Turmoil in the Testing Industry

Thomas Toch

As the testing industry buckles under the weight of NCLB’s testing demands, states are opting for fast and cheap assessments that focus on basic skills.

Standardized achievement tests are crucial to No Child Left Behind's school reform effort because the legislation requires states to use these tests to measure whether students meet state standards. When insufficient percentages of students pass state tests, schools are judged as failing to make adequate yearly progress. And if they fail to make sufficient progress two or more years in a row, schools—and the educators who work in them—face increasingly severe consequences. At the heart of this accountability system is extensive testing of students in grades 3–8 and in one high school grade in reading and math.

Because schools tend to teach what's tested—especially when the test scores have consequences for teachers and principals—the content of the tests required by No Child Left Behind (NCLB) has become the focus of teaching and learning in public school classrooms throughout the United States. That focus would be fine if states administered tests that measured the sorts of skills and knowledge that would lead to a first-class education for every public school student, the result that NCLB advocates have asserted the law would produce.

But states don't usually administer those kinds of tests. The magnitude of NCLB's testing requirements, the law's demanding deadlines, insufficient federal funding, and other factors have produced a different result: Many states have adopted tests that can be constructed quickly and inexpensively. These tests primarily measure low-level skills, such as recall and restatement of facts, at the expense of synthesis, analysis, and other higher-order skills. Educators increasingly are focusing on the same low-level reading and math skills in their classrooms.

NCLB's goal is to raise instructional standards by requiring states to set challenging expectations for what students should know and be able to do. But many of the tests that states have introduced under NCLB are leading instruction in the opposite direction.

Heavy Demands on the Testing Industry

Creating high-quality tests is difficult and labor-intensive. The process involves determining the length and content of a test, hiring curriculum experts to write questions, and ensuring that the questions align with state standards. Test makers field-test the test items on thousands of students to ensure that these items don't discriminate against groups of students but do discriminate between strong and weak students, a complex mathematical task. Test makers also have to ensure that every multiple-choice question has only one correct answer and that the questions reflect an appropriate range of difficulty. They must perform another complex mathematical computation to ensure that the same scores on different tests represent the same
level of performance. Then the tests must be edited, printed, and distributed.

It’s a demanding process under the best of circumstances, and this complex test-making infrastructure is buckling under the weight of NCLB’s testing demands.

Moreover, the need to align tests to state standards forced the testing industry to custom-build the majority of the tests that were scheduled to be in place at seven grade levels in every state in spring 2006. And because a growing number of states release portions of their tests to the public after administering them each year, testing companies have to generate vastly larger pools of credible test questions and do so far more quickly. Many in the industry say that they can’t find enough qualified people to do the work.

The surge in state testing under NCLB has created a severe shortage of psychometricians—the specialists who do the heavy statistical lifting in test making. Only a handful of these experts, who are trained in measurement theory and statistics at the University of Iowa, Michigan State, the University of Massachusetts, and a dozen or so other colleges, enter the workforce each year.

Testing companies also face immense pressures at the back end of the testing cycle. In the pre-NCLB era, states and school systems gave testing companies months to score standardized tests because the results rarely had immediate consequences. Now, completed answer sheets are routed from schools to testing company scoring centers, where results are tabulated and then uploaded directly to state education department or school system computers. State agencies must then analyze the results, grade schools and school systems on the basis of whether sufficient percentages of their students as a whole and in every subgroup have met state standards on the tests, and package the ratings in reports that NCLB requires them to supply to school systems. School systems, in turn, must route the state ratings to schools and parents.

All these reports must be completed in time for parents to place their children in tutoring or in different public schools before the start of the next school year, an opportunity that NCLB grants students in schools that fail to make adequate yearly progress. With many schools starting in August, the entire testing and state rating process must be completed by mid-July in many places—only three or four weeks after the end of the typical public school year.

It would be difficult enough to successfully complete this process with long time lines. But many state policymakers, under pressure to give students as much time as possible to prepare for NCLB’s high-stakes tests, are demanding that schools administer tests late in the school year.

Lobbying by local educators persuaded the Ohio legislature in 2005 to move the state’s two-week testing window from March to May, beginning in 2007. The legislature also mandated that Ohio’s testing contractors—the American Institutes of Research and Measurement Incorporated—report scores on the tests by June 15—two weeks earlier than in the past. Some states want even quicker turnarounds. Michigan fired Measurement Incorporated in 2005 for months-long delays in scoring the state’s tests. Pearson Educational Measurement, the state’s new contractor, is required to get test results to local school systems within 30 days.

Shoestring Budgets

State departments of education and their testing contractors must meet NCLB’s testing mandates on shoestring budgets. Eduventures, a Boston-based research firm, estimates that states spent $517 million in the 2005–2006 school year on NCLB testing (Jackson & Bassett, 2005). Some testing company executives peg the number even higher, at $700–$750 million. But that’s still a small portion of the approximately $500 billion the United States spends on public elementary and secondary education annually (U.S. Department of Education, 2005). Indeed, a study by Harvard economist Caroline Hoxby (2002) found less than one-quarter of 1 percent of a state’s public school spending typically went to its statewide testing programs.

As a result, many statewide tests are not getting sufficient psychometric scrutiny to ensure that they accurately measure student and school performance, say testing experts. "States and
contractors should be doing a lot more validity studies, to be sure that what the tests are saying about student achievement is accurate,” says Scott Marion, vice president of the New Hampshire–based Center for Assessment, a nonprofit test-consulting firm that advises 15 state testing agencies. “But they aren't doing it.”

Lack of time, money, and skilled staff have also led a substantial number of states to introduce tests that are not fully aligned with state standards. Instead of building tests from scratch, some states hire testing companies to “augment” the Stanford and other national norm-referenced tests with questions that cover topics in state standards. But the tests aren't always what they should be. According to Marion, test publishers almost always claim that their tests are at least 85 percent aligned with state standards, "But our studies show that the percentage of alignment is a lot lower, 50 percent. As a result, teachers are teaching stuff that they can't be sure is on the tests.”

**A Race to the Bottom**

The most troubling classroom consequence of the tumult in the testing industry is the pressure placed on states and their testing contractors to build tests that measure primarily low-level skills. Simple tasks are easier and cheaper to test.

Test questions that measure lower-order competencies do have a role to play; schools need to understand students' grasp of basic skills. But because teachers have so much riding on their students' test results, tests that stress basic skills encourage teachers to emphasize these skills in their classrooms at the expense of other more demanding standards.

Such tests also give a skewed picture of student achievement. Scores on reading tests that measure mainly literal comprehension tend to be higher than scores on tests that require students to evaluate what they've read—for example, by reading two passages and identifying themes common to both. The same is true in math. In a study by Lorrie Shepard (1997), 80 percent of a national sample of 8th graders correctly identified the product of 9 × 9, but only 40 percent correctly answered a problem asking them to calculate the square footage of a 9 × 9 foot room. Many state tests, as a result, are likely to suggest that students are achieving at higher levels than they really are. The tests may create glass ceilings for higher-achieving students, who have less opportunity to demonstrate the extent of their abilities. And when the scores of low-achieving students rise, this achievement ceiling may create the illusion that performance gaps among groups of students are closing when in fact they are not.

Multiple-choice questions *can* measure higher-level skills, but writing such questions is difficult and time-consuming. Most testing experts would prefer to measure students' grasp of more advanced abilities through open-ended or constructed-response questions that require students to produce their own answers. But such questions are more expensive and slower to process than their multiple-choice counterparts. To score more open-ended tests, states and their testing contractors must first establish rubrics for judging students' responses. They have to hire and train test graders to field-test the rubrics and then again to score the open-ended questions themselves. Testing companies spend between two days and one week training their test graders to ensure that answers of comparable quality receive the same scores from different graders.

Scoring open-ended questions requires both technology and people: Students' responses are electronically scanned so that they can be evaluated by the hundreds of graders who sit at banks of computers in sprawling scoring centers, working their way through hundreds of responses at a rate of 20 to 30 per hour. The result is that it costs anywhere from 50 cents to 5 dollars to score a constructed-response question, compared with pennies for a multiple-choice question, says Gary Cook, a research scientist at the University of Wisconsin's Center for Education Research who served as Wisconsin's testing director.

The cost differential is not lost on the state legislators who control state education department budgets. In 2004, Pearson brought the membership of the Michigan House and Senate education
committees to Iowa City to tour the company’s high-tech facility for scoring multiple-choice answer sheets. The legislators were wowed by the speed and low cost of the process they witnessed; once back in Michigan, they promptly pushed the state’s testing officials to drop open-ended questions from the state’s tests, says Edward Roeber, Michigan’s testing director.

As a result, there are few open-ended questions on many of the new state tests, say testing experts. According to John Olson, director of psychometric and research services at Harcourt Assessment, who spent a decade working on the federally funded National Assessment of Educational Progress (NAEP),

states are shifting from constructed response to multiple-choice because of the cost and time of scoring constructed-response questions…. During the 1990s states had more top-end questions, more challenging, NAEP-like questions. They tested student ability over a wide range; they used more constructed-response questions. There was a lot more attention to making high-quality tests.

Mississippi and Kansas eliminated non-multiple-choice questions from their state tests in 2005–2006. In all, 15 states serving 42 percent of U.S. students used reading and math tests in 2005–2006 that had no open-ended questions, says Education Week (Olson, 2005).

Where to Go from Here

U.S. Secretary of Education Margaret Spellings has echoed President Bush’s claims about NCLB’s potential to produce a “first-class education” for all students, but she has sought to side-step the issue of the low quality of many state tests under NCLB. Connecticut has had high-quality tests with many open-ended questions since the 1980s, including math questions that require students to write explanations of their answers. In 2005, the state sued the U.S. Department of Education over the cost of NCLB testing, arguing that Connecticut’s share of federal NCLB testing monies is inadequate to fund tests of the same caliber as the state’s past tests. In a letter to Betty Sternberg, then Connecticut’s commissioner of education, Spellings responded,

Some of the costs of [Connecticut’s testing system] are attributable to state decisions [regarding the types of tests it uses]. While these decisions are educationally sound, they do go beyond what was contemplated by NCLB. (Spellings, 2005)

The confrontation between Connecticut and the U.S. Department of Education was predictable. The General Accounting Office (now the Government Accountability Office) issued a report that included three widely varying estimates of what states would have to spend to comply with NCLB testing mandates between 2003 and 2008: The cost would be $1.9 billion if states used tests with only multiple-choice questions, $3.9 billion if they used multiple-choice questions and some open-ended items, and 5.3 billion if they used tests with a larger percentage of open-ended items (Shaul, 2003). But the agency didn’t predict the harmful consequences for U.S. classrooms of the low-cost option—an option that Secretary Spellings has sought to defend and that many state appropriators have chosen.

Correcting the problems of poor test quality will require a commitment to vastly improving the testing infrastructure in public education—to building a system of high-quality tests that deliver dependable accountings of student and school performance, encourage schools to aim higher, and supply educators with timely information on students' strengths and weaknesses. That will mean investing resources commensurate with testing’s central role in school reform today. If policymakers don’t acknowledge the problem and take steps to address it, the chances of achieving NCLB’s lofty goal of first-rate instruction for all students are slim.

References


*Author’s note:* A more detailed discussion of the issues raised in this article is available in the Education Sector report *Margins of Error: The Education Testing Industry in the No Child Left Behind Era* ([www.educationsector.org](http://www.educationsector.org)). Nearly three dozen testing industry executives, state testing officials, and other testing experts shared their candid assessments of the testing industry for this report.

**Thomas Toch** is Cofounder and Co-director of Education Sector, a nonpartisan think tank in Washington, DC; 202-552-2841; [ttoch@educationsector.org](mailto:ttoch@educationsector.org).

---

Copyright © 2006 by Association for Supervision and Curriculum Development

© Copyright ASCD. All rights reserved.