Testing What Has Been Taught

Helpful, High-Quality Assessments Start with a Strong Curriculum

By Laura S. Hamilton

In recent years, standardized, large-scale tests of student achievement have been given a central role in federal, state, and local efforts to improve K–12 education. Despite the widespread enthusiasm for assessment-based reforms, many of the current and proposed uses of large-scale assessments are based on unverified assumptions about the extent to which they will actually lead to improved teaching and learning, and insufficient attention has been paid to the characteristics of assessment programs that are likely to promote desired outcomes. Moreover, advocates of assessment-based reform often hold unrealistic expectations for what these assessments can and cannot do.

In light of the recently developed Common Core State Standards and the ongoing work to develop assessments aligned to those standards, now is a good time to pause and consider our state and federal assessment policies. If we are to actually improve schools, researchers and policymakers must address a few essential questions: How many purposes can one assessment serve? Can assessments meaningfully be aligned to standards, or is something more detailed, like a curriculum, necessary to guide both teachers and assessment developers? What would the key features of an assessment system designed to increase student learning and improve instruction be? While current assessment knowledge is not sufficient to fully answer these questions, in this article I offer an overview of what is known and several suggestions for improving our approach to assessment.

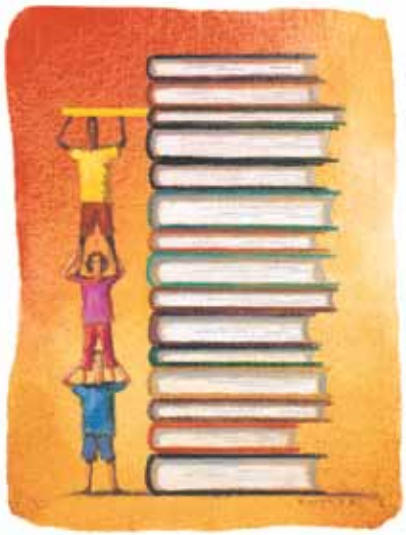
Purposes of Assessment

Large-scale assessments of student achievement are currently being used to serve a number of purposes in K–12 education. Broadly speaking, these purposes can be described as focusing

Laura S. Hamilton is a senior behavioral scientist with the RAND Corporation and an adjunct associate professor in the University of Pittsburgh's Learning Sciences and Policy program. She has directed several large studies, including an investigation of the implementation of standards-based accountability in response to No Child Left Behind. She is currently working with the National Center on Performance Incentives to investigate teachers' responses to pay-for-performance programs, and she serves on the committee that is revising the Standards for Educational and Psychological Testing.

ILLUSTRATIONS BY PAUL ZWOLAK

on providing information, imposing accountability, or some combination of the two. Increasingly, policymakers and others are placing multiple demands on large-scale testing programs to serve a wide variety of information and accountability purposes, and to inform decision making and induce change at different levels of the education system. Unfortunately, tests are seldom designed to address multiple purposes at once. Policymakers and the public must recognize that when a test designed for one purpose (e.g., to identify students' strengths and weaknesses in algebra) is used for another purpose (e.g., to decide which students will be promoted to ninth grade or which teachers will receive bonuses), the resulting test scores may not provide valid information for both purposes. The use of the test to make decisions for purposes other than those for which it was validated is generally unwarranted.[1]



Research indicates that teachers and other staff reallocate time and resources toward tested content.

Efforts to validate large-scale assessments are not able to keep pace with the public policies expanding their use. Though many policymakers are not heeding researchers' warnings, there is evidence that most such assessments may not be serving *any* of their purposes adequately. At the classroom level, teachers tend to find that most accountability-focused tests are less useful than other information (such as homework, teacher-developed tests, or classroom observations) for informing instruction. In addition, the attachment of high stakes to existing tests has led to unintended and probably undesirable consequences (discussed below).

The Effects of High-Stakes Testing

Because much of today's policy debate focuses on externally mandated assessments for use as tools of accountability, we can apply lessons learned from the past few decades, when accountability testing became nearly ubiquitous in public K–12 education. In brief, research (conducted by various individuals and organizations across numerous districts, states, and nations) indicates that teachers and other school and district staff reallocate resources (including time) toward tested content and away from untested content.[2] This reallocation occurs across subjects, across topics within subjects, and even across students when the performance of some students counts more than that of others for accountability purposes (e.g., some schools have provided extra help to students just below the cut score for proficient).[3]

The form of resource reallocation that has probably generated the most concern is the excessive emphasis on test-taking skills; it consumes time that should be spent teaching content. However, this is not the only form, and may not even be the most common. Reallocation also takes the form of increases in time spent engaging in instructional activities that are directed toward what is tested and how it is tested—such as focusing on short reading passages with closed-ended comprehension questions—and decreases in time spent on activities that are not tested—such as reading novels or writing extended essays. Because most large-scale tests rely on multiple-choice items or other formats that tend to emphasize discrete skills and knowledge rather than complex, extended problems, reallocation is likely to reduce the amount of class time and resources devoted to these more complex skills and processes.*

Reallocation is often thought of as something teachers do, but the decisions that lead to reallocation are often made at higher levels of the education system. Teachers report drawing on a variety of instructional resources (such as curriculum and pacing guides, test-preparation materials, professional development, and mandatory interim assessments), and school, district, and state administrators often design these resources to emphasize tested content.[5] Worse, these resources are not always well aligned or designed in ways that promote high-quality instruction. For example, while some teachers have access to high-quality formative assessment systems that are linked to their local curricula and provide clear guidance for next steps, others obtain their interim data from mandatory assessments that do not provide formative feedback and may not be well aligned with what they are teaching.

The key lesson of all this research is that *what is tested influences what is taught, in significant and sometimes unexpected, problematic ways*. For example, one well-documented problem is score inflation. Scores on high-stakes tests tend to increase much more rapidly than scores on low- or no-stakes tests, as educators alter their instruction to better prepare students for the high-stakes test. Some of these score increases are legitimate and welcomed; some are the result of anything from drilling in test-taking strategies to outright cheating. The term "score inflation" refers to any score increase that is not caused by an increase in students' learning of the skills and knowledge that the test is intended to measure.

Since at least the 1980s, one popular "solution" to the sometimes negative influence of testing on teaching has been calls for "tests worth teaching to," based on the notion that if tests were of high quality and measured complex skills and process, instruction would follow suit. This idea resulted in the wave of performance-based assessments in the 1990s. Evidence from some states' performance-based assessment programs suggests that these assessments can lead to some of the desired outcomes, such as increased emphasis on problem solving,[6] but for the most part these efforts have failed to lead to fundamental changes in how teachers deliver instruction.[7] Most states have backed away

*It is worth pointing out that the findings regarding reallocation in response to high-stakes performance measures are not limited to education. They have been observed in sectors as varied as health care, transportation, and emergency preparedness.[4]

from performance-based assessment because of costs and technical problems (e.g., states that implemented portfolio assessments found that scoring tended to be inconsistent and expensive[8]). Moreover, evidence suggests that simply adopting performance-based assessment does not eliminate the problems of narrowing what is taught or score inflation.[9] Although some have claimed that the Advanced Placement (AP) and International Baccalaureate (IB) programs might be considered successful implementations of the idea of tests worth teaching to, both of those programs' exams are aligned to well-defined course content. So, while their tests are generally high in quality and doing well on these tests is a legitimate goal of AP and IB courses, the key to these programs appears to be well-aligned instructional materials and assessments—not assessments alone.

This brings us to another popular "solution": standards. A number of factors have contributed to the appeal of standards-based teaching. One of these may have been the negative influence of high-stakes testing as a result of the minimum-competency testing movement. Standards may have seemed like a logical way to counter the narrowing of the curriculum and emphasis on lower-order, tested skills and content. However, efforts to promote more cognitively demanding instruction by building complex skills and knowledge into state or district content standards have been thwarted by the very tests used to assess those standards. Most states claim that their assessments are aligned with their standards, but these ostensibly aligned tests often sample only a subset of the standards,† with disproportionate emphasis on the lower-level content that is easier to test.[10] Because standards and high-stakes tests are not fully aligned, educators understandably tend to rely more on the tests than on the standards for instructional guidance.[11]

After 20 years of trying to align standards and tests, it is time to question whether this is even possible—at least in a meaningful way. Most standards are not highly specific or detailed. Typically, they are broad outcome statements that are wide open to interpretation. Assessments, however, are highly specific and detailed. Herein lies the problem with assessments aligned to standards: a teacher may faithfully and effectively teach to the standards all year and her students may learn a great deal, but her students may still do poorly on the test simply because the teacher and the test developer interpreted the standards differently. A curriculum, by specifying what knowledge and skills to teach and to test, could reduce the severity of this problem.

Clearly, assessment-based reforms (1) have not fully achieved policymakers' goals, and (2) have led to unintended consequences. These findings raise concerns about the extent to which assessment can be viewed as a means for improving educational outcomes. At the same time, assessment clearly plays an important role in providing information that helps teachers and other educators improve. Moreover, because testing affects what is taught, assessment has the potential to contribute to positive educational change *if* it is designed and implemented appropriately.

†Another problem is the low quality of the standards themselves, which tend to be either too vague to guide instruction or too detailed to be covered in one school year. For more on the problems with most states' standards, see the Spring 2008 issue of *American Educator*, available at www.aft.org/newspubs/periodicals/ae/spring2008.

Building a Better Assessment System

There is no research evidence to tell us definitively how to build an assessment system that will promote student learning and be resistant to the negative consequences that are common in high-stakes testing programs. One promising approach is to start with a detailed, coherent curriculum that is aligned with rigorous content standards, and then build an assessment system that measures the skills and knowledge emphasized in the curriculum. (Of course, using curriculum to guide assessment development would require a more consistent curriculum policy than currently exists in our states, a topic discussed throughout this issue of *American Educator*.) While it's inevitable that assessment will continue to drive instructional decisions, the less desirable consequences may be mitigated by providing educators with a high-quality curriculum and a set of supports like sample lesson plans and quizzes, ongoing professional development, and more time to confer with colleagues. Ensuring that all the components are well aligned should give teachers confidence that if they teach the curriculum effectively, the result will be improved student learning as measured by the assessments.

While assessment will continue to drive instruction, the consequences may be mitigated by providing educators a high-quality curriculum and supports like sample lesson plans and time to confer with colleagues.

The tendency to engage in practices that narrow the curriculum and cause score inflation stems in large part from a belief among educators that delivering the entire existing curriculum (or standards, in districts and schools that do not have a curriculum) will not ensure adequate coverage of the tested material. Teachers and principals understand that many aspects of their curricula/standards are not included on the accountability tests and that some of the tested material is not included in the curricula/standards (at least for that grade level).[12] A better-aligned system, modeled in part after the AP and IB programs (combined with some of the other suggestions discussed below), might help to assuage teachers' concerns about coverage and enable them to worry less about what is likely to be on the test.

This idea is not inconsistent with earlier notions of standards-based reform,[13] which advocated for alignment among not just standards and assessment, but standards, assessment, curriculum, and professional development. Many advocates of standards-based reform argued that standards should drive the development of *both* the curriculum and the assessments. While this makes sense in theory, in practice most standards are not written at a level of specificity that promotes the development of aligned curricula or assessments.[14] To date, no state has even

developed a statewide curriculum, much less based its assessment on a curriculum.

Even if a superb curriculum and well-aligned, high-quality assessment had been developed, our work would not be done. A sound accountability policy requires multiple sources of information and supports: not all of the outcomes that we want schools to promote can be measured easily or cheaply through large-scale assessments, and not all desired changes can be induced through improvements in assessment alone. Decision makers who understand the strong influence that high-stakes tests exert may, understandably, wish to rely heavily on assessment as a means to promote school improvement. For assessment to serve this role effectively, it must be designed in a way that supports rather than detracts from teachers' efforts to engage in high-quality instruction. Research on the effects of various assessment-design features is limited, so any effort that relies heavily on assessment as a tool for school improvement should be carried out with caution. Nonetheless, it is worth reviewing what is known and looks promising. Here are four approaches to designing assessment and accountability policies that are likely to support school improvement.

First, an accountability system that is designed to reward or penalize districts, schools, or individuals on the basis of their performance should not rely exclusively on tests. Although there is extensive research being conducted to guide improvements in large-scale testing, it is likely that society will continue to expect schools to promote outcomes (like critical thinking and responsible citizenship) that cannot be measured well using tests. In addition, even if the perfect assessments could be designed, it is not realistic to expect that it would be practical or desirable to

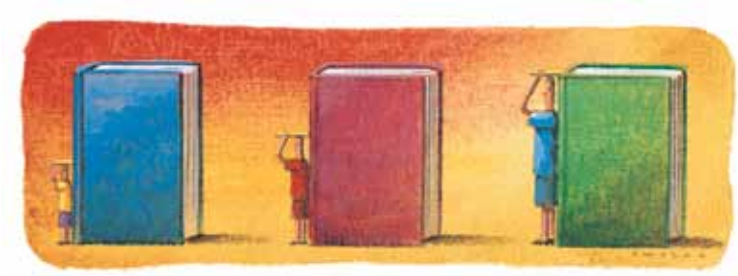
spend the time and money required to administer tests representing the full range of outcomes of interest. Accountability systems could supplement tests with non-test-based indicators of processes or outcomes, such as college-preparatory course taking, high school and college graduation rates, and apprenticeship completion rates. And, these systems could be designed in concert with current efforts by several teams of researchers and practitioners to develop improved test and nontest measures of teaching quality. When we look beyond tests alone to meet our information and accountability needs, a wide range of better options become available.

Of course, any supplemental measure should be evaluated using the same criteria for validity and reliability that are applied to test-based measures, and unintended consequences should be identified and addressed. One potential advantage of nontest indicators, such as peer and administrator observations and critiques of instruction, is that they might serve a more useful professional development function than test scores have, by providing teachers with clear, constructive feedback on their teaching. But if new measures (or rubrics) are used for both professional development and accountability purposes, investigations need to be designed to examine the validity of scores from those measures in light of each of those purposes, as well as the consequences that arise. Some problems, such as the tendency to focus on what is measured at the expense of what is not measured, are unlikely to be eliminated completely, so it will be important to monitor for undesirable consequences and modify the system as necessary to address them.

Second, for assessment and accountability to be useful, policymakers must consider ways to improve the quality of informa-

tion from the tests themselves, and to mitigate the expected negative effects of using tests for high-stakes purposes. In particular, designers of testing programs should take steps to reduce the likelihood of curriculum narrowing and score inflation. As mentioned above, basing the test on a detailed curriculum instead of broad standards will probably help. Another promising approach is to design tests to minimize predictability from one administration to another, so that focusing instruction on particular item formats or styles will not be viewed as likely means to raising scores. A single test administered at one point in time can sample only a fraction of the material in the curriculum, so varying this material over time, along with the types of items designed to measure it, should result in reduced curriculum narrowing and score inflation. In short, *if teachers had a high-quality curriculum and supporting materials at hand, and if the test were well-aligned but unpredictable, then teachers would probably just focus on helping all students master the skills and knowledge specified in the curriculum*. Of course, the problem of testing higher-order knowledge and skills would remain, but in the near future technology may offer new opportunities to design cost-effective and high-quality performance-based measures.[15]

When we look beyond tests alone to meet our information and accountability needs, a wide range of better options become available.



Third, any accountability system that seeks to support instructional improvement ought to include a high-quality formative assessment system—one that is aligned with the curriculum and provides clear instructional guidance rather than simply predicting students' scores on the state test.[16] But the assessment itself is just the beginning. The results must be accessible and available in a way that facilitates effective day-to-day use to guide instruction and be accompanied by ongoing professional development.

Finally, a number of other considerations need to be addressed when designing the testing components of an accountability policy, such as whether to focus the system on student or educator performance, on individual or group performance, on current achievement or growth, and on fixed targets or participant rankings.[17] These need not be such stark tradeoffs, but they do need to be considered. Many policymakers seem to want to say "All of the above," but such an unfocused and unwieldy accountability system would be very unlikely to promote school improvement.

Despite these challenges (and the dozens of more technical challenges that I have not addressed), it is likely that test-based accountability will be with us for some time. No doubt the policymakers who enthusiastically support such accountability are truly committed to school improvement—so they ought to see that heeding educators' and researchers' concerns about the purposes, meaningful uses, and technical limits of assessments is worthwhile. Working together, we can develop a program of large-scale assessment that addresses the information needs of educators, particularly at the classroom level, while also contributing to improved accountability policies. ❑

Endnotes

1. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (Washington, DC: American Psychological Association, 1999); and Michael T. Kane, "Validation," in *Educational Measurement*, 4th ed., ed. Robert L. Brennan (Westport, CT: American Council on Education/Praeger, 2006), 17–64.

2. For reviews of relevant literature, see Laura S. Hamilton, "Assessment as a Policy Tool," *Review of Research in Education* 27, no. 1 (2003): 25–68; Jane Hannaway and Laura S. Hamilton, *Accountability Policies: Implications for School and Classroom Practices* (Washington, DC: Urban Institute, 2008); and Brian M. Stecher, "Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practice," in *Making Sense of Test-Based Accountability in Education*, ed. Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein (Santa Monica, CA: RAND, 2002).

3. See, for example, Jennifer Booher-Jennings, "Below the Bubble: 'Educational Triage' and the Texas Accountability System," *American Educational Research Journal* 42, no. 2 (2005): 231–268.

4. Brian M. Stecher, Frank Camm, Cheryl L. Damberg, Laura S. Hamilton, Kathleen J. Mullen, Christopher Nelson, Paul Sorensen, Martin Wachs, Allison Yoh, and Gail L. Zellman, *Toward a Culture of Consequences: Performance-Based Accountability Systems for Public Services* (Santa Monica, CA: RAND, 2010).

5. Laura S. Hamilton, Brian M. Stecher, Jennifer Lin Russell, Julie A. Marsh, and Jeremy Miles, "Accountability and Teaching Practices: School-Level Actions and Teacher Responses," in *Strong States, Weak Schools: The Benefits and Dilemmas of Centralized Accountability*, ed. Bruce Fuller, Melissa K. Henne, and Emily Hannum, vol. 16, *Research in Sociology of Education* (St. Louis, MO: Emerald Group Publishing, 2008), 31–66; and Brian M. Stecher, Scott Epstein, Laura S. Hamilton, Julie A. Marsh, Abby Robyn, Jennifer Sloan McCombs, Jennifer Russell, and Scott Naftel, *Pain and Gain: Implementing No Child Left Behind in Three States, 2004–2006* (Santa Monica, CA: RAND, 2008).

6. Suzanne Lane, Carol S. Parke, and Clement A. Stone, "The Impact of a State Performance-Based Assessment and Accountability Program on Mathematics Instruction and Student Learning: Evidence from Survey Data and School Performance," *Educational Assessment* 8, no. 4 (2002): 279–315.

7. John B. Diamond, "Where the Rubber Meets the Road: Rethinking the Connection between High-Stakes Testing Policy and Classroom Instruction," *Sociology of Education* 80, no. 4 (2007): 285–313; and William A. Firestone, David Mayrowetz, and Janet Fairman, "Performance-Based Assessment and Instructional Change: The Effects of Testing in Maine and Maryland," *Educational Evaluation and Policy Analysis* 20, no. 2 (1998): 95–113.

8. See Daniel Koretz, Brian M. Stecher, Stephen P. Klein, and Daniel McCaffrey, "The Vermont Portfolio Assessment Program: Findings and Implications," *Educational Measurement: Issues and Practice* 13, no. 3 (1994): 5–16.

9. Daniel Koretz and Sheila Barron, *The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS)* (Santa Monica, CA: RAND, 1998).

10. Robert Rothman, Jean B. Slattery, Jennifer L. Vranek, and Lauren B. Resnick, *Benchmarking and Alignment of Standards and Testing* (Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, 2002).

11. Brian M. Stecher and Tammi Chun, *School and Classroom Practices during Two Years of Education Reform in Washington State* (Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, 2001).

12. Stecher et al., *Pain and Gain*.

13. Marshall S. Smith and Jennifer O'Day, "Systemic School Reform," in *The Politics of Curriculum and Testing: The 1990 Yearbook of the Politics of Education Association*, ed. Susan H. Fuhrman and Betty Malen (New York: Falmer Press, 1991), 233–267.

14. Heidi Glidden, "Common Ground: Clear, Specific Content Holds Teaching, Texts, and Tests Together," *American Educator* 32, no. 1 (Spring 2008): 13–19.

15. Edys S. Quellmalz and James W. Pellegrino, "Technology and Testing," Science 323, no. 5910 (2009): 75–79; and Bill Tucker, *Beyond the Bubble: Technology and the Future of Student Assessment* (Washington, DC: Education Sector, 2009), 1–9.

16. Marianne Perie, Scott Marion, Brian Gong, and Judy Wurtzel, *The Role of Interim Assessments in a Comprehensive Assessment System: A Policy Brief* (Washington, DC: Achieve, Aspen Institute, and National Center for the Improvement of Educational Assessment, 2007).

17. See, for example, Michael J. Podgursky and Matthew G. Springer, "Teacher Performance Pay: A Review," *Journal of Policy Analysis and Management* 26, no. 4 (2007): 909–950.