issue of "equality in outcomes" can arise. The question of fairness arises when these tests are used to exclude a subgroup of students from a desired result or certification, like earning a high school diploma. For example, if a subgroup of advantaged students is more likely to pass a state exit exam than a subgroup of disadvantaged students, the

[44] "Cultural or linguistic factors" may include the use of unnecessarily complicated vocabulary. For example, on a multiple choice assessment, instead of using the word "piranha" in a question, the assessment could use the word "fish." "Piranha" is a more complicated term that students who are English learners are less likely to understand. For more information, see Jamal Abedi and Edynn Sato, *Linguistic Modification*, U.S. Department of Education, 2008, http://www.ncela.us/files/rcd/BE024210/Linguistic_Modification.pdf.

[45] Differential item functioning is a statistical characteristic of an item that demonstrates whether it is measuring different abilities of different subgroups of participants.

advantaged students are more likely to graduate from high school, receive a diploma, pursue higher education, and obtain a job. The disadvantaged students are less likely to graduate from high school, which further disadvantages them in their pursuit of higher education or job attainment. "Equality in outcomes" is more likely to be a concern with high-stakes assessments, such as state assessments and state exit exams, than with low-stakes assessments, such as NAEP and international assessments.

Fairness as Opportunity to Learn

Fairness as opportunity to learn is particularly relevant to educational assessment. Many educational assessments, particularly state assessments used in accountability systems, are aligned with state standards and designed to measure what students know as a result of formal instruction. All students within a state are assessed against the same content and performance standards for accountability. Thus, the question arises: if all students have not had an equal opportunity to learn, is it "fair" to assess all students against the same standard? If low scores are the result of a lack of opportunity to learn the tested material, it might be seen as a systemic failure rather than a characteristic of a particular individual, school, or LEA.

The difficulty with affording all students equal opportunity to learn is defining "opportunity to learn." Is exposure to the same curriculum enough to give students the opportunity to learn? Even if all students are exposed to the same curriculum, does the overall school environment influence a student's opportunity to learn? If students are exposed to the same curriculum within the same school environment, does the quality of the classroom teacher influence a student's opportunity to learn?

Using Assessment Results: Avoiding Inappropriate Inferences

Test users have a responsibility to examine the validity, reliability, and fairness of an assessment to make appropriate inferences about student achievement. There is no checklist that will help determine if an inference is appropriate. Instead, test users are to conduct a thoughtful analysis of the assessment in terms of the construct; purpose; type of scores it reports; and evidence concerning its validity, reliability, and fairness; as well as the context in which the assessment results will be used. If these issues are not carefully considered, inappropriate inferences can lead to a variety of unintended consequences.

The sections that follow provide some guidance in the form of sample questions that can be used to consider the appropriateness of inferences about test scores. These guidelines are not intended to be an exhaustive list of considerations but rather a starting point for examining the appropriateness of conclusions drawn from assessments.[46]

Construct

Sample questions about the construct include the following: What is the content area being assessed (e.g., reading, mathematics)? What is the specific construct that is being measured

[46] The most comprehensive resource available for making judgments about educational and psychological testing is AERA, APA, NCME, "Standards for Educational and Psychological Testing" (Washington, DC: American Psychological Association, 2014).

within the content area (e.g., mathematics computation, mathematical problem solving, measurement, geometry)? Does the construct measure general knowledge within a content area, or is it specifically aligned with the curriculum?

Understanding the construct of an assessment can have important implications when comparing the results of two tests. Consider, for example, two of the international assessments mentioned earlier, PISA and TIMSS. Both assessments measure mathematics achievement, but they measure different mathematical constructs. PISA was designed to measure general "mathematical literacy," whereas TIMSS is curriculum-based and was designed to measure what students have learned in school. Students in a particular country may perform well on PISA and poorly on TIMSS, or vice versa. Because the tests measure different mathematical constructs, the assessments are likely sensitive to how mathematics is taught within the country. Thus, if the score for the United States was above the international average on a TIMSS assessment and below the international average for a subsequent PISA assessment, it would not be appropriate to infer that mathematics achievement in the United States is declining, because TIMSS and PISA measure different constructs.[47]

Purpose

Sample questions about the purpose include the following: What was the intended purpose of the assessment when it was designed (e.g., instructional, predictive, diagnostic, evaluative)? How will teachers, administrators, and policymakers use the results (e.g., formative assessment vs. summative assessment)?

Understanding the original purpose of the assessment can help test users determine how the results may be interpreted and how the scores may be used. For example, a state assessment that was designed for evaluative purposes may not lend itself to using scores to modify and adapt instruction for individual students. Most state assessments are primarily summative assessments, and it is difficult to use them in a formative manner because the results may not be reported in a timely fashion to the teachers and the items may not be sensitive to classroom instruction. Alternatively, an interim assessment that was designed for predictive purposes may report results in a more timely manner and allow teachers to target their instruction to students who scored poorly.

Interim assessments are often aligned with state summative assessments; however, scores on interim assessments are best not considered definitive indicators of what state assessment scores will be. For example, some summative assessments do not have associated interim assessments. LEAs may choose to use an interim assessment that measures the same construct as a summative assessment (e.g., reading comprehension); however, the measure may not be well-aligned with the summative assessment. If students score poorly on the interim assessment, it is not necessarily indicative that they will score poorly on the summative assessment. There may also be difficulties with the timing of an interim assessment. Classroom instruction has different pacing, depending on the school, teacher, and abilities of the students. If an LEA sets the timeline for the interim assessment, it is possible that some schools or teachers would have not yet covered the content on the interim assessment. Students would likely score poorly on the interim assessment, but if the content is covered and learned later in the year, the students may score well on the summative assessment.

[47] Other reasons that this conclusion would be inappropriate include differences in countries participating in the assessments, in sampling procedures of students participating in the assessments, and in the level of development of participating countries.

Scores

Sample questions about scores include the following: Does the score reported compare a student's performance to the performance of others (e.g., NRT)? Does the score reported compare a student's performance to a criterion or standard (e.g., CRT, scaled score, performance standard)? Does the score determine whether a student is "proficient" or has "met expectations" within a certain content area (e.g., performance standards)? Does the score show growth or progress that a student made within a content area (e.g., vertically scaled score)?

Misinterpreting scores is perhaps the most common way to make an inappropriate inference. To avoid this, a test user would fully investigate the scale of the assessment and the way in which scores are reported. If scores are reported from NRTs, a student's score can be interpreted relative to the normative sample, which is a group of the student's peers. NRTs cannot, however, determine whether a student met a predetermined criterion or whether a student is proficient within a particular content area. If scores are reported from CRTs, either in the form of scaled scores or performance standards, a student's score can be interpreted relative to a predetermined standard or criterion. When using scaled scores from CRTs, it is possible to make meaningful comparisons between students and subgroups of students. If vertically scaled scores are available, it is possible to measure student growth and make meaningful inferences about how much a student has learned over time.

Making appropriate inferences from performance standards can be particularly difficult. Because of the use of performance standards in state assessments, it is important for test users to understand what they do and do not report. Performance standards are used primarily because they can be easily aligned with the state content standards and provide some meaningful description of what students know. Performance standards are hard to interpret, however. Students are classified into categories based on their performance on an assessment, but all students within the same category did not score equally well. Furthermore, scores from performance standards do not lend themselves to measuring a student's growth. A student can score at the lower end of the "met expectations" category, make considerable progress over the next year, and still be in the "met expectations" category at the end of the year. Alternatively, a student could score at the high end of the "did not meet expectations" category, make minimal progress over the next year, and move up into the "met expectations" category. Because of these qualities of performance standards, test users should be cautious about equating the performance of students within the same category, and about making assumptions concerning growth based on movement through the categories.

Technical Quality

Sample questions about technical quality include the following: Did the test developers provide statistical information on the validity and reliability of the instrument? What kind of validity and reliability evidence was collected? Does that evidence seem to match the purpose of the assessment? Have the test developers reported reliability evidence separately for all the subgroups of interest? Was the issue of fairness and bias addressed, either through thoughtful reasoning or statistical procedures?

Commercially available assessments are typically accompanied by a user's manual that reports validity and reliability evidence. Smaller, locally developed assessments do not always have an accompanying manual, but test developers typically have validity and reliability evidence available upon request. It is a fairly simple process to determine whether evidence has been provided, but a much more difficult task to evaluate the quality of the evidence. A thorough discussion of how to evaluate the technical quality of an assessment is beyond the scope of this

report.[48] In light of the current uses of assessments in schools, however, some issues are noteworthy:

- Because schools are required to report state assessment results for various subgroups (i.e., students with disabilities and ELs), it is important that validity and reliability be investigated for each subgroup for which data will be disaggregated. Doing so will reduce the likelihood of bias in the assessment against a particular subgroup.

  The type of reliability evidence provided should be specific to the assessment. For example, an assessment with constructed responses, such as essay tests or performance assessments, will have a degree of subjectivity in scoring. In this case, it is important to have strong evidence of inter-scorer reliability. In other cases, such as when the assessment format consists of multiple choice or fill-in-the-blank items, inter-scorer reliability may be of lesser importance.

- A test like the SAT that relies on several alternate forms should report alternate-form reliability. Without a high degree of alternate-form reliability, some students will take an easier version of an assessment and others will take a more difficult version. Unequal forms of the same assessment will introduce bias in the testing process. Students taking the easier version may have scores that are positively biased and students taking the harder version may have scores that are negatively biased.
- No assessment is technically perfect. All inferences based on an observed score will be susceptible to measurement error, and some may be susceptible to bias.

Context of the Assessment

Sample questions about the context include the following: Is it a high-stakes or a low-stakes assessment? Who will be held accountable (e.g., students, teachers, schools, states)? Is the validity and reliability evidence strong enough to make high-stakes decisions? Are there confounding factors that may have influenced performance on the assessment? What other information could be collected to make a better inference?

The context in which an assessment takes place may have implications for how critical a test user must be when making an inference from a test score. In a low-stakes assessment, such as a classroom-level formative assessment that will be used for instructional purposes, conducting an exhaustive review of the reliability and validity evidence may not be worthwhile. These assessments are usually short, conducted to help teachers adapt their instruction, and have no major consequences if the inference is not completely accurate. On the other hand, for a high-stakes assessment like a state exit exam for graduation, it is important to examine the validity and reliability evidence of the assessment to ensure that the inference is defensible. Consider the consequences of a state exit exam with poor evidence of validity due to a high degree of construct irrelevance. Students would be tested on content outside of the construct and may perform poorly, which may prevent them from earning a high school diploma. Or, consider a state exit exam with poor evidence of reliability due to a high degree of measurement error. Students who are likely to score near the cut score of the assessment may pass or fail largely due to measurement error.

[48] For a comprehensive discussion on evaluating the technical quality of assessments, see AERA, APA, NCME, "Standards for Educational and Psychological Testing" (Washington, DC: American Psychological Association, 2014).

Sometimes when inferences for a high-stakes decision are being made, certain protections are placed on the testing process or the test result. For example, some states allow students to take a state exit exam for high school graduation multiple times to lower the probability that measurement error is preventing them from passing. Or, in some cases, a state will consider collecting additional data (such as a portfolio of student work) to determine whether a student has met the requirements for receiving a high school diploma. For other high-stakes decisions, such as differentiating the performance of public schools within state accountability systems, states use the results from state assessments plus other indicators (e.g., attendance rates, graduation rates, and school climate measures). When making a high-stakes decision, using multiple measures of achievement can lead to a more valid inference. While all measures should have adequate technical quality, the use of multiple measures provides protection against making an invalid inference based on one measure that may not have the strongest evidence to support its validity and reliability. If multiple measures are used, it is less likely that one measure disproportionately influences the overall result.[49]

Closing Remarks

Students in elementary and secondary education participate in a wide range of educational assessments. Assessments are an important tool at many levels—from making instructional decisions in the classroom to making policy decisions for a nation. When used correctly, educational assessments contribute to the iterative process of teaching and learning and guide education policy decisions.

Currently, a primary focus of educational assessment is tracking student academic achievement and growth in schools. Assessment is a critical component of accountability systems such as those required under Title I-A of the ESEA. At times, the results of these assessments are used to make high-stakes decisions that affect students, teachers, LEAs, and states. It is therefore important to understand the purpose of educational assessments and to give consideration to the appropriateness of inferences based on assessment results.

[49] When using multiple measures, the overall result will depend on the type, number, and weighting of the measures. It is possible that if one measure is weighted heavily, it may disproportionately influence the overall result. However, the addition of multiple measures necessarily takes weight away from other primary measures, which likely leads to a more balanced overall result. In a system that uses multiple measures, *each* measure should have strong evidence of validity and reliability.

Appendix. Glossary

alternate-form reliability	A reliability statistic that measures the degree to which scores from alternate forms of the same assessment are consistent.
assessment	Any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects, or programs.
balanced assessment system	A balanced assessment system may include assessments for instructional, diagnostic (identification), predictive, and evaluative purposes. It may include both formative and summative assessments.
bias	In a statistical context, a systematic error in a test score. In discussing fairness in testing, bias may refer to construct underrepresentation or construct irrelevance of test scores that differentially affect the performance of various subgroups of test takers.
confidence interval	In educational assessment, a range of values that is likely to contain a student's score. The size of the confidence interval depends on the level of confidence desired (e.g., 95% confidence) in the interpretation of test scores. Higher levels of confidence create larger confidence intervals.
construct	The concept or characteristic that a test is designed to measure.
construct irrelevance	The extent to which test scores are influenced by factors that are irrelevant to the construct that the test is intended to measure. Such extraneous factors distort the meaning of test scores from what is implied in the proposed interpretation.
construct underrepresentation	The extent to which a test fails to capture important aspects of the construct that the test is intended to measure. In this situation, the meaning of test scores is narrower than the proposed interpretation implies.
criterion-referenced score	A score from a test that allows its users to make interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to the performance of others. Examples of criterion-referenced interpretations include comparisons to cut scores (performance standards), interpretations based on expectancy tables, and domain-referenced score interpretations.
fairness	In testing, the principle that every test taker should be assessed in an equitable way.
formative assessment	A type of assessment that is used during the learning process in order to improve curricula and instruction. It is a process of assessment that teachers use within the classroom to determine gaps in a student's knowledge and to adjust instruction accordingly. Formative assessment takes place within a relatively short time frame and is mainly used to inform the teaching process.
generalizability	The extent to which one can draw conclusions for a larger population based on information from a sample population. Or, the extent to which one can draw conclusions about a student's ability in an entire content area based on a sample of test items from that content area.
inference	In assessment, a meaningful conclusion based on the results of the assessment.
inter-scorer agreement	A reliability statistic that measures the degree to which independent scorers agree when assessing a student's performance.

interim assessment	A type of assessment that falls between formative assessment and summative assessment. The term is not widely used but sometimes describes assessments that are used to evaluate a student's knowledge and skills within a limited time frame and to inform decisions at the classroom, school, and district level. Interim assessments may serve a variety of purposes, including instructional, predictive, or evaluative, depending on how they are designed.
internal consistency	A reliability statistic that measures the correlation between related items within the same assessment.
mean	The arithmetic average of a group of scores.
measurement error	Inaccuracy in an assessment instrument that can misrepresent a student's true score through fluctuations in the observed score. Measurement error reduces the reliability of the inference based on the observed score. Measurement error is not the same as bias, which is systematic error in the assessment instrument that tends to misrepresent scores consistently in one direction.
normative group	A group of sampled individuals designed to represent some larger population, such as test takers throughout the country. The group may be defined in terms of age, grade, or other demographic characteristics, such as socioeconomic status, disability status, or racial/ethnic minority status.
norm-referenced score	A score from a test that allows its users to make interpretations in relation to other test takers' performance within the normative group.
observed score	A score that is a result of an assessment; a reported score. In measurement, the observed score is often contrasted with the true score.
performance standard	An objective definition of a certain level of performance in some content area in terms of a cut score or a range of scores on a test. The performance standard often measures the level of proficiency within a content area.
reliability	The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group.
scaled score	A standardized score that exists on a common scale that can be used to make comparisons across students, across subgroups of students, and over time. A scaled score is a way to report a score from a criterion-referenced test.
standard deviation	A statistic that shows the spread or dispersion of scores in a distribution of scores. The more widely the scores are spread out, the larger the standard deviation.
standard error of measurement (SEM)	The standard deviation of an individual's observed scores from repeated administrations of a test under identical conditions. Because such data cannot generally be collected, the standard error of measurement is usually estimated from group data. The standard error of measurement is used in the calculation of confidence intervals.
summative assessment	In education, summative assessments are generally given at the end of a lesson, semester, or school year to "sum up" what the student knows and has learned.
test-retest reliability	A reliability statistic that measures the stability of a student's score over time.
true score	In classical test theory, the average of the scores that would be earned by an individual on an unlimited number of perfectly parallel forms of the same test. In educational assessment, a hypothetical, error-free estimation of true ability within a content area.
validation	The process through which the validity of the proposed interpretation of test scores is investigated.
validity	The degree to which accumulated evidence and theory support specific interpretations of test scores based on proposed uses of a test.

variability	The spread or dispersion of scores in a group of scores; the tendency of each score to be unlike the others. The standard deviation and the variance are the two most commonly used measures of variability.
variance	A measure of the spread or dispersion of scores. The larger the variance, the further the scores are from the mean. The smaller the variance, the closer the scores are to the mean.
vertical scaling	A measurement process that places achievement test scores within the same subject but at different grade levels onto a common or "vertical" scale. The use of a vertical scale provides a comparable measure of students' achievement growth from year-to-year.

Sources: These definitions are based primarily on those included in the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), "Standards for Educational and Psychological Testing" (Washington, DC: American Psychological Association, 2014). The one exception is the definition for a "balanced assessment system," which can be found at http://www.ccsso.org/Documents/Balanced%20Assessment%20Systems%20GONG.pdf.

Author Contact Information

Rebecca R. Skinner
Specialist in Education Policy
rskinner@crs.loc.gov, 7-6600

Acknowledgments

Erin Lomax, former CRS analyst and current independent contractor to CRS, was the lead author on this report.