Assessment in Elementary and Secondary Education: A Primer

(name redacted)
Specialist in Education Policy

October 12, 2012

Congressional Research Service
7-....
www.crs.gov
R40514

CRS Report for Congress
Prepared for Members and Committees of Congress

Summary

In recent years, federal education legislation has placed an increased emphasis on assessment in schools. Perhaps most notably, Title I-A of the Elementary and Secondary Education Act (ESEA), as reauthorized by the No Child Left Behind Act (NCLB), requires states to test all students annually in grades 3 through 8 and once in high school in the areas of reading and mathematics. These assessments are used as key indicators in an accountability system that determines whether schools are making progress with respect to student achievement. To receive Title I funding, states must also participate in the National Assessment of Educational Progress (NAEP), a standards-based national test given at grades 4 and 8. The Individuals with Disabilities Education Act (IDEA) requires states to use assessments to identify students with disabilities and track their progress according to individualized learning goals. In addition to assessments required by federal law, elementary and secondary school students generally participate in many other assessments, which range from small-scale classroom assessments to high-stakes exit exams.

This report provides a framework for understanding various types of assessments that are administered in elementary and secondary schools. It broadly discusses various purposes of educational assessment and describes comprehensive assessment systems. Common assessment measures currently used in education are described, including state assessments, NAEP, and state exit exams. The report also provides a description and analysis 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. Finally, this report provides a brief analysis of the use of assessments in accountability systems, including implications for curriculum, students, and testing.

Contents

Overview 1
Assessment Framework 2
- Purposes of Educational Assessment 2
  - Instructional 2
  - Predictive 3
  - Diagnostic (Identification) 3
  - Evaluative 3
- Comprehensive Assessment System: Formative and Summative Assessments 3
  - Formative Assessment 4
  - Summative Assessment 5
  - Relationships Between Formative and Summative Assessment 6
- Scores: How are Assessment Results Reported? 6
  - Norm-Referenced Tests 7
  - Criterion-Referenced Tests 7
  - Performance Standards 8
  - Professional Judgment 9

Current Assessments in Elementary and Secondary Schools 9
- State Assessments 10
- National Assessment of Educational Progress 12
- Assessments to Identify Students for Special Services 13
  - Students with Disabilities 13
  - LEP Students 14
  - Summary 14
- State Exit Exams 14
- Benchmark Assessments 15
- International Assessments 16
- Linking Assessments 17

Technical Considerations in Assessment 18
- Validity 19
- Reliability 20
  - Reliability Coefficient 20
  - Range of Uncertainty—Confidence Intervals 21
  - Consistency of Classification 22
- Fairness 23
  - Fairness as a Lack of Bias 23
  - Fairness as Equitable Treatment in the Testing Process 24
  - Fairness as Equality in Outcomes of Testing 24
  - Fairness as Opportunity to Learn 25
- Using Assessment Results: Avoiding Inappropriate Inferences 25
  - Construct 25
  - Purpose 26
  - Scores 26
  - Technical Quality 27
  - Context of the Assessment 28

Use of Assessments in Accountability Systems: Implications for Curriculum, Students, and Testing 28

Implications for Curriculum 29
Implications for Students 30
Implications for Testing 31

Appendixes

Appendix A. Glossary 33
Appendix B. Acronym Reference 36

Contacts

Author Contact Information 36

Overview

In recent years, federal education legislation has placed an increased emphasis on assessment in schools. Perhaps most notably, Title I-A of the Elementary and Secondary Education Act (ESEA), as reauthorized by the No Child Left Behind Act (NCLB; P.L. 107-110), has required all states that receive Title I-A funds[1] to test all public school students annually in grades 3 through 8 and once in high school in the areas of reading and mathematics.[2] These assessments are used as key indicators in an accountability system that determines whether schools are making progress with respect to student achievement. To receive Title I-A funding, states must also participate in the National Assessment of Educational Progress (NAEP), a standards-based national test given at grades 4 and 8. The Individuals with Disabilities Education Act (IDEA; P.L. 108-446) requires states to use assessments to identify students with disabilities and track their progress according to individualized learning goals. In addition to assessments required by federal law, elementary and secondary school students generally participate in many other assessments, which range from small-scale classroom assessments to high-stakes exit exams.

This report provides a framework for understanding various types of assessments that are administered in elementary and secondary schools. It broadly discusses various purposes of educational assessment and describes comprehensive assessment systems. Common assessment measures currently used in education are described, including state assessments, NAEP, and state exit exams. The report also provides a description and analysis 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. Finally, this report provides a brief analysis of the use of assessments in accountability systems, including implications for curriculum, students, and testing.

While this report does not comprehensively examine all of the assessment provisions in federal education laws, it summarizes several of the major provisions and draws on examples from federal laws, such as IDEA and NCLB, to help situate assessment concepts in the context of federal policies.

It should be noted that at this writing, NCLB provisions are still in effect in several states. Whereas in other states, alternative educational accountability systems are being developed in response to a flexibility package made available by the Secretary of Education (hereafter referred to as the Secretary) in September 2011. The flexibility package provides states with waivers exempting them from certain NCLB accountability requirements if, in their place, states develop alternative accountability systems. As it is not feasible to characterize the array of features associated with the accountability systems being developed across states, the accountability system employed under NCLB is the principal example used to discuss the application of assessments in accountability systems. A great many of the concepts related to the uses of assessments in NCLB are likely to be relevant to the use of educational assessments in other test-based accountability systems.

[1] Currently all states receive Title I-A funds.

[2] States must also administer standards-based science assessments at least once in each of three grade level ranges (3-5, 6-9, and 10-12).; however, the results of the science assessment is not used in the federal accountability system.

Assessment Framework

Educational assessment is a complex endeavor involving gathering and analyzing data to support decision-making about students and the evaluation of academic programs and policies. The most common type of assessment used in current education policy is achievement testing. Although educational assessment involves more than achievement testing, this report will use the words "assessment" and "test" interchangeably.

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 comprehensive assessment systems, and the scoring of assessments. After outlining a general assessment framework, this report will discuss current assessments in elementary and secondary schools, technical considerations in assessment, innovation in assessment, and the use of assessments in test-based accountability systems.

A Glossary is provided at the end of this report to provide definitions of common assessment and measurement terms. The Glossary provides additional technical information that may not be addressed within the text of the report. An Acronym Reference is also provided at the end of this report to provide an easily accessible list of common education and testing acronyms.

Purposes of Educational Assessment

Educational assessment does not take place in a vacuum. Generally, assessments are designed with a specific purpose in mind, and the results should be used for the intended purpose. It is possible that a test was designed for multiple purposes, and results can be interpreted and used in multiple ways. Often, however, test results are used for multiple purposes when the test itself was designed for only one. This "over-purposing" of tests is a major issue in education and can undermine test validity. In the sections below, four general purposes of assessment are discussed: instructional, predictive, diagnostic (identification), and evaluative.

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.[3] 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.

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

Predictive

Predictive assessments are used to determine the likelihood that a student or a school will meet a particular predetermined goal. One common type of predictive assessment used by schools and districts is a benchmark 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 goal. Similarly, entire schools or districts that are not on-track can undertake larger, programmatic changes to improve the likelihood of achieving the end goal.

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 services 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.

Evaluative

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

Assessments in accountability systems are conducted for an evaluative purpose. These assessments are administered to determine the outcome of a particular policy objective (e.g., determining a percentage of students who are proficient in reading). For example, under NCLB, state assessments have been used for evaluative purposes to determine whether schools have made Adequate Yearly Progress (AYP).[4] State assessments will be discussed in more detail throughout this report.

Comprehensive Assessment System: Formative and Summative Assessments

One assessment cannot serve all the purposes discussed above. A comprehensive assessment system is necessary to cover all the purposes of educational assessment. One type of comprehensive assessment system is a combination of formative assessments and summative assessments. Generally speaking, formative assessments are those that are used during the

[4] For more information on AYP, see CRS Report RL32495, *Adequate Yearly Progress (AYP): Implementation of the No Child Left Behind Act*, by (name redacted).

learning process in order to improve curriculum and 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 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, predictive, and diagnostic purposes). If an assessment has been designed to evaluate the effects or the outcome of curriculum, instruction, or policy, the assessment is being used in a summative manner (i.e., for diagnostic or evaluative purposes).

Formative Assessment

Formative assessment has received a lot of attention in recent years. That said, it is reasonably clear that there is not universal agreement over what constitutes a "formative assessment" in the field of education. It seems that teachers, administrators, policymakers, and test publishers use the term "formative assessment" to cover a broad range of assessments, from small-scale classroom assessments that track the learning of individual students to large-scale benchmark assessments that track the progress of a whole school or district to determine if they will meet certain policy goals. The confusion over exactly "What is formative assessment?" has led some in the testing industry to avoid the term altogether[5] and others to offer alternative names for certain types of formative assessment.[6] In this section, various types of assessments that have been described as formative assessments will be discussed, including classroom, interim, and benchmark assessments.

Formative assessments are often used in the classroom. They can be as informal as teacher questioning strategies or as formal as written examinations. Teachers use formative assessments for both instructional and predictive purposes. The results of formative assessment can be used to determine deficits 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."

Perhaps in response to the apparent success of formative assessment at the classroom level, test publishers began promoting commercial formative assessment products in the form of interim assessments and benchmark assessments. Some testing experts believe that referring to interim and benchmark assessments as "formative" is inaccurate, but others believe that these assessments can be used in a formative way to determine how school or district practices need to change in order to meet policy goals. The latter position considers the use of interim or benchmark assessments as formative assessments at the school or district level as opposed to the

[5] Scott J. Cech, "Test Industry Split Over 'Formative' Assessment," *Education Week*, September 17, 2008, at http://www.edweek.org/ew/articles/2008/09/17/04formative_ep.h28.html.

[6] 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 Assessment, Inc., November 2007.

classroom level. Instead of adjusting teaching practices to increase student learning, this type of formative assessment would require adjusting school or district practices to increase student achievement across the board. Interim and benchmark assessments can track the progress of students, schools, and districts toward meeting predetermined policy goals. For example, schools and districts have used benchmark assessments to determine if they are on-track to meet AYP goals as defined by NCLB.

The term "interim assessment" has been suggested to characterize assessments that fall between those which are purely formative and summative assessments.[7] Under this characterization, interim assessments are assessments used to track student achievement and to inform decisions at the classroom, school, or district level. Interim assessments can report on student achievement at the individual level or in the aggregate. The content and timing of the assessment is usually determined by the school or district, not the teacher, making it a less flexible classroom tool than a teacher-controlled, classroom-level formative assessment. Interim assessments can be used to inform classroom practice, but because teachers have less control over timing and content, the true value of interim assessment may lie at the school or district level. These assessments are usually used for predictive purposes—to determine whether a student, school, or district is likely to succeed on a later summative assessment and to identify those students who may need more intensive instruction. Another use of interim assessment may be to evaluate a short-term instructional program or a small aspect of the overall curriculum.

A benchmark assessment is a type of interim assessment that is widely used in schools and districts. Like other types of interim assessment, benchmark assessments are primarily used to predict the likelihood of success on a later summative assessment and to identify those students who may need more intensive instruction. They are also used to determine whether a student, school, or district is on-track to meet certain policy objectives, such as AYP.

Summative Assessment

Summative assessments are tests given at the end of a lesson, semester, 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 or media are based on summative assessments—state assessments, NAEP, international assessments, and state exit exams. 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 in order to graduate. Furthermore, under NCLB, all states used high-stakes assessments used to determine AYP. Although in this instance, the assessments have high stakes for schools and school districts, not for individual students.

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 district-level consequences attached. On a larger scale, NAEP and international assessments are used to get an overall picture of national and

[7] 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 Assessment, Inc., November 2007.

international achievement, but, again, there are no major consequences associated with the results.

Relationships Between Formative and Summative Assessment

Ideally, formative and summative assessments are administered in a comprehensive assessment system. In order for teachers and school administrators to use formative assessment to increase student achievement and predict outcomes on summative assessments, the two types of assessment must be closely aligned in terms of the test content and goals. One way to measure whether the assessments are in alignment is to determine the ability of a formative assessment to predict achievement on a summative assessment (i.e., determine predictive validity).

The REL Mid-Atlantic conducted an analysis of the predictive validity of benchmark assessments.[8] This analysis looked at the extent to which common, commercially developed benchmark assessments predicted performance on state assessments in Delaware, Maryland, New Jersey, Pennsylvania, and Washington, D.C. A review of four common assessments found that only one of these benchmark assessments showed strong evidence of predictive validity with state assessments. Moreover, none of the benchmark assessments demonstrated evidence of predictive validity for state assessments in Maryland and New Jersey. The ability of benchmark assessments to predict later performance on state assessments, therefore, seems to depend heavily on the benchmark assessment used and the state in which the assessment takes place. If this pattern is indicative of national use of benchmark assessments, there is evidence to suggest that these benchmarks are not serving a formative function within a comprehensive assessment system. Without having a strong predictive relationship between benchmark assessments and state assessments, school and district personnel may be unable to use the information from the benchmark assessment to predict future achievement on summative assessments, such as state assessments used in accountability systems.

Scores: How are Assessment Results 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 "advanced" in a certain subject. Misinterpreting test scores or misunderstanding the way in which scores are reported can lead to unintended negative consequences, such as making an inappropriate conclusion regarding the effectiveness of a program or policy. The following sections describe common methods of score reporting in educational assessment, including scores from norm-referenced tests (NRTs), scores from criterion-referenced tests (CRTs), performance standards, and professional judgment. A brief discussion of the advantages and disadvantages of each method is provided.

[8] Richard S. Brown and Ed Coughlin, The Predictive Validity of Selected Benchmark Assessments Used in the Mid-Atlantic Region (Issues & Answers Report, REL 2007-No.017), Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory, Mid-Atlantic, November 2007.

Norm-Referenced Tests

An NRT is a standardized test in which results compare the performance of an individual with 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 (age, gender, ethnicity, or grade in school). Comparisons are made using two statistical properties of the normative sample: the mean and the standard deviation.[9]

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 (e.g., z-scores, scale scores, or T-scores).

Commercially available cognitive and achievement tests are often norm-referenced. For example, the Stanford Achievement Test Series (SAT10) is an NRT and was used in a national evaluation of the Reading First program.[10] Language proficiency tests used to identify students with Limited English Proficiency (LEP), such as the IPT Family of Tests, are NRTs. Tests to measure cognitive ability of students with disabilities, such as the Wechsler Intelligence Scale for Children (WISC), are also NRTs.

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 be easily 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 criticize 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 curriculum. In 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

[9] 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.

[10] Beth C. Gamse, Robin Tepper Jacob, Megan Horst, Beth Boulay, and Fatih Unlu, *Reading First Impact Study Final Report* (NCEE 2009-4038), Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, November 2008; For more information on Reading First, see CRS Report RL31241, *Reading First and Early Reading First: Background and Funding*, by (name redacted) and CRS Report RL33246, *Reading First: Implementation Issues and Controversies*, by (name redacted).

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 of professional judgment. Mastery can be defined in many ways. 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 designed to differentiate between students or compare an individual student to a normative group. Because comparisons are not being made, CRTs report either scale scores or raw scores, depending on how the assessment was designed. CRT results may be reported as grades, pass/fail, number correct, percentage correct, or performance standards. They may be measured through the use of multiple choice formats, short answer, 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 instructional purposes. They can be directly linked to the 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. The disadvantage of CRTs is that they do not typically facilitate good comparisons across schools, districts, 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, that would require (a) consistent standards across schools, districts, and states, and (b) consistent definitions of "mastery" across schools, districts, and states.

Performance Standards

Interest in CRTs increased throughout the 1990s due to the emphasis on standards-based reform in education.[11] Performance standards are a type of score reporting that evolved from CRTs and standards-based reform. A CRT can often report results as either a scale score or a performance standard. 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 proficiency. For example, NAEP uses a system of performance standards with three achievement levels: basic, proficient, and advanced. Additionally, state assessments have used performance standards to determine AYP under the NCLB accountability system. Definitions are provided for each performance standard, describing the competencies and abilities associated with the label.

Performance standards have the same advantages of CRTs. Performance standards can be directly linked to the curriculum and results can be used for planning, modifying, and adapting instruction. The main difference between reporting a score as a CRT or a performance standard is the "proficiency label," 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 the score.

[11] For more information on standards-based reform, see CRS Report R41533, *Accountability Issues and Reauthorization of the Elementary and Secondary Education Act*, by (name redacted) and (name redacted).

Performance standards provide the context. If the proficiency cut score was predetermined to be 240, a score of 242 would be above the cut score, and therefore the student would be considered proficient in the content area.

Although they provide a meaningful context for assessment results, performance standards are criticized for their somewhat arbitrary cut scores. Cut scores are usually determined through a process of consensus and professional judgment, but 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 denotes "proficiency". One student may score 238 and not be considered proficient, while another student may score 242 and be considered proficient. 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 "advanced" 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 "proficient."

Professional Judgment

Occasionally, assessment calls for professional judgment. On a daily basis within classrooms, teachers ask questions and make judgments about students' knowledge based on students' responses. In some cases, teachers may use their professional judgment to refer a child for a special education evaluation or a language assessment. Another application of professional judgment may be the use of a rubric to evaluate a student's performance against a predetermined standard. Although scoring rubrics can be quite prescriptive, there are occasionally value-laden decisions that require teachers to make judgments about the degree to which a student met the standard. Professional judgment has the advantage of being directly tied to the curriculum and sensitive to individual student performance, however, it is subjective and susceptible to personal biases.

Another type of professional judgment that is used in educational assessment is a process of professional consensus used to set performance standards. For example, the National Assessment Governing Board (NAGB) is responsible for setting the policy for NAEP. One of the activities of NAGB is to set appropriate student achievement levels (i.e., performance standards) that denote varying levels of proficiency. This process of defining and reviewing performance standards includes the professional judgment of a representative panel of teachers, education specialists, and members of the general public.

Current Assessments in Elementary and Secondary Schools

Students in elementary and secondary schools are assessed using a wide range of tests. The following sections describe some of the common types of assessments used in elementary and secondary schools; these assessments are situated within the framework described above. First, assessments that are required by federal law are discussed, followed by a discussion of assessments that are required by state policies, assessments that are administered at the discretion of local districts, and voluntary assessments.

The first three sections provide a discussion of assessments that are or have been required by federal law: state assessments for AYP, NAEP, and assessments to identify students for special services. In the next section, state exit exams are discussed. These assessments are not required by federal law, but state policies often require that students participate in these assessments as a high school graduation requirement. Next, benchmark assessments are discussed. Benchmark assessments are also not required by federal law, but they are widely used by districts and states. The next section provides a discussion of international assessments. These assessments are voluntary assessments in which schools and students are periodically selected at random to participate. The final section describes the practice of linking assessments—a statistical technique used to compare scores across different tests.

State Assessments[12]

Since the reauthorization of the ESEA by the NCLB, a good deal of focus has been placed on state assessments used to calculate required annual progress (i.e., AYP).[13] Under NCLB, states that participate in the Title I-A program[14] have been required to administer standards-based assessments in reading and mathematics to students in each of grades 3-8, plus at least once in grades 10-12. Beginning with the 2007-2008 school year, states were also required to administer standards-based science assessments at least once in each of three grade level ranges (3-5, 6-9, and 10-12). NCLB provisions have required that at least 95% of all students (and at least 95% of students in all demographic subgroups used for AYP determinations) participate in the state assessment in order for a school or LEA to make AYP.

NCLB has allowed states to develop individualized state standards and individualized state assessments that appropriately measure these standards. Results from the state assessments have been used in the determination of AYP within the NCLB accountability system. Although state assessments may be individualized, they have been subject to several legislative requirements. NCLB has required that state assessments be used for the purposes for which they are valid and reliable, and that they must meet professionally recognized technical standards. Under NCLB, assessments must be aligned with challenging academic content and academic achievement standards, and they must produce coherent results that report whether students attained the achievement standards. Achievement standards, under NCLB, must include, at a minimum, three levels. In a three-level system, these achievement levels are often referred to as basic, proficient, and advanced. NCLB has required the state educational agency (SEA) to provide evidence to the Secretary that the chosen assessments are consistent with the above requirements, including providing evidence of the technical quality of the instrument.

[12] As has been noted, in September 2011, the U.S. Department of Education (ED) announced the availability of an ESEA flexibility package for states and described the principles that states must meet to obtain the included waivers (see http://www.ed.gov/esea/flexibility). The waivers would provide flexibility with regard to accountability and general administrative provisions. To obtain a waiver, a state must implement college- and career-ready standards and annual, high-quality assessments aligned with these standards that measure student growth in grades 3-8 and once in high school in the areas of reading and mathematics. These are the same grade levels and subject areas required by current law to determine AYP. Under the waiver program, however, results from these assessments would not be required to determine AYP but rather used in a new statewide accountability system.

[13] For more information, see CRS Report RL31407, *Educational Testing: Implementation of ESEA Title I-A Requirements Under the No Child Left Behind Act*, by (name redacted) and (name redacted).

[14] Currently, all states participate in ESEA, Title I-A.

Under NCLB, all students with disabilities have been required to participate in state assessments. The majority of students with disabilities have participated in the general state assessment; however, a subset of students with disabilities may take alternate assessments. The requirements for the administration of alternate assessments and the use of alternate assessments in the NCLB accountability system are outlined in regulations issued by the U.S. Department of Education.[15]

In school year 2007-2008, the Council of Chief State School Officers (CCSSO) reported data on various features of state assessments required under NCLB. The CCSSO reported that states were using both NRTs and CRTs in their state assessment systems. Under NCLB, if a state chooses to use an NRT, it must use an "augmented" NRT, which is aligned with state content and performance standards. In the 2007-2008 school year, at the elementary and middle school level, 2 states were using NRTs only, 35 states were using CRTs only, and 14 states were using a combination of NRTs and CRTs. At the high school level, 3 states were using NRTs only, 37 states were using CRTs only, and 11 states were using a combination of NRTs and CRTs. Examples of NRTs used for state assessments include the Iowa Test of Basic Skills (ITBS) and the SAT 10. Examples of CRTs used for state assessments include the Texas Assessment of Knowledge and Skills (TAKS) and the New England Common Assessment Program (NECAP).[16]

The state assessments were using a variety of test formats. Of 42 states with reported data, 7 states were using multiple choice only, 1 state was using extended response only, and 34 states were using a combination of formats. Of the 34 states that were using a combination of testing formats, 34 states were using multiple choice, 31 states were using extended response, 24 states were using short answer, and 4 states were using fill-in-the blank.

In September 2010, ED awarded grants to two consortia of states[17] to develop new state assessments.[18] Consistent with NCLB state assessment requirements, the consortia are developing reading and mathematics assessments for grades 3 through 8 and once in high school.[19] At this time, 44 states and the District of Columbia have joined at least one of the two consortia working on developing common assessments.[20] The goal of both consortia is to implement common assessments that are aligned with the common core standards for reading and mathematics by school year 2014-2015.[21]

---

[15] 34 C.F.R. §200. For more information on alternate assessments, see CRS Report R40701, *Alternate Assessments for Students with Disabilities*, by (name redacted).

[16] This information is available through the Council of Chief State School Officers' state assessment profiles for school year 2007-2008.

[17] The two consortia of states are the Smarter Balanced Assessment Consortium (SBAC; http://www.smarterbalanced.org/) and the Partnership for the Assessment of Readiness for College and Career (PARCC; http://www.parcconline.org/about-parcc).

[18] See the press release from the U.S. Department of Education: http://www.ed.gov/news/press-releases/us-secretary-education-duncan-announces-winners-competition-improve-student-asse. Funding for the grants was made available by the Race to the Top program (RTTT), authorized by the American Recovery and Reinvestment Act (P.L. 111-5). For more information on RTTT, see http://www2.ed.gov/programs/racetothetop/index.html. For more information on the RTTT assessment program, see http://www2.ed.gov/programs/racetothetop-assessment/index.html.

[19] The assessment consortia are not required to develop science assessments.

[20] The following states did not join one of the two consortia working on common assessments: Alaska, Minnesota, Nebraska, Texas, Virginia, and Wyoming. For more information, see footnote 18. A state's commitment to participate in the common assessment development process does not necessarily translate into eventual adoption and use of the assessments.

[21] For more information about common standards, see CRS Report R41533, *Accountability Issues and Reauthorization of the Elementary and Secondary Education Act*, by (name redacted) and (name redacted) and (continued...)

State assessments are summative assessments used for evaluative purposes. Results from the augmented NRTs and CRTs are reported as the percentage of students reaching a performance standard (e.g., basic, proficient, advanced). Schools are held accountable for the percentage of students scoring "proficient" or above. The goal of NCLB is to achieve 100% proficiency by the end of school year 2013-2014.[22]

The percentage of proficient students cannot be compared across states. Because each state had the discretion to develop its own assessment and choose its own cut scores denoting proficiency, there is no common measure of proficiency. If a State A chose a low cut score to denote proficiency and State B chose a high cut score to denote proficiency, State A may have a higher percentage of students reaching proficiency than State B. It would not be appropriate to conclude, however, that State A had higher student achievement levels overall.

National Assessment of Educational Progress

The NAEP is a series of assessments that have been administered since 1969. NAEP tests are administered to students in grades 4, 8, and 12, and they cover a variety of content areas, including reading, mathematics, science, writing, and, less frequently, geography, history, civics, social studies, and the arts. NAEP policies are established by NAGB, which is responsible for selecting the areas to be assessed, designing the assessment methodology, and developing guidelines for reporting and disseminating results. NAGB is an independent, bipartisan group of governors, state legislators, local and state school officials, educators, business representatives, and members of the general public. NAEP is administered uniformly to students within states (and some large urban districts) and serves as a common metric for understanding student achievement across the nation. NAEP is administered and scored by ED with assistance from contractors.[23]

There are several types of NAEP assessments. The NAEP *national assessment* began in 1969 and tests students in grades 4, 8, and 12 in all nine content areas (reading, mathematics, science, writing, geography, history, civics, social studies, and the arts); however, each subject is not assessed during every administration.[24] The NAEP *state assessment* began in 1990. The state assessment is administered every other year to students in grades 4 and 8 in the areas of reading and mathematics. States that receive Title I-A funding under NCLB are required to participate in these assessments. The NAEP *long-term trend* (LTT) assessments are given every four years and track the trends in reading and mathematics achievement since the 1970s. It is administered to a nationally representative sample of students ages 9, 13, and 17. The LTT differs from other NAEP assessments in that it has used identical questions since its original administration. The consistency of the questions allow for tracking overall national progress over time.

NAEP is a summative assessment used for evaluative purposes. NAEP is a CRT and, like state assessments, the results are reported as the percentage of students reaching a performance

(...continued)

http://www.corestandards.org/.

[22] States that have received a waiver under the ESEA Flexibility packaged offered by ED are not held accountable for this goal.

[23] The Commissioner from the National Center for Education Statistics (NCES), U.S. Department of Education is responsible by law for carrying out NAEP activities.

[24] For a schedule of NAEP national assessments, see http://nces.ed.gov/nationsreportcard/about/assessmentsched.asp.

standard (e.g., basic, proficient, advanced). Unlike state assessments, however, the performance standards of basic, proficient, and advanced are consistent, which allows comparisons to be made across states. NAEP does not report results for individual students because no student is administered the entire NAEP assessment. Students who are selected to participate take only a sample of the possible items. The scores from all students are aggregated to produce results for groups and subgroups. NAEP reports achievement results for groups of students by grade and content area (e.g., grade 4 reading and grade 8 reading) and by subgroup (e.g., gender, ethnic minorities, students with disabilities).

It is important to note that the meaning of "proficiency" is not consistent across state assessments and NAEP. For any given state, the percentage of students who are proficient in reading on the state assessment and the percentage of students who are proficient in reading on NAEP can vary greatly. As discussed earlier, states had the discretion to design their own assessments and choose their own cut scores to denote proficiency. State assessments were not based on NAEP assessment frameworks and states did not choose cut scores consistent with NAEP. It is possible, however, to compare NAEP scores across states. For example, it is valid to compare the percentage of proficient students on a NAEP 4th grade reading assessment in State A to the percentage of proficient students on a NAEP 4th grade reading assessment in State B.

Assessments to Identify Students for Special Services

Schools are required by law to provide special services to eligible students with disabilities and LEP students. To receive special services through the schools, a student must be found "eligible" for services based on a battery of assessments. For eligible students with disabilities, IDEA requires that students receive special education and related services. For eligible LEP students, ESEA requires that students receive supplemental English language instruction. IDEA and ESEA provide general guidelines to determine a student's eligibility for services, but states and districts have some flexibility in the use and interpretation of assessments for eligibility determinations.

Students with Disabilities

To be covered under IDEA, a student with a disability must meet two criteria. First, the student must be in one of several categories of disabilities,[25] and second, the student must require special education and related services as a result of the disability in order to benefit from public education. The law does not, however, provide educational definitions of these disability categories. Each state is required to develop its own educational definition of these disability categories outlined by IDEA. Because the definitions of these disability categories vary across states, the assessments used to identify a student with a disability is specific to the state and the suspected disability.

Although the actual assessments can vary across states, IDEA specifies several requirements for special education evaluations.[26] In conducting an evaluation, a local educational agency (LEA)

[25] IDEA defines a student with a disability as a student, "with mental retardation, hearing impairments (including deafness), speech or language impairments, visual impairments (including blindness), serious emotional disturbance (referred to in this title as 'emotional disturbance'), orthopedic impairments, autism, traumatic brain injury, other health impairments, or specific learning disabilities." For a more comprehensive definition, see IDEA, §602(3).

[26] IDEA, §614

must use a variety of assessment tools and strategies to gather relevant functional, developmental, and academic information, including information provided by the parent. The decision of which assessments to use depends on the disability in question and the domain to be assessed (i.e., functional, developmental, or academic). Most students with disabilities are assessed on a variety of skills and competencies outside of traditional academic assessments. For example, it is common in the assessment of students with disabilities to measure skills such as basic language, behavior and social competency, cognitive functioning, and motor skills. LEAs are responsible for interpreting the scores of these assessments and determining a student's eligibility for services based on state definitions of disability.

LEP Students

ESEA defines an LEP student as a student whose native language is a language other than English and whose difficulties in speaking, reading, writing, or understanding the English language may inhibit the individual from meeting proficient levels of achievement, succeeding in the classroom, or participating fully in society.[27] The law does not provide national eligibility criteria for LEP students. Based on the federal definition, each state determines their own eligibility criteria for LEP students to receive supplemental English language instruction.

Because the eligibility criteria vary across states, the assessments used to identify a student as an LEP student also varies across states. In general, there are two major components: a home language survey and an English language assessment. LEAs are responsible for interpreting the responses on the home language survey and interpreting scores on the English language assessment to determine LEP eligibility based on state criteria. In addition to this initial assessment, Title I-A of NCLB requires that LEP students be assessed annually in English language skills (i.e., reading, writing, speaking, and listening).

Summary

Assessments used to identify students for special services can be used as either formative or summative assessments, depending on their purpose and administration. Regardless of whether the assessment is formative or summative, these assessments are used for diagnostic purposes in order to determine a comprehensive profile of strengths and weaknesses for individual students. These assessments are a mixture of NRTs and CRTs, depending on the needs of the student and the eligibility criteria of the state.

State Exit Exams

Though not required by federal law, an increasing number of states require students to pass exit exams to graduate from high school. A state "exit exam" typically refers to one or more tests in different subject areas, such as language arts, mathematics, science, and social studies. These tests can usually be taken more than once throughout high school. Exit exams can take several forms, including minimum competency exams, comprehensive exams, end-of-course exams, or some combination of the three. A minimum competency exam focuses on basic skills below the high school level. Comprehensive exams are aligned with state standards and typically assess 9th or

[27] For a more comprehensive definition, see ESEA, §9101(25).

10th grade knowledge in several subject areas. End-of-course exams assess knowledge related to specific high school courses, such as Algebra I or U.S. History. In general, there has been a movement away from minimum competency exams toward comprehensive exams and end-of-course exams. The Center on Education Policy (CEP) publishes information state policies regarding high school exit exams annually.[28]

Very few studies of the impact of state exit exams on student achievement have been conducted. A recent national study examined the relationship between high school exit exams and achievement in reading and mathematics as measured by NAEP.[29] The results of this study indicate that high school exit exams do not lead to increases in reading and mathematics achievement and may reduce graduation rates. Furthermore, authors of this study reported that students who receive a diploma in states with required exit exams are not more successful in the labor market than students who receive diplomas in states that do not require exit exams. Other state-level studies have reported similar findings. Several reports on California's exit exam found that the exit exam may increase dropout rates and decrease enrollment in postsecondary education.[30]

State exit exams are summative assessments used for evaluative purposes. These assessments have high stakes for individual students, but there are no school- or district-level consequences associated with the results. Most exit exams are CRTs, aligned with state standards and specific curricula. Because exit exams can be taken more than one time, it is possible that they serve as a formative assessment for instructional purposes, as well. Performance on the first exit exam could serve to modify or adapt instruction in a formative way. States' use of exit exams as a formative assessment, however, has not been studied.

Benchmark Assessments

Benchmark assessments are mid-year assessments that are usually administered to determine whether a school is on-track to meet its end-of-year goals. They are used at the discretion of school administrators and the particular type of benchmark assessment is chosen at the local level. Since NCLB increased the consequences associated with performance on state assessments, there has been more demand for commercially developed benchmark assessments that are aligned with state content standards and assessments. Typically, SEAs or LEAs hire the original test publisher or an independent contractor to conduct alignment studies between benchmark assessments and state content standards. Examples of commercially developed benchmark assessments include 4Sight Math and Reading, STAR Math and Reading, Study Island Math and Reading, and TerraNova Math and Reading.

[28] CEP's annual reports on state high school exit exams can be found at http://www.cep-dc.org/index.cfm?DocumentTopicID=7.

[29] Eric Grodsky, John R. Warren, and Demetra Kalogrides, "State High School Exit Examinations and NAEP Long-term Trends in Reading and Mathematics, 1971-2004.," *Educational Policy*, (in press).

[30] D.E. (Sunny) Becker and Christina Watters, "Independent Evaluation of the California High School Exit Examination (CAHSEE): 2007 Evaluation Report, October 2007 at http://www.cde.ca.gov/ta/tg/hs/documents/evalrpt07.pdf; Andrew C. Zau and Julian R. Betts, "Predicting Success, Preventing Failure: An Investigation of the California High School Exit Exam," 2008 at http://www.ppic.org/content/pubs/report/R_608AZR.pdf.

Benchmark assessments are usually considered formative assessments and occasionally referred to as "interim assessments."[31] Most benchmark assessments report scores as performance standards. They can be used for instructional purposes if teachers use the results to identify deficits in students' knowledge and modify their instruction accordingly. Benchmark assessments are also used for predictive purposes, and well-designed benchmark assessments are closely aligned with a state assessment so that schools can predict the likelihood of making desired progress or making AYP.[32]

International Assessments

International assessments allow educators, administrators, and policymakers to get a sense of how students in the United States perform relative to other countries. Since the mid-1990s, students in the United States have participated in several international assessments, including the Program for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study (TIMSS). Participation in international assessments is voluntary and the countries that choose to participate can vary from one administration to the next. Generally, a representative sample of schools and a representative sample of students within schools are selected to participate in international assessments.

The primary purpose of PISA is to report on broad subject-area "literacy" that is not directly tied to a particular curriculum or content framework. PISA assesses 15-year-olds' performance in reading literacy, mathematics literacy, and science literacy. The first administration of PISA was in 2000; it is administered every three years.[33]

PIRLS is an assessment of reading achievement, behavior, and attitudes of 4th grade students in the United States and students who are in the equivalent of 4th grade in other countries. It was first administered in 2001 to students in 35 countries. PIRLS reports results in two ways. First, it reports national averages, which allow countries to be compared to each other. Second, PIRLS reports the percentage of students in each country that reach international benchmarks.[34] Like PISA, PIRLS does not report results for individual students.[35]

TIMSS is an assessment of science and mathematics achievement of students in grades 4 and 8 in the United States and equivalent grades in other countries. It has been administered every four years since 1995. Like the other international assessments, TIMSS reports national averages which allow countries to be compared to each other, but it does not report results for individual

[31] See the earlier discussion of "Formative Assessment" in this report.

[32] The general level of alignment of benchmark assessments and state assessments has not been thoroughly studied. There is some evidence to suggest that performance on benchmark assessments does not strongly predict performance on state assessments, which may suggest a lack of alignment. See, for example, Richard S. Brown and Ed Coughlin, The Predictive Validity of Selected Benchmark Assessments Used in the Mid-Atlantic Region (Issues & Answers Report, REL 2007-No.017), Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory, Mid-Atlantic, November 2007.

[33] For the most recent results of the PISA, see http://nces.ed.gov/surveys/pisa/.

[34] International benchmarks are independently determined for the PIRLS assessment by percentile groups (i.e., top 10%, upper quartile, median, lower quartile).

[35] For the most recent results of PIRLS, see http://nces.ed.gov/surveys/pirls/.

students.[36] Several states have started using TIMSS results in international benchmarking studies. Within the context of these studies, some states can compare student performance within their state to other countries that participated in TIMSS.[37]

The aforementioned international assessments are summative assessments used for evaluative purposes. Usually, international assessment results are reported as a simple rank ordering of countries taking the assessment. Results describe which countries scored above the "international mean" and which scored below the mean. Because the "international mean" is highly variable depending on the countries that participate from administration to administration, it is quite possible that any particular country can score above the mean on one administration and below the mean on the next administration. Furthermore, this shift can happen even when students in that country make large gains in achievement.

Linking Assessments

Students in elementary and secondary schools participate in many different assessments. Although each assessment has a unique purpose, some feel that students are required to participate in too many assessments, potentially detracting from instructional time at school. Due, in part, to the time involved in administering assessments, there has been interest in statistically linking assessments so that results can be compared across different assessments. If linkages between assessments could be established, students could participate in one assessment, and the score could potentially be used to estimate how well they would score on other assessments.

The process of linking assessments incorporates the use of statistical analyses to connect the scores from one test with those of another, regardless of the equivalence of the scales. The most common example of linking is the Fahrenheit and Celsius temperature scales. These two systems express temperature using two different scales, but they are easily linked with a simple equation. This process is fairly simple because there is a common understanding of what constitutes "temperature." The process of linking educational assessments, however, is much more complex because there is less agreement on what constitutes "reading achievement" or "mathematics achievement."

If statistical linkages could be achieved, it would allow policymakers to interpret performance across state assessments using a common metric. Comparing all states to a common metric would, in essence, allow states to be compared to each other, even if they use different assessments and different performance standards. Furthermore, linking techniques based on a common metric may allow individual states to be compared to international assessments. If linking assessments became a common practice, it could greatly reduce the number of assessments administered to students because student performance could be compared across assessments without the necessity of all students participating in the same assessment. Administering fewer assessments may reduce the testing burden and increase the amount of time schools could use for instruction.

Congress asked a committee from the National Research Council (NRC) to study the feasibility of developing a common metric to link scores from existing commercial and state assessments to

[36] For the most recent results of TIMSS, see http://nces.ed.gov/timss/.

[37] For more information on the TIMSS benchmarking studies, see http://nces.ed.gov/timss/benchmark.asp.

each other and to NAEP.[38] To determine the feasibility of linking these assessments, the committee considered the validity and practicality of making statistical linkages. After a review of studies attempting to link NAEP with state assessments and an independent evaluation, the NRC concluded that (1) comparing state assessments to each other using a common metric was not feasible, and (2) linking state assessments to NAEP is problematic and any inferences drawn from these links may be misleading. The inferences drawn from linking two tests can be adversely affected by differences in content, format, and the use of the tests being linked, as well as the consequences attached to the two tests. If two tests vary greatly along these dimensions (as many state assessments and NAEP tend to vary) the inferences drawn from the linkage may not be valid.[39]

There has been some progress, however, linking assessments in limited contexts. For example, studies have demonstrated that it may be feasible to link NAEP results for grade 8 mathematics to results on the TIMSS.[40] Statistical linking is more feasible in this case because NAEP and TIMSS use similar constructs, testing frameworks, and scoring, and they test students in the same grade. The statistical linkage between NAEP and TIMSS allows individual states to be compared to other countries in mathematics performance. This type of linkage could be particularly useful, given the recent increased interest in international benchmarks.[41]

Technical Considerations in Assessment

This section will discuss technical considerations in assessment, such as validity, reliability, and fairness. It is generally the responsibility of the test developer to investigate technical characteristics of an assessment and to 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 to 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. An inference is a conclusion that is drawn from the result of a test. Inferences may be either appropriate or inappropriate based on a number of technical and contextual factors. This section will conclude with a discussion on 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 vs. low-stakes assessments.

[38] P.L. 105-78, §306

[39] Michael J. Feuer, Paul W. Holland, and Bert F. Green, et al., *Uncommon Measures: Equivalence and Linkage Among Educational Tests* (Washington, D.C.: National Academy Press, 1999).

[40] Gary W. Phillips, *Expressing International Educational Achievement in Terms of U.S. Performance Standards: Linking NAEP Achievement Levels to TIMSS*, American Institutes for Research, Washington, DC, April 2007, http://www.air.org/news/documents/naep-timss.pdf; Gary W. Phillips, *Chance Favors the Prepared Mind: Mathematics and Science Indicators for Comparing States and Nations*, American Institutes for Research, Washington, DC, November 14, 2007, http://www.air.org/publications/documents/phillips.chance.favors.the.prepared.mind.pdf.

[41] Michele McNeil, "Panel to Spur International Benchmarks," *Education Week*, September 17, 2008.

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. Validity is the degree to which a certain inference from a test is appropriate and meaningful.[42] The question to be asked, therefore, is "Is the inference being drawn from the test result valid?" The distinction between these questions may seem unimportant, but consider the following example. Often times, teachers, administrators, or policymakers would like to support multiple conclusions from the same assessment. Some of these conclusions, or inferences, may be valid and others may not. Consider the SAT Reasoning Test, which is taken by many high school students. The SAT is a college entrance examination, and its purpose is to measure critical thinking skills that are needed for success in college. 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 possible valid inference from this result is that seniors from School A are more likely to succeed in college. There are, however, 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, one could infer 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 when a new use is proposed for an existing assessment, a process of validation should occur. 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 assessment is designed to measure 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 an assessment was designed to measure addition and subtraction 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

[42] For a thorough discussion of validity, see AERA, APA, NCME, "Standards for Educational and Psychological Testing," (Washington, DC: American Psychological Assiciation, 1999).

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 advantaged subgroups may be more likely to score well on a test than students from disadvantaged subgroups, even if both subgroups have equal knowledge of the construct itself. The construct irrelevance, therefore, may lead to an invalid inference that advantaged students outperform disadvantaged students in a given construct (in this example, addition and subtraction skills).

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. Reliability and measurement error are inversely related. The lower the measurement error, the higher the reliability. Furthermore, as reliability increases, it increases the likelihood that a student's observed score and a student's true score are reasonably equivalent.

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, which

is used as a college entrance examination, 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 the SAT. Test-retest coefficients measure the stability of an individual student's score over time. If the NAEP reading subtest 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 is 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. Internal consistency coefficients 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, and 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 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. Often times, results can 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. Under the NCLB accountability system, some states have used confidence intervals to report the results of state assessments. 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)[43], and the desired level of confidence. The confidence interval is reported as

[43] 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 (continued...)

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 hypothetical 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, a 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) and 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 very 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 or pass vs. fail), the consistency of classification is of interest.

Within school settings, consistency of classification is particularly important when using performance standards to place students in achievement levels based on state assessments (i.e., basic, proficient, advanced). For example, if the classification of students into achievement levels for AYP 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.[44]

(...continued)

from group data. The standard error of measurement is used in the calculation of confidence intervals.

[44] Daniel Koretz, "Error and Reliability: How Much We Don't Know What We're Talking About," in *Measuring Up:* (continued...)

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 "disability" and, if students are determined to be eligible, they receive special education services. Some research has begun to investigate the consistency of states' "disability" classifications over time. In an Office of Special Education Program's annual report to Congress, it was reported that approximately 17% of elementary students who received special education services in the year 2000 no longer received such services in 2002.[45] Similar results have been reported for students with disabilities in preschool.[46] While it is possible that students become "declassified" and ineligible for special education services due to their improvement in academic skills, it is likely 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.[47]

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 all members of the population should be treated equally. The notion of fairness as "equal treatment", however, 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.[48]

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 example, bias can arise when cultural or linguistic factors influence test scores of individuals within a subgroup, despite the individual's inherent ability. 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, that may be able to detect bias in specific test items, however, such techniques cannot directly address

(...continued)

What Educational Testing Really Tells Us (Cambridge, MA: Harvard University Press, 2008), pp. 143-178.

[45] U.S. Department of Education , *28th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act, 2006*, Vol. 1, Washington, DC, 2006, http://www.ed.gov/about/reports/annual/osep/2006/parts-b-c/28th-vol-1.pdf.

[46] Elaine Carlson, Tamara Daley, and Amy Shimshak, et al., *Changes in the Characteristics, Services, and Performance of Preschoolers with Disabilities from 2003-04 to 2004-05*, U.S. Department of Education, PEELS Wave 2 Overview Report, Washington, DC, June 10, 2008, http://ies.ed.gov/ncser/pdf/20083011.pdf.

[47] 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.

[48] For a comprehensive discussion of fairness in testing, see AERA, APA, NCME, "Standards for Educational and Psychological Testing," (Washington, DC: American Psychological Assiciation, 1999).

the bias in the interpretation of assessment results. Test bias, if present, undermines the validity of the inferences based on assessment results.

It is important to note that 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 students and the 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 assuring 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 LEP students.

Finally, 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 districts, it is common practice to familiarize students with sample test questions or provide examples of actual test questions from previous assessments. In other districts, 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 districts 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. On the other hand, 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 certain privileges, 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 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 to determine AYP, are aligned with school curriculum 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 AYP. 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 district.

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. Unfortunately, there is no simple checklist that will help determine if an inference is appropriate. Instead, test users must conduct a thoughtful analysis of the construct of the assessment, purpose of the assessment, the type of scores reported by the assessment, the evidence concerning the validity, reliability, and fairness of the assessment, and 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 test users may wish to ask themselves before making an inference about a test score. These guidelines are not intended to be an exhaustive list of considerations but rather a starting point for learning to draw appropriate conclusions from assessments.

Construct

Questions about the construct: What is the content area being assessed (e.g., reading, mathematics)? What is the specific construct that is being measured 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, the international assessments described above, PISA and TIMSS. Both assessments measure mathematics achievement, but they measure different mathematical constructs. PISA was designed to measure basic "mathematical literacy" whereas TIMSS is curriculum-based and was designed to measure what students have learned in school. Results from the 2006 PISA administration reported that the average U.S. score in mathematics was lower than international average. Results from the 2007 administration of the

TIMSS reported that the average U.S. score in mathematics was higher than the international average. Based on these results, a novice test user may be tempted to conclude that within one year, students in the U.S. improved in mathematics achievement compared to other countries. There are several reasons why this conclusion is inappropriate, one of which is that PISA and TIMSS measure very different constructs.[49]

Purpose

Questions about the purpose: 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 will 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 strictly 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, a benchmark 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. Benchmark assessments are often aligned with state assessments, however, scores on benchmark assessments should not be considered definitive indicators of what state assessment scores will be.

Scores

Questions about scores: 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, performance standard)? Does the score determine whether a student is proficient within a certain content area (e.g., performance standards)? Does the score show growth or progress that a student made within a content area?

Misinterpreting scores is perhaps the most common way to make an inappropriate inference. To avoid an inappropriate inference, a test user should 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 criterion-referenced scores or performance standards, a student's score can be interpreted relative to a predetermined standard or criterion. CRTs and performance standards, however, were not designed to make particularly meaningful comparisons between students who participated in the same assessment.

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

Because of the use of performance standards in state assessments, it is particularly 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 both a score and some meaningful description of what students know. Performance standards, however, can be particularly difficult to interpret. Students are classified into categories, such as basic, proficient, or advanced, based on their performance on an assessment. All students within the "proficient" category, however, did not score equally well. Furthermore, scores from performance standards do not lend themselves to interpret a student's growth. A student can score at the lower end of the proficient category, make considerable progress over the next year, and still be in the proficient category at the end of the year. Alternatively, a student could score at the high end of the basic category, make minimal progress over the next year, and move up into the proficient category. Because of these qualities of performance standards, test users should be very cautious equating the performance of students within the same category and making assumptions about growth based on movement through the categories.

Technical Quality

Questions about technical quality: Did the test developers provide statistical information on the validity and reliability of the instrument? Was the issue of fairness and bias addressed, either through thoughtful reasoning or statistical procedures? 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?

Commercially available assessments are 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 should 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.[50] In light of the current uses of assessments in schools, however, some issues are noteworthy.

First, because schools are required to report state assessment results for various subgroups (i.e., students with disabilities and LEP students), 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.

Second, 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 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 Reasoning Test which 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

[50] 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 Assiciation, 1999).

may have scores that are positively biased and students taking the harder version may have scores that are negatively biased.

Third, 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

Questions about the context: Is this 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 about 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 a worthwhile endeavor. These assessments are usually short, conducted to help teachers adapt their instruction, and have no 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.

Sometimes when making an inference for a high-stakes decision, certain protections are placed on the testing process or the test result. For example, in terms of a state exit exam for high school graduation, some states allow students to take the assessment multiple times to lessen 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. In other high-stakes assessments, such as state assessments for AYP, some states use confidence intervals in addition to observed scores to report student achievement. Several states have chosen to use 95% or even 99% confidence intervals to increase the certainty of inferences based on test scores.

Use of Assessments in Accountability Systems: Implications for Curriculum, Students, and Testing

NCLB greatly increased the emphasis on student assessment. Under NCLB, student scores on state assessments are used as key indicators in an accountability system that determines whether schools are making progress with respect to student achievement. Some have viewed this shift towards test-based accountability as a positive move because it places more emphasis on developing rigorous content standards in reading, mathematics, and science and teaching to the standards. Test-based accountability as implemented by NCLB also leads to increased attention on traditionally underperforming subgroups of students, including disadvantaged students,

students with disabilities, and LEP students. On the other hand, test-based accountability has been criticized for narrowing the curriculum and focusing all instruction on the tested subjects of reading and mathematics at the expense of other subjects. The current practice of test-based accountability may also create incentives to set low expectations for proficiency and to focus on a subset of children who are near the proficiency level instead of focusing on children at all achievement levels. Another criticism is that the increased emphasis on test-based accountability can lead to score inflation, which may inhibit policymakers from measuring the actual impact of accountability.

In the sections below, the potential positive and negative implications of test-based accountability are discussed, including implications for curriculum, students, and testing. It is important to note that the issues discussed below are specific to current test-based accountability systems under NCLB; however, all test-based accountability systems may not have the same positive and negative implications.

Implications for Curriculum

One potentially positive outcome of test-based accountability has been an increased focus on state-level content standards and teaching to those standards. Under NCLB, states have been required to develop rigorous content standards for reading and mathematics, which many see as core subjects that will lead to improved learning in other content areas. There is some evidence that using content standards and assessments can help teachers focus their instruction and obtain feedback on the effectiveness of their instruction.[51] There are also some data to suggest that high-performing schools have a stronger alignment between state content standards and school curriculum. Furthermore, schools that include teachers in the development of these standards tended to have a higher degree of teacher "buy-in" to the standards.[52]

On the other hand, test-based accountability may be affecting the curriculum in less desirable ways. One criticism of test-based accountability systems is that they lead to a narrowing of the curriculum. There are several ways in which these systems might narrow the curriculum. First, the time spent administering the actual assessments, sometimes called the "testing burden," could detract from classroom instruction. Second, test-based accountability systems may lead to increases in test preparation, leaving less time for instruction. There is some evidence to suggest that teachers feel pressure to "teach to the test" and engage in test preparation activities at the expense of instruction. Test-preparation activities take several forms, including altering typical classroom assignments to conform to the format of an expected response on the state assessment (e.g., if the state assessment requires a five paragraph constructed response, teachers may assign a disproportionate number of five-paragraph essays).[53] Third, in test-based accountability systems, teachers report reallocating instructional time towards tested subjects and away from non-tested subjects. Surveys of teachers have consistently reported that their instruction emphasizes reading

[51] L. Mabry, J. Poole, and L. Redmond, et al., "Local Impact of State Testing in Southwest Washington," *Education Policy Analysis Archives*, vol. 11, no. 22 (July 18, 2003), http://epaa.asu.edu/epaa/v11n22/.

[52] Deepa Srikantaiah, Ying Zhang, and Lisa Swayhoover, *Lessons from the Classroom Level: Federal and State Accountability in Rhode Island*, Center on Education Policy, November 25, 2008, http://www.cep-dc.org/document/docWindow.cfm?fuseaction=document.viewDocument&documentid=249&documentFormatId=3846.

[53] Deepa Srikantaiah, Ying Zhang, and Lisa Swayhoover, *Lessons from the Classroom Level: Federal and State Accountability in Rhode Island*, Center on Education Policy, November 25, 2008, http://www.cep-dc.org/document/docWindow.cfm?fuseaction=document.viewDocument&documentid=249&documentFormatId=3846.

and mathematics over other subjects like history, foreign language, and arts.[54] Although there is consistency in the survey results, it is difficult to understand the extent to which instructional time is reallocated towards tested subjects. Most of these reports do not collect data through experimental observation, but rather they rely on self-reported data from teachers and administrators, which is often less reliable.

Implications for Students

NCLB has required states to disaggregate student assessment data for major subgroups, including racial/ethnic groups, economically disadvantaged students, students with disabilities, and LEP students. Under NCLB, schools are held accountable for the performance of each of these subgroups, and each subgroup shares a common goal of reaching 100% proficiency in reading and mathematics by 2014. Designing an accountability system in this way has increased the attention given to the achievement of certain subgroups that may have been previously masked by overall student performance. In general, disaggregating data by subgroups has been seen as a positive step in terms of equity in education because the performance of all subgroups "counts" towards AYP. Supporters of disaggregation believe that it leads to increased access to rigorous academic curriculum for students who otherwise may not have had access to such curriculum due to low expectations of performance.

Along with the increased attention to subgroups of students, there has been increased attention to the achievement gaps between white students and minority students and between economically advantaged students and disadvantaged students. Over the last several decades, a general goal of public education has been to "close the achievement gap," and thus, improve equity in education. By disaggregating assessment results, NCLB has led to consistent measurement of the achievement gap and allows researchers to examine the size of the achievement gap over time.

One of the unintended consequences of NCLB accountability is the way instruction may be focused on students just below the "proficient" level, possibly at the expense of other students.[55] Under NCLB's test-based accountability system, the goal is for 100% of students to reach proficiency by 2014. In an effort to raise the percentage of proficient students, schools and teachers may target instructional time and resources towards those students who are near proficiency. Since time and resources are a zero-sum game, fewer instructional resources may be available for students who are far below proficiency or even those who achieve at advanced levels. This disincentive to focus instructional resources on all children has led to possible alternative methods of measuring achievement, including growth models.[56] Within certain

[54] Laura Hamilton, "Assessment as a Policy Tool," *Review of Research in Education*, vol. 27 (2003), pp. 25-68; Laura S. Hamilton and Mark Berends, *Instructional Practices Related to Standards and Assessment*, RAND, WR-374-EDU, Washington, DC, April 2006, http://www.rand.org/pubs/working_papers/2006/RAND_WR374.pdf; Patricia Velde Pederson, "What is Measured Is Treasured: The Impact of the No Child Left Behind Act on Nonassessed Subjects," *Clearing House: A Journal of Educational Strategies, Issues and Ideas*, vol. 80, no. 6 (July/August 2007), pp. 287-291; Jennifer McMurrer, *Choices, Changes, and Challenges: Curriculum and Instruction in the NCLB Era*, Center on Education Policy, December 2007, http://www.cep-dc.org/_data/n_0001/resources/live/07107%20Curriculum-WEB%20FINAL%207%2031%2007.pdf.

[55] Jennifer Booher-Jennings, "Below the Bubble: "Educational Triage" and teh Texas Accountability System," *American Educational Research Journal*, vol. 42, no. 2 (Summer 2005), pp. 231-268; Laura S. Hamilton and Mark Berends, *Instructional Practices Related to Standards and Assessment*, RAND, WR-374-EDU, Washington, DC, April 2006, http://www.rand.org/pubs/working_papers/2006/RAND_WR374.pdf.

[56] See CRS Report RL33032, *Adequate Yearly Progress (AYP): Growth Models Under the No Child Left Behind Act*, (continued...)

accountability systems, the use of growth models may give teachers and schools credit for student growth, even if the growth occurs far below the proficiency level.

Implications for Testing

Test-based accountability systems use high-stakes assessments to make decisions about students, teachers, and schools. Under NCLB, individual schools are held accountable for student achievement, and if schools fail to meet their AYP goals, there are consequences. In an effort to avoid these consequences, schools often make conscious efforts to prepare students for high-stakes assessments. Although these efforts are often undertaken with good intentions, they can lead to score inflation. Score inflation is a phenomenon in which scores on high-stakes assessments tend to increase at a faster rate than scores on low-stakes assessments. The validity of an inference is greatly reduced when score inflation is present.

Test preparation can take many forms, and it is difficult to distinguish appropriate test preparation from inappropriate test preparation. Many schools provide test preparation to young students who have little experience with standardized testing, and this form of test preparation can actually increase the validity of a test score because it is less likely that students will do poorly due to unfamiliarity with the testing process. Other test preparation strategies, such as working more effectively or working harder, are also usually desirable. Test preparation begins to affect validity in a negative way, however, when there are excessive amounts of alignment between test items and curriculum, excessive coaching of a particular type of item that will appear on the test, or even outright cheating.

Studying the prevalence of score inflation is difficult because school districts may be reluctant to give researchers access to test scores for the purpose of investigating possible inflation. Nevertheless, several studies have documented the problem of score inflation by comparing gains on state assessments (high-stakes) to those made on NAEP (low-stakes).[57] Studies have consistently reported discrepancies in the overall level of student achievement, the size of student achievement gains, and the size of the achievement gap. The discrepancies indicate that student scores on state assessments may be inflated and that these inflated scores may not represent true achievement gains as measured by another test of a similar construct. In this case, the validity of the inference made from state assessments may be questioned.

One possible way to reduce the problem of score inflation is to consistently use a low-stakes "audit" assessment, such as NAEP, to corroborate gains on state assessments.[58] If gains on state assessments generalize to another "audit" assessment, it increases the likelihood that gains are

(...continued)

by (name redacted).

[57] B. Fuller, K. Gesicki, and E. Kang, et al., *Is the No Child Left Behind Act Working? The Reliability of How States Track Achievement*, University of California, Berkeley PACE, Working Paper 06-1, Berkeley, CA, 2006; S.P. Klein, Linda S. Hamilton, and Daniel F. McCaffrey, et al., *What Do Test Scores in Texas Tell Us?*, RAND, Santa Monica, CA, 2000; Daniel Koretz and S. I. Barron, *The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS)*, RAND, MR-792-PCT/FF, Santa Monica, CA, 1998; Robert L. Linn and C. Haug, "Stability of School-building Accountability Scores and Gains," *Educational Evaluation and Policy Analysis*, vol. 24, no. 1 (2002), pp. 29-36.

[58] Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, MA: Harvard University Press, 2008), pp. 247-248.

due to true achievement gains. This type of corroboration may help policymakers separate the policies that lead to true student achievement from those that lead to score inflation.

Appendix A. 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.
benchmark assessment	A type of interim assessment that is either commercially developed or created by school districts for the purpose of predicting the likelihood that students will meet a predetermined future goal, such as passing the annual state assessment. It can be used as either a formative or summative assessment, depending on the timing of the test and how the results are used by teachers.
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 curriculum 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 on an entire content area based on a sample of test items from that content area.
inference	A meaningful conclusion based on the results of an assessment.
inter-scorer agreement	A reliability statistic that measures the degree to which two 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.
score inflation	A phenomenon in which scores on high-stakes assessments tend to increase at a faster rate than scores on low-stakes assessments. Score inflation can be influenced by both positive factors (such as working harder or teaching more efficiently) or negative factors (such as excessive test preparation or cheating).
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	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. They are used for evaluative purposes.
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 true score is 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 entailed by 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.

Appendix B. Acronym Reference

AYP	Adequate Yearly Progress
CRT	Criterion-referenced Test
ED	U.S. Department of Education
ESEA	Elementary and Secondary Education Act
IDEA	Individuals with Disabilities Education Act
IEP	Individualized Education Program
LEA	Local Educational Agency
LEP	Limited English Proficiency
NAEP	National Assessment of Educational Progress
NAGB	National Assessment Governing Board
NCLB	No Child Left Behind
NRT	Norm-referenced Test
PIRLS	Progress in International Reading Literacy Study
PISA	Program for International Student Assessment
SEA	State Educational Agency
TIMSS	Trends in International Mathematics and Science Study

Author Contact Information

(name redacted)
Specialist in Education Policy
/redacted/@crs.loc.gov, 7-....