STATE ASSESSMENT SYSTEMS

CONTEXT

Standards-based accountability is a widely accepted framework for public education. Every state now has education standards, although the current array of accountability approaches is characterized by significant variation, as Diana Pullin noted. Content and performance standards range widely in rigor, as does the implementation of test-based accountability (see National Research Council, 2008). By design, assessments play a key role in standards-based accountability, and because standards are not working exactly as they were intended to, Pullin suggested, “assessments can be more powerful in driving what happens in schools than standards themselves.” Recent research has indicated that the influence of assessments on curriculum and instruction has increased since 2001, when NCLB was passed, and that tests themselves have changed in significant ways (see, e.g., Jennings and Rentner, 2006; McMurrer, 2007; Lai and Waltman, 2008; Sunderman, 2008). Several presenters provided perspectives on the history and current status of assessment and accountability systems.

The idea that assessments should be used to evaluate not only individual students’ progress, but also the quality of instruction and the performance of educators more generally, is one with longstanding roots, Joan Herman noted. Edward Thorndike, who published pioneering books on educational measurement in the first decades of the 20th century, viewed his work as useful in part because it would provide principals and teachers with a tool for improving student learning. Ralph Tyler, known for innovative work in educational evaluation in the 1940s, posed the idea that objectives ought to drive curriculum and instruction and that new kinds of assessments (beyond paper-and-pencil tests) were needed to transform learning and the nature of educational programs. Other contributions to thinking about evaluation include Benjamin Bloom’s 1956 taxonomy of educational objectives, the development of criterion-referenced testing in the 1950s, mastery learning in the 1960s and 1970s, minimum competency testing in the 1970s and 1980s, and performance assessment in the 1990s. All of these, Herman suggested, have been good ideas, but they have not had the effects that had been hoped for.

Changes in Tests

Most recently, as Scott Marion detailed, NCLB has had a very clear impact on many aspects of the system. Prior to 2002, for example, states were required to test at one grade each in the elementary, middle, and high school years. NCLB required testing in grades 3 through 8 as well as in high school. Marion argued that this increased testing resulted in improvements in state standards. The new requirement compelled states to define coherent content standards by grade level, rather than by grade span, and to articulate more precisely what the performance standards should be for each grade. Testing at every grade has
INTRODUCTION

also opened up new possibilities for measuring student achievement over time, such as value-added modeling.\(^3\)

The design of states’ tests has also been affected, most notably in a dramatic reduction in the use of matrix sampling designs because they do not provide data on individual students. For a long time, many test designers used matrix sampling to produce data about the learning of large groups of students (such as all children in a single grade) across a broad domain. With this sampling approach, tests are designed so that no one student answers every question (which would require too much testing time for complete coverage), but, taken together, different students’ responses to different questions support inferences about how well the group as a whole has learned each aspect of the domain tested. One advantage of matrix sampling is that each student takes fewer test items—because student-level scores are not produced—which allowed developers to include more complex item types. This approach makes it possible to better assess a broad academic domain because the inclusion of more complex item types is likely to yield more generalizable inferences about students’ knowledge and skills. However, reporting individual student scores on matrix-samples content is problematic because different students have responded to different questions. Thus, because states are required under NCLB to report results for individual students, matrix sampling is much less commonly used than it had been.

The types of test questions commonly used have also changed, Marion observed, with developers relying far less on complex performance assessments and more on multiple-choice items. He cited evidence from the Government Accounting Office (2009) that the balance between multiple-choice and open-ended items (a category that includes short constructed-response items) has shifted significantly in favor of the multiple-choice format as states have responded to the NCLB requirements. Many states still use items that could be described as open ended, but use of this term disguises important differences between a short constructed-response item worth a few points and an extended, complex performance task that contributes significantly to the overall score. To illustrate the difference, he showed sample items—a four-page task from a 1996 Connecticut test that called for group collaboration and included 16 pages of source materials, contrasted with mathematics items from a 2009 Massachusetts test that asked students to measure an angle and record their result or to construct a math problem and solve it.

Marion was not suggesting that the shorter items—or others like them—are necessarily of inferior quality. Nevertheless, he views this shift as reflecting an increased focus on breadth at the expense of depth. The nature of state assessments under NCLB signals that the most important goal for teachers is

\(^3\)“Value-added modeling” is statistical analysis in which student data that are collected over time are used to measure the effects of teachers or schools on student learning.
to ensure that students have an opportunity to learn a broad array of content and skills, even if the coverage is superficial. It is important to ask, if this is the case, whether the types of processes students use to solve multiple-choice items are truly the basis for the 21st century “college- and work-ready” skills that policy makers stress. For example, Marion pointed out, the 1996 Connecticut extended item begins by asking the students to break into small groups and discuss their approach to the task—a challenge much closer to what is expected in many kinds of work than those that are posed by most test items.

States have also changed their approaches to high school testing, though the situation is still in flux. There has been a modest increase in the number of states using end-of-course examinations (given whenever students complete a particular course)—instead of survey tests that all students take at a certain point (such as at the end of particular grades). A few states have begun using college entrance tests as part of their graduation requirements.

**Interim Assessments**

Another development, discussed by Margaret Goertz, has been a marked increase in the use of interim assessments, particularly at the district level. These assessments measure students’ knowledge of the same broad curricular goals that are measured in annual large-scale assessments, but they are given more frequently and are designed to give teachers more data on student performance to use for instructional planning. Interim assessments are often explicitly designed to mimic the format and coverage of state tests. They may be used not only to guide instruction, but also to predict student performance on summative state assessments, to provide data on a program or approach, or to provide diagnostic information about a particular student. Researchers stress the distinction between interim assessments and formative assessments, however, because the latter are typically embedded in instructional activities and may not even be recognizable as assessments by students (Perie and Gong, 2007).

Districts have vastly increased their use of interim assessments in the past 10 years, Goertz noted (see Stecher et al., 2008), and draft guidelines for the Race to the Top initiative encouraged school districts to develop formative or interim assessments as part of comprehensive state assessment systems. However, there have been very few studies of how interim assessments are actually used by individual teachers in classrooms, by principals, and by districts, or of

---

4Formative assessments are those designed primarily to provide information that students can use to understand the progress of their learning and that teachers can use to identify areas in which students need additional work. Formative assessments are commonly contrasted with summative assessments, those designed to measure the knowledge and skills students have attained by a particular time, usually after the instruction is complete, for the purpose of reporting on progress. Interim assessments are assessments, which may be formative or summative, that are given at intervals to monitor student progress.
their effects on student achievement, perhaps in part because these are relatively new tools. Moreover, Goertz pointed out, many of the studies that are cited in their favor were actually focused on formative assessments. Moreover, she said, studies are needed to provide technical and other validity evidence to support inferences made from interim assessments.

In surveys, teachers have reported that the results of interim assessments helped them monitor student progress and identify skill gaps for their students and led them to modify curriculum and instruction (Clune and White, 2008; Stecher et al., 2008; Christman et al., 2009). Goertz noted that a study of how teachers used curriculum-based interim assessments in elementary mathematics in two districts showed that teachers did use the data to identify weak areas or struggling students and to make instructional decisions (Goertz, Olah, and Riggan, 2009). The study also showed that teachers varied in their capacity to interpret interim assessment data and to use it to modify their teaching. The study also found that few of the items in the interim assessments provided information that teachers could readily use, and few actually changed their practice even as they retaught material that was flagged by the assessment results. For example, many teachers focused on procedural rather than conceptual sources of error.

Marion noted that the limited research available provides little guidance for developing specifications for interim assessments or for support and training that would help teachers use them to improve student learning. There is tremendous variability in the assessments used in this way, and there is essentially no oversight of their quality, he noted. He suggested that interim assessments provide fast results and seem to offer jurisdictions eager to respond to the accountability imperative in an easy way to raise test scores.

**Multiple Purposes for Assessment**

Another significant change, Goertz pointed out, is that as demands on state-level assessments have increased in a time of tight assessment budgets, tests are increasingly being used for a number of different purposes. Table 1-1 shows some common testing purposes by goal and by the focus of the information collected. In practice, the uses may overlap, but the table illustrates the complexity of the role that assessments play.

To clarify the implications of putting a single test to multiple uses, Goertz highlighted the design characteristics that are most important for two of these uses, informing instruction and learning and external accountability. For informing instruction and learning, a test should be designed to provide teachers with information about student learning on an ongoing basis, which they can easily interpret and use to improve their instruction. For this purpose, an assessment needs to provide information that is directly relevant to classroom instruction and is available soon after the assessment is given. Ideally, this kind of assess-
ment would provide continuous information: if it is embedded in instruction it can provide continuous feedback. For this purpose, statistical reliability is not as important as relevance and timeliness. In contrast, when test data are to be used for external accountability, the assumption is that information about performance will motivate educators to teach well and students to perform to high standards. Therefore, incentives and sanctions based on test results are often used to stimulate action, which means that the tests have stakes for both students and educators. So, when accountability is the goal, several test characteristics are of critical importance: alignment of test questions to standards; standardization of the content, test administration, and scoring to support fair comparisons; and the fairness, validity, and reliability of the test itself.

The tension between these two purposes is at the heart of many of the problems that states have faced with their assessment programs, and it is a key challenge for policy makers to consider as they weigh improvements to accountability systems.

The growing tendency to use assessments for multiple purposes may be explained in part by the loss of assessment staff in a time of tight education budgets. Marion reported that most states have seen an approximately three-fold increase in testing requirements without a corresponding increase in personnel (Government Accounting Office, 2003; Toch, 2006). As a result, many states have moved from internal test design and development to outside

---

**TABLE 1-1** Uses of Assessment

<table>
<thead>
<tr>
<th>Use</th>
<th>Student</th>
<th>Teacher</th>
<th>School</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>Instructional decisions; placement; allocation of educational services</td>
<td>Professional development and support</td>
<td>Resource allocation; technical assistance</td>
</tr>
<tr>
<td>Inform Teaching and Learning</td>
<td></td>
<td>Focus, align, redirect content; instructional strategies</td>
<td>Instructional focus, align curriculum to skills or content; school improvement and planning</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Certification of individual achievement</td>
<td>Teacher preparation programs; teacher pay</td>
<td>Program evaluation</td>
</tr>
<tr>
<td>Public Reporting</td>
<td>Transcripts</td>
<td></td>
<td>Parent or community action</td>
</tr>
<tr>
<td>External Accountability</td>
<td>Promotion; high school graduation</td>
<td>Renewal; tenure; pay</td>
<td>Sanctions and rewards</td>
</tr>
</tbody>
</table>

SOURCE: Goertz (2009, p. 3).
INTRODUCTION

vendors, and, he suggested, remaining staff have less time to work with vendors and to think about innovative approaches to testing.

A number of other factors help explain recent changes in the system, Marion suggested. NCLB required rapid results, and the “adequate yearly progress” formula put a premium on a “head-counting” methodology (measuring how many students meet a particular benchmark by a particular time), rather than considering broader questions about how well students are learning. However, the law did not, in his view, provide adequate financial support for ongoing operational costs. He also said that there has been insufficient oversight of technical quality, so that the validity of particular assessments for particular purposes has received inadequate attention. Marion also noted that because results for multiple-choice and open-ended items are well correlated, many people mistakenly believe that this is evidence that they are interchangeable. An era of tight funding has made it easy for policy makers to conclude that open-ended items and other innovative approaches are not worth their higher cost, he said, without necessarily understanding that such assessment methods make it possible to assess skills and content that cannot be assessed with multiple-choice items.

THE CURRENT SYSTEM

This outline of some of the important recent changes in assessment systems provided the foundation for a discussion of strengths and weaknesses of the current system and targets for improvement. Goertz and Marion presented very similar messages, which were seconded by discussant Joan Herman.

Strengths

Attention to Traditionally Underserved Student Populations Including all students in assessments—to ensure that schools, districts, and states are accountable for their results with every group—was a principal goal of NCLB. As a result, although much work still needs to be done, assessment experts have made important progress in addressing the psychometric challenges of testing students with disabilities and English language learners. Progress has been made in understanding the validity of assessments for both of these groups, which are themselves very heterogeneous. Test designers have paid attention to the challenges of more explicitly specifying the constructs to be measured and removing construct irrelevant variance from test items (e.g., by reducing the reading burden in tests of mathematics so that the measure of students’ mathematics skills will not be distorted by reading disabilities). As policy makers and psychometricians have worked to strike an appropriate balance between standardization and technical quality, more assessments are available to measure academic skills—not just functional skills—for special populations. Improved
understanding of patterns of language acquisition and the relationship between language skills and academic proficiency have supported the development of better tools for assessing English-language learners across many domains.

**Increased Availability of Assessment Data** The premise of NCLB is that if states and districts had more data to document their students’ mastery of educational objectives, they would use that information to improve curricula and instructional planning. States and districts have indeed demonstrated a growing sophistication in the analysis and use of data. Improved technology has made it easier for educators and policy makers to have access to data and to use them, and more people are using them. However, the capacity to draw sound inferences from the copious data to which most policy makers and educators now have access depends on their training. As discussed below, this capacity has in many cases lagged behind the technology for collecting data.

**Improved Reporting** The combination of stricter reporting requirements under NCLB and improved technology has led states and districts to pay more attention to their reporting systems since 2002. Some have made marked improvements in presenting data in ways that are easy for users to understand and use to make effective decisions.\(^5\)

**Weaknesses**

**Greater Reliance on Multiple-Choice Tests** In comparison with the assessments of the 1990s, today’s state assessments are less likely to measure complex learning. Multiple-choice and short constructed-response items that are machine scorable predominate. Though valuable, these item types assess only a limited portion of the knowledge and skills that are called for in current standards.

**More Focus on Tested Content Than on Standards** Particularly in low-performing schools, test-based accountability has focused attention on standards, especially the subset of academic standards and content domain that is covered by the tests. Although this focus has had some positive effects, it has also had negative ones. States and districts have narrowed their curricula, placing the highest priority on tested subjects and on the content in those subjects that is covered on tests. The result has been emphasis on lower-level knowledge and skills and very thin alignment with the standards: for example, Porter, Polikoff, and Smithson (2009) found very low to moderate alignment between state assessments and standards—meaning that large proportions of content standards are not covered on the assessments (see also Fuller et al., 2006; Ho, 2008).

\(^5\)Marion cited the Colorado Department of Education’s website as a good example of innovative data reporting (see http://www.schoolview.org/ [accessed January 2010]).
INTRODUCTION

More Narrow Test Preparation Because of the considerable pressure to make sure students meet minimum requirements on state assessments, many observers have noted an increased focus on so-called “bubble kids,” those who are scoring just below cutoff points. A focus on drilling these students to get them above the passing level may often come at the expense of other kinds of instruction that may be more valuable in the long run. Discussants suggested that this focus on test preparation is particularly prevalent in schools serving poor and traditionally low-performing students; and the emerging result is a dual curriculum, in which already underserved children are not benefiting from the rigorous curriculum that is the ostensible goal of accountability (see, e.g., McMurrey, 2007). This approach often results in less attention to the needs of both high- and low-performing students.

Insufficient Rigor Many researchers and analysts regard current state assessments as insufficiently rigorous. Analysis of their cognitive demand suggests that they tend to focus on the lower levels of cognitive demands as defined in state standards and that they are less difficult than, for example, NAEP (see, e.g., Ho, 2008; Cronin et al., 2009). In general, the multiple-choice and short-answer items on which many state tests rely heavily are most frequently used to assess the recall of factual knowledge (rather than to assess students’ abilities to synthesize or analyze knowledge, for example), and basic skills (rather than more complex thinking skills).

Challenges

Many of the challenges that presenters and discussants identified as most pressing mirrored the strengths and weaknesses. They identified opportunities not only to address the weaknesses, but also to build on many of the strengths in the current system.

Purposes of Testing For Goertz, any plans for improving assessment and accountability systems should begin with clear thinking about several questions: “What do we want to test and for what purpose? What kinds of information do we want to generate and for whom? What is the role of a state test in a comprehensive assessment system? What supports will educators need?” Assessments, Goetz said, whatever their nature, communicate goals to students and teachers. They signal what is valued, in terms of the content of curriculum and instruction and in terms of types of learning. Everyone in the system listens to the signal that testing sends, and they respond accordingly. Goertz suggested that current approaches may be coherent, in a sense, but many assessments are sending the wrong signals to students and teachers, because insufficient attention has been given to varying purposes for which they are actually being used.
**The System as a Whole** If tests are bearing too much weight in the current system, several participants said it is logical to ask whether every element of an accountability system must be represented by a test score. Measures of students’ opportunity to learn and student engagement, as well as descriptive measures of the quality of the teaching and learning, may provide a valuable counterbalance to the influence of multiple-choice testing. It is important to balance the need for external accountability against other important goals for education.

Thus, in different ways, Marion, Goertz, and Herman each suggested that it is important to evaluate the validity of the entire system, seeking evidence that each element of the system serves its intended purpose. The goal for an accountability system should be to provide appropriate evidence for all intended users and to ensure that those users have the capacity and resources to use the information. The key question, Herman said, is clear: “Can we engineer the process well enough that we minimize the negative and maximize the positive consequences?” Clearly, she stressed, it does not make sense to rely on an annual assessment to provide all the right data for every user—or to measure the breadth and depth of the standards.

Looking at the system as a whole will entail not only consideration of intended and unintended consequences of assessments, but also a clear focus on the capacity of each element of the system to function as intended. However, Goertz pointed out, a more innovative assessment system—one that measures the most important instructional goals—cannot by itself bring about the changes that are desired. Support for the types of curriculum and instruction that foster learning, as well as such critical elements as teacher quality, is also needed.

**Staff Capacity** Many workshop participants spoke about the importance of developing a “culture of data use.” Even as much more data has become available, insufficient attention has been paid to developing teachers’ and administrators’ capacity to interpret it accurately and use it to support their decision making. Ideally, a user-friendly information management system will focus teachers’ attention on the content of assessment results so they can easily make correct inferences (e.g., diagnose student errors) and connect the evidence to specific instructional approaches and strategies. Teachers would have both the time to reteach content and skills students have not mastered and the knowledge of effective strategies to target the gaps.

**System Capacity** Looking more broadly at the capacity issue, Marion noted again that here has been a “three- or four-fold increase in the number of tests that are given without any corresponding increase in assessment personnel.” Yet performance or other kinds of innovative assessments require more person-hours at most stages of the process than do multiple-choice assessments. These issues were discussed in the next session of the workshop, described in Chapter 2.
INTRODUCTION

Reporting of Results  Although there have been improvements in reporting, it has generally received the least attention of any aspect of the assessment system. NCLB has specific reporting requirements, and many jurisdictions have better data systems and better technology as a result. Nevertheless, even the best reports are still constrained by the quality of the data and the capacity of the users to turn these data into information, decisions, and actions.