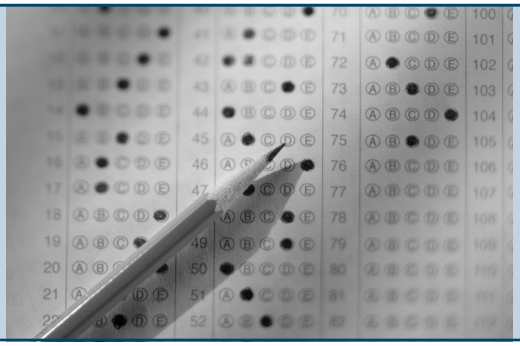
EDUCATION**SECTOR** REPORTS



MARGINS OF ERROR:

The Education Testing Industry in the No Child Left Behind Era

by Thomas Toch

Table of Contents

1101 Pennsylvania Ave, N.W., Fifth Floor, Washington, DC 20004
202.756.4944 • www.educationsector.org

State standards and standardized tests have become dominant forces in American public schooling. For most of its history, public education in the U.S. was a local matter, with local schools and school systems setting their own educational priorities. But in the wake of mounting evidence that the preparation most students received from public schools wouldn't suffice in a postindustrial economy, and with the conscience of the nation having been transformed by the civil rights movement, policymakers began to pursue a new paradigm, one that sought to establish statewide public school standards and hold local educators accountable if their students fell short of these standards. Standardized tests, used to measure student performance against the new state expectations, are the linchpin of this strategy of standards-based reform.

The No Child Left Behind Act of 2001 (NCLB) solidified standards-based reform as a national priority, part of a bold attempt by federal policymakers to force state and local educators to improve the education of minorities and other students that public schools traditionally hadn't served very well. The legislation required that by the spring of 2006 states test nearly every public school student in grades three through eight and in one high school grade to gauge whether students have met standards in reading and math—a task requiring some 45 million standardized tests annually.

To comply with NCLB, 23 states that have not yet fully implemented the law's testing requirements will administer some 11.4 million new tests during the 2005-06 school year alone, half in reading, half in math. Within two years, states must begin testing students in a minimum of one elementary, middle, and high school grade in science under NCLB, requiring at least another 11 million tests.[1]

Standardized test scores form the basis of NCLB's accountability mechanisms—school report cards, tutoring and school-choice options for students, and serious consequences for low-performing schools. Increasingly, as a result, the content of statewide tests has become the focus of teaching and learning in public school classrooms throughout the nation, to the point where many schools have begun to do much more testing than is required by NCLB, in an effort to prepare their students for the high-profile NCLB-mandated exams.

But this surge in testing has created immense challenges for both the industry that writes, scores, and reports the vast majority of the new statewide tests and the state agencies charged with carrying out NCLB's requirements. NCLB's test-based accountability system has given local educators powerful incentives to help students whom public education has long neglected. But the scale of the NCLB testing requirements, competitive pressures in the testing industry, a shortage of testing experts, insufficient state resources, tight regulatory deadlines, and a lack of meaningful oversight of the sprawling NCLB testing enterprise are undermining NCLB's pursuit of higher academic standards.

Symptoms of the turmoil in the testing industry aren't difficult to find: Newspapers carry accounts of testing companies giving students college scholarships to atone for the fact that scoring errors deprived them of their high school diplomas; of scoring errors sending thousands of students to summer school when they had in fact passed their tests; of months-long scoring delays; of administrators losing their jobs for low scores on tests that, had they been scored correctly, would have shown improvements in student achievement.[2]

These problems have damaged the credibility of standards-based reform in the eyes of many educators and parents, and they have attracted the attention of the Office of Inspector General at the U.S. Department of Education, which has announced plans to examine the extent of test-scoring and reporting mistakes under NCLB.[3]

But there are deeper, more structural problems stemming from the tremendous expansion of statewide standardized testing that haven't made headlines in the buildup to the full implementation of NCLB's testing requirements in spring 2006. Many states are constructing tests that don't fully measure student and school performance against state standards. And they are using tests that measure mostly low-level skills, a move that encourages teachers to make the same low-level skills the priority in their classrooms at the expense of the higher standards that NCLB has sought to promote.

The testing infrastructure that undergirds NCLB's accountability system must be improved, as this report makes clear, and if steps aren't taken to do so, teachers and principals will lose valuable tools to improve instruction, and both NCLB's work on behalf of public education's neediest students and standards-based reform itself will be increasingly at risk. Statewide testing, envisioned under NCLB as a key part of the solution to what ails public schools, is fast becoming part of the problem in public education.

States Adding Reading and Math Tests in 2005-06

STATE	TOTAL TESTS	STATE	TOTAL TESTS
Connecticut	311,286	New Jersey	619,588
Illinois	974,160	New York	1,704,592
Kansas	315,138	Ohio	850,850
Kentucky	443,828	Oklahoma	190,066
Maine	126,474	Pennsylvania	864,686
Massachusetts	456,750	Rhode Island	150,796
Michigan	1,062,907	Vermont	89,104
Minnesota	383,214	Virginia	565,096
Missouri	566,802	Washington	629,156
Montana	90,602	Wisconsin	512,240
Nevada	188,358	Wyoming	59,147
New Hampshire	198,032		
TOTAL			11,352,872

Source: Editorial Projects in Education Research Center; National Center for Education Statistics; Education Sector calculations

The Industry

The testing industry is surprisingly small, given its outsize role in public education today. Eduventures Inc., a Boston-based research firm, estimates that the value of tests, testing services, and test-prep materials purchased in 2006 will be $2.3 billion. But that includes purchases by school systems and schools, state-level testing, and college-admissions testing and test prep. Total expenditures for developing, publishing, administering, grading, and reporting NCLB-required statewide tests, Eduventures estimates, will be $517 million in the 2005-06 school year.[4] Some testing company executives peg the number somewhat higher, at $700 million to $750 million, still a small portion of the approximately $500 billion the United States spends on public elementary and secondary education annually.[5]

A handful of companies capture some 90 percent of the statewide testing revenue, Eduventures estimates. They include Pearson Educational Measurement, a subsidiary of London-based publisher Pearson PLC; CTB/McGraw-Hill, a division of the New York-based publishing and information conglomerate McGraw-Hill Cos.; Harcourt Assessment Inc., owned by Anglo-Dutch publishing giant Reed Elsevier; Riverside Publishing, a division of the privately owned publisher Houghton Mifflin Co.; and the nonprofit, Princeton-based Educational Testing Service (ETS), best known as maker of the SAT college-admissions test. They are "full-service" companies that create tests; align them with state standards; ensure they are technically sound; publish, distribute, and score them; and analyze results.

Other, smaller full-service companies have entered the statewide testing business more recently, including Measurement Inc., Questar Educational Systems, Data Recognition Corp., and non-profits Measured Progress, Northwest Evaluation Association, and American Institutes of Research. And there is a growing number of niche companies that focus on aspects of the testing enterprise such as test-question writing or test scoring.

The Major Players

CTB/McGraw-Hill
Major Test: TerraNova
Activities: Development, Administration, Scoring, Reporting
State Testing Contracts: 23
K-12 Tests Administered, 2005: 16.5 million
Test Items Written, 2005: 167,000
Percent of Nation's Students Taking a CTB test: 35
K-12 Testing Revenue: N/A
Corporate Parent: The McGraw-Hill Companies

Educational Testing Service
Major Test: None. Creates Custom-designed Statewide Exams
Activities: Test Development, Reporting
State Testing Contracts: 15
K-12 Tests Administered, 2005-06: 7 million
K-12 Testing Revenue: $150 million
Key Fact: Best Known as Creator of the SAT; New Player in K-12 Testing
Corporate Parent: Non-profit

Harcourt Assessment
Major Test: Stanford 10
Testing Activities: Development, Administration, Scoring, Reporting
State Testing Contracts: 22
K-12 Tests Administered, 2005-06: 9.5 Million
Test Items Written, 2004: Nearly 85,000
Open-ended Questions Scored, 2005-06: 40 Million
K-12 Testing Revenue: N/A
Corporate Parent: Reed Elsevier

Pearson Educational Measurement
Major Test: None. Creates Custom-designed Statewide Exams
Activities: Development, Administration, Scoring, Reporting
State Testing Contracts: 20
K-12 Tests Administered, 2005-06: 40 million
K-12 Testing Revenue: N/A
Key Fact: Nation's Largest Test Scorer
Corporate Parent: Pearson PLC

Riverside Publishing
Major Tests: Iowa Test of Basic Skills
Activities: Design, Development, Scoring, Assessment Management
State Testing Contracts: 4
K-12 Tests Administered, 2005-06: N/A
K-12 Testing Revenue: N/A
Key Fact: Top Player in Formative-assessment Market
Corporate Parent: Houghton Mifflin Company

Source: Testing companies

The Path to NCLB

The major players have been around for a long time. Harcourt's Stanford 10 test and CTB-McGraw Hill's TerraNova tests date to the 1920s, Riverside Publishing's Iowa Test of Basic Skills to the 1930s.

But in keeping with the tradition of local control in public education, publishers for decades sold their elementary- and secondary-school achievement tests only to schools and school systems, where they were used to compare local student performance with that of representative national samples of students—so-called norm groups.

The publishers' local sales staff sold the tests at the same time they sold textbooks, because the major publishers were in both businesses. There was thus a patchwork of different tests in place in every state rather than a single, statewide testing system. There were typically no consequences for local educators for their students' performance on the tests. And there wasn't any attempt to measure student performance against state standards, because with local school systems establishing their own educational agendas, there weren't any statewide standards.

All that began to change in the late 1960s. The Elementary and Secondary Education Act of 1965 (ESEA), of which NCLB is the latest reauthorization, called for evaluation of federal programs for disadvantaged students and set aside funding for the task. A nascent accountability movement also took shape in the 1970s, as state lawmakers, in the face of reports that many students weren't learning and demands that state officials address the problem, started to require statewide testing programs as a way of ensuring that students had "minimum competencies" in core subjects; they wanted to know, for example, that sixth-graders were performing at least as well as typical fourth- or fifth-graders. Michigan created the first statewide standardized testing program, in 1969, and Florida created the second, in 1971.

The minimum-competency movement expanded to other states during the 1970s, with lawmakers typically requiring testing at two or three grade levels each year. And the movement spread more rapidly in the early 1980s with the publication of *A Nation At Risk* and several other national studies that laid bare the troubled state of the nation's public schools. New funding for school reforms began to flow from state coffers in the wake of the reports, and

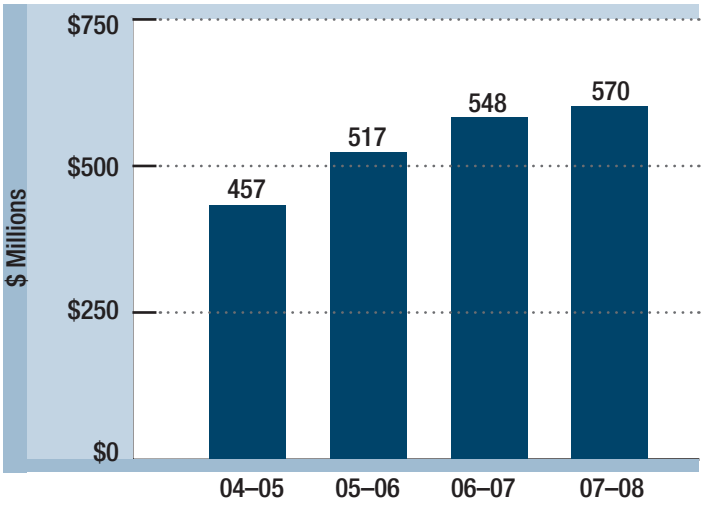
lawmakers wanted evidence that their investments were paying dividends, so they mandated more testing.

By the end of the 1980s, frustration with the pace of local reforms had led President George H.W. Bush to convene a national summit with the nation's governors in Charlottesville, Va., to explore ways to promote reform on a larger scale. Bush and the governors, led by then-Arkansas Gov. Bill Clinton, helped win bipartisan support for standards-based reform by establishing as a national goal that students demonstrate "competency" in core subjects such as English and math in grades four, eight, and 12. By the time Congress reauthorized the Elementary and Secondary Education Act in 1994, Clinton was in the White House and standards and accountability were the watchwords of reform.

Known as the Improving America's Schools Act (IASA), the Clinton administration's ESEA-reauthorization legislation required every state to put in place both standards and tests in reading and math at three grade levels, and about two-thirds of the states had done so by the end of the second Clinton presidency.[6] Most of the tests were designed to measure whether students mastered states' standards, rather than how they compared with students nationally; they were "criterion-referenced" tests rather than the "norm-referenced" tests that most states had introduced previously.

State Spending on Standardized Testing

Millions spent per academic year on the development, publishing, administration, grading and organizing of state exams.



Source: The State of the K-12 State Assessment Market, Eduventures, Inc., 2005

NCLB built on the Clinton-era accountability measures. It more than doubled the amount of testing required of the states, from three grade levels to seven; it established much tighter deadlines for introducing new tests; it required that results be broken down by a range of subgroups of students in every school; and, most significant, it linked serious consequences for schools to student test scores. Today, under NCLB, more students are tested more often than at any time in the nation's history, and the stakes are far higher.

'Harder Than the Dickens'

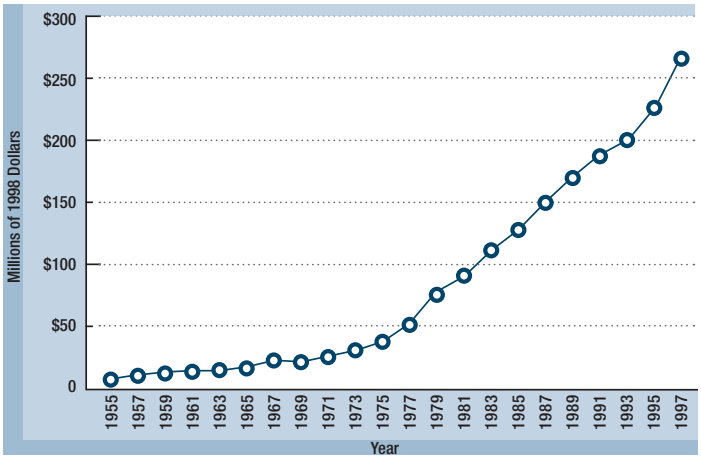
Creating high-quality tests is difficult and labor intensive. The process involves determining the length and content of a test, hiring curriculum experts to write questions, and ensuring that the questions align with state standards so the questions test what students are supposed to know. Then questions are field-tested on thousands of students to ensure that they don't discriminate against groups of students but do discriminate between strong and weak students, a complex mathematical task that requires comparing how students do on other questions with how they perform on the questions being trial-tested. Test-makers also have to ensure that every multiple-choice question has one, and only one, correct (or clearly best) answer and that the questions on a test reflect an appropriate range of difficulty. Another complex statistical computation has to be performed to ensure that the same scores on different tests represent the same level of performance. Then tests have to be edited, printed, and distributed to every public school in the country.

It's a demanding process under the best of circumstances. "Hundreds of people have to touch every item," says Gary Cook, a research scientist at the University of Wisconsin's Center for Education Research, who served as Wisconsin's testing director and as vice president of state accounts at Harcourt Assessment. The difficulty of writing questions that can clear testing's many hurdles has typically resulted in a majority of them being dropped—even those drafted by the most experienced item writers, says H.D. Hoover, who was the principal author of the Iowa Test of Basic Skills (ITBS) for nearly two decades. "Building a good test looks easy, but it's harder than the dickens," says Hoover. Ohio now administers 400 different forms of its statewide tests in order to field-test enough questions to keep its bank of test items stocked. As a result, it costs anywhere from $300 to $1,000 to develop a simple multiple-choice question, the least

expensive type of test item. State tests typically have 50 to 100 questions per subject per grade.

This complex test-making infrastructure is buckling under the weight of NCLB's testing demands. "There's way too much demand and not enough supply," says Hoover.

The Rise of Standardized Testing: Test Sales, 1955 to 1997



Source: The Bowker Annual of Library and Book Trade Information, 1970-98; Association of American Publishers, 1970-1998

When tests were purchased by schools and school systems, and in the early years of the standards movement, when most states used the Stanford and other major norm-referenced tests as their statewide exams, test publishers produced a new battery of tests only every six to eight years, because they didn't release test items and were thus able to recycle their tests. The result was a manageable demand for test items and ample time to vet them. NCLB has changed that.

The law's requirement that states align their tests to challenging state standards is an important step toward clarifying classroom expectations. But it is forcing the testing industry to custom-build the majority of the tests that must be in place at seven grade levels in every state this spring. And because a growing number of states release at least portions of their tests once they have been administered each year in order to give educators and parents more-detailed reports of student performance, testing companies have to generate vastly larger pools of credible test questions and do so within far shorter timelines. In the view of many in the industry, they can't find enough qualified people to do the work.

Testing companies "are desperately looking for people to write test items," says Hoover. "It's hard to do well, and it's hard to recruit people to do it." Adds Daniel Koretz, a testing expert at the Harvard Graduate School of Education: "Testing company executives tell me, 'We're having a hell of a time finding the caliber of people we need.' "

The surge in state testing under NCLB has created a severe shortage of the specialists who do the analyses of how test items perform in field trials and the other heavy statistical lifting in test-making. Though the work of these experts, who are trained in measurement theory and statistics and are known as psychometricians, is crucial to creating high-quality tests, only a handful of them enter the workforce each year from the University of Iowa, Michigan State, the University of Massachusetts, and the dozen or so other campuses that train them—under a dozen a year, reports a survey of doctoral degrees earned nationwide between 1995 and 2003 by the National Opinion Research Corp. An additional 35 Ph.D.s were awarded annually in the related field of "statistics, testing and education measurement." Cook, the University of Wisconsin testing expert, refers to NCLB as the "No Psychometrician Left Unemployed Act." [7]

The dearth of testing experts isn't hard to explain. Psychometrics is a highly technical, mathematics-based discipline that doesn't pay particularly well by private-sector standards (about $120,000 a year in top industry slots and much less in state testing agencies). And many potential recruits, undergraduates studying educational and quantitative psychology at colleges of education, are discouraged from entering the field by education school professors, many of whom are opposed to the rise of standardized testing in public education, says Hoover. "They say we're the bad guys."

To make matters worse, a growing number of psychometricians are working on non-educational tests, says James Impara, president of the National Council on Measurement in Education, a professional organization for education measurement experts. Over 1,000 occupations from accounting to firefighting now require licensure or certification, Impara says, and many of them give tests to applicants.

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-

long windows in which 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 computers or, as in Michigan, back to school systems and from there to the state agencies.

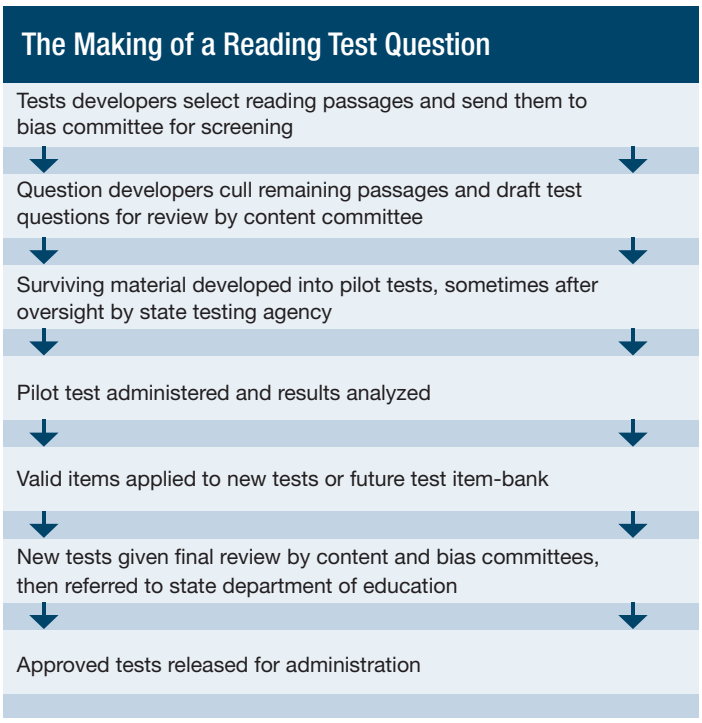
States' testing staffs calculate the percentages of students meeting state standards in reading and math. Once they do this for every NCLB student subgroup (students are grouped by race/ethnicity, family income, disability, and language proficiency) in every tested grade in every public school and school system, they 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, what NCLB calls "adequate yearly progress." Then the state agencies 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—in time for parents to place their children in tutoring or in different public schools prior to 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 up in August, that means that the entire testing and state rating process must be completed within six weeks from the end of the typical public school year.

This would be difficult enough to do successfully with long timelines. But many state policymakers, under pressure from public educators to give students as much time as possible to prepare for NCLB's high-stakes tests, are demanding that tests be administered late in the school year and that testing companies nonetheless complete their scoring and reporting in time to place underachieving students in summer school and in time for states to do their adequate yearly progress calculations ahead of the midsummer deadline for public reporting, says Jeff Galt, chief executive officer of Harcourt Assessment.

Lobbying by local educators led 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, Washington, D.C.-based American Institutes of Research and North Carolina-based Measurement Inc., both small companies, report scores on the tests by June 15—two weeks faster than in the past. And some states want even quicker turnarounds. Michigan ended its contract with Measurement Inc. in 2005 in the wake of months-long delays in scoring the state's tests. Pearson, the state's new contractor, has to get test results to local school systems within 30 days.

Despite the fact that testing companies have sought to upgrade their test-processing infrastructure in recent years, the pressure put on them by the volume of testing and the new scoring deadlines is "immense," says Scott Marion, vice president of the New Hampshire-based Center for Assessment, a nonprofit test-consulting firm that advises 15 state testing agencies. And it is intensified, says Stuart Kahl, president and CEO of Measured Progress, a New Hampshire-based testing company, by the fact that companies often must spend weeks after students take tests tracking down test booklets that school systems have failed to forward, resolving discrepancies between enrollment figures and the number of students' tests, and "cleaning up" basic student biographical information required to report test results under NCLB.

Headlines about scoring blunders are one measure of the overwhelming demands of the scale and speed of test processing required by NCLB. Another is that over half of the school systems in a 2005 national survey by the



Source: Scott Marion, Center for Assessment

Glossary of Terms

Validity: The extent to which tests accurately measure the knowledge or skills that the tests are intended to measure.

Constructed response: A test item that requires students to provide the answer to a question, as opposed to a multiple-choice question, where students choose among possible answers that the test creator provides. Constructed-response questions can be as simple as fill in the blank (e.g., 9 x 9 = ?) or require more complex answers, such as a written essay.

Rubric: A tool used to score answers to a test question. Most commonly used for scoring answers to more complicated, constructed-response questions, rubrics help ensure consistent grading from different graders by describing the specific elements of the answer needed for students to receive various score levels.

Criterion-referenced test: An assessment that measures the extent to which students have mastered a specific body of knowledge and skills, such as standards-based tests that determine if students have reached certain predefined levels of proficiency in a subject.

Norm-referenced test: An assessment that measures a student's performance relative to that of a representative national sample of students called a norming sample. Results on norm-referenced tests are often expressed in percentiles. A student scoring at the 60th percentile on a norm-referenced test, for example, scored better on the test than 60 percent of the students in the norming sample.

Scaling: A process of converting raw test results, such as the number of correct answers, into a score that can be used to compare results from different students or different versions of a test. Student scores on the SAT, for example, are converted to a scale where the minimum score for each section is 200 and the maximum score is 800.

Equating: The process of placing scores from different versions of the same test on a common scale, so that student results on those tests can compared on a fair, "apples to apples" basis. Equating, for example, ensures that a score of 700 on the SAT in 2005 is comparable to a score of 700 in 2006.

Reliability: A measure of the consistency and dependability of a test score's representation of a student's knowledge and skills. There are various dimensions of reliability, such as the consistency of test results when comparing the administration of the same test at different points in time, or comparing the results of different questions that measure the same skill, or comparing the scores given by different graders to the same answers.

Source: Scott Marion, Center for Assessment; Pearson Educational Measurement

Center on Education Policy said that late reports by state departments of education created "serious or moderate" problems in meeting NCLB's start-of-the-school-year deadline for informing parents of their children's eligibility to attend higher-performing public schools.[8]

Market Pressures

The testing industry is facing these challenges in a time of tight budgets and thin margins. A study by Harvard economist Caroline Hoxby revealed that states typically spend less than one-quarter of 1 percent of public school revenues on their statewide testing programs.[9] In 2005-06, combined federal, state, and local per-student spending in public education averaged over $8,000. Despite testing's tremendous influence on what students are taught and how teachers teach in the nation's public schools, and despite the importance of testing to school reform under NCLB, states spend between $10 and $30 per student on their testing programs, says Harcourt's Galt and other industry experts. Eduventures estimates that schools and school systems spend twice that amount on test-prep materials.[10]

The major testing companies weren't fazed by state testing budgets when they were selling large quantities of the ITBS, Stanford, and other national norm-referenced tests to schools and school systems. These "catalogue sales," as they are known in the testing industry, were lucrative, says Cook, who coordinated bids for state testing contracts at Harcourt. Publishers would invest $3 million to $6 million in a testing series and earn $15 million to $20 million over the five- to eight-year life of the tests, he says, because they were able to keep the development cost of every test booklet they sold low by using the same tests for a number of years.

But schools and school systems are buying far fewer of the major norm-referenced tests in the NCLB era of statewide testing. Sales of such tests are down 30 percent to 70 percent, says Cook. In the new NCLB marketplace, the publishers must make customized criterion-referenced tests that measure students' grasp of each state's unique academic expectations. Such tests, Galt says, can be five times as expensive to construct as the ITBS and other norm-referenced tests. And the fact that many states own the copyrights to their tests and release them to the public after the tests have been administered has wiped out the economies that publishers enjoyed when they were able to use their norm-referenced tests for several years

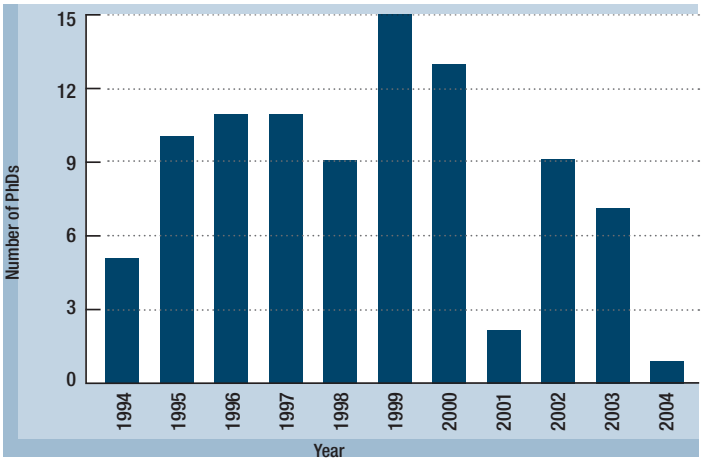
running. The result, predictably, has been much lower profit margins. “When you are building a new test every year,” says Hoover, the ITBS author, “it’s difficult to get your money back.”

To compensate, the major testing companies are vying aggressively—both with one another and with Measured Progress, Data Recognition Corp., and the other new players—for as many state NCLB testing contracts as possible, in an attempt to achieve efficiencies through scale.

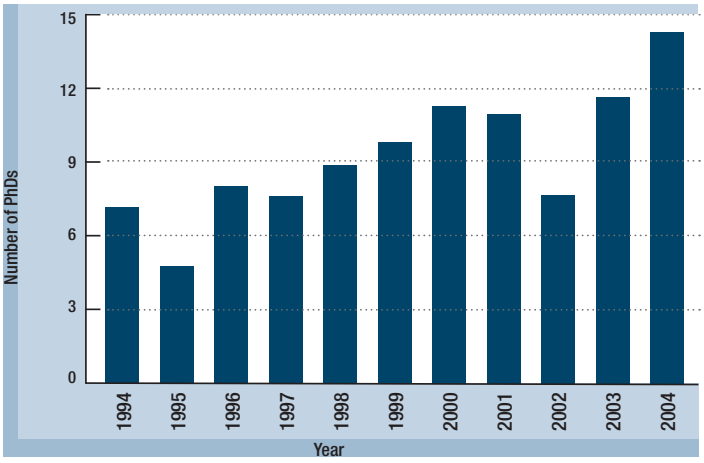
But this intense competition has led to more pressure on profit margins. Princeton-based ETS lost $18 million on its first NCLB testing deal, a three-year, $175 million contract with California, the nation’s largest market, says Anthony Carnevale, a former ETS vice president. Nonetheless, California tentatively approved a new three-year deal with ETS in late 2005, for even less money. John Oswald, an ETS senior vice president and general manager of the company’s elementary- and secondary-education division, says ETS lost money on its first California contract, made a slight profit during a one-year extension to the deal, and expects the new contract to be profitable.

Talent Search

Ph.D.s granted in Psychometrics, 1994 to 2004



Ph.D.s granted in Educational Assessments/Testing/ Measurement, 1994 to 2004



Source: National Opinion Reasearch Corporation, 2004

“With three or four companies bidding, one always has a reason to low-ball to get into the state,” says Kahl of Measured Progress. Pearson won a three-way competition in 2005 for Michigan’s testing business with a $48 million bid on a three-year contract. The second company, Data Recognition, bid $84 million, and the third, Measurement Inc., $114 million. Says Jon Twing, senior vice president for test and measurement services at Pearson, where he heads the company’s 200-person division that develops state tests: “We are reducing our rates in bids because our competitors are doing the same thing. It cuts profit margins.”

“There’s a tremendous amount of competition to get volume to cover fixed costs,” says Galt, Harcourt’s president. In response to higher testing volume and tighter scoring deadlines, Harcourt is spending $50 million over three years on printers, scanners, software, and other test-processing infrastructure—“most of which,” Galt says, “remain idle for 10 or 11 months of the year.” Doug Kubach, president and CEO of Pearson Educational Measurement, says there’s “hyper-competition in the industry.”

Penalty clauses in state testing contracts that have become more common and more prescriptive since NCLB became law are also squeezing testing company profit margins. Pearson’s 2005 deal with Michigan, for instance, stipulates that the company must pay the state four cents per student per day for every NCLB test it fails to score and return to school systems within four weeks—a potential maximum fine of $100,000 a day. Pearson’s penalty clock began ticking in early December of that year, and the company quickly racked up two weeks’ worth of fines in several parts of the state, says Roeber, adding that, “We’ll probably save some money on our testing program this year.”

Meanwhile, NCLB has spawned a secondary testing market that is further taxing the testing industry’s capacity. The tremendous pressure on schools and

school systems to have their students to do well on NCLB tests and the advent of technology that permits companies to very rapidly give school superintendents, principals, and teachers detailed breakdowns of student test results have produced a burgeoning market among school systems for so-called formative tests—short tests that are administered throughout the school year to help educators respond quickly to student weaknesses. The rise of formative testing has increased demand for banks of test questions dramatically, industry experts say. Sales increased 50 percent between 2003 and 2006, to $323 million, Eduventures estimates.

Industry Leaders

Nine companies capture more than 95 percent of the expenditures from states for tests and testing services.

COMPANY & PRIMARY STATE CONTRACTS

Education Testing Service
California
New Jersey

Harcourt Assesment
Arizona
Deleware
Hawaii
Idaho
Illnois
Mississippi
New Mexico
Rhode Island
South Dakota
Virginia
Wyoming

Riverside Publishing
Arkansas
Iowa
Louisiana

CTB/McGraw-Hill
Alabama
Alaska
California
Colorado
Connecticut
District of Columbia
Florida
Indiana
Kentucky
Mississippi
Missouri
New Mexico
New York
North Dakota
South Dakota
Tennessee
West Virginia
Wisconsin

Measured Progress
Maine
Massachusetts
Montana
Nevada
New Hampshire
Utah
Vermont

Questar Education System
Arkansas
Minnesota

Pearson Education Measurement
Minnesota
New Jersey
Texas
Washington

Northwest Evaluation Association
Idaho

Data Recognition Corp.
Alaska
Pennsylvania

Source: The State of the K-12 State Assessment Market, April 2005, Eduventures Inc.

Uneducated Consumers

State departments of education are ultimately responsible for carrying out NCLB's testing mandates, and they are, if anything, in a weaker position than the testing industry to respond to the surge in testing under NCLB. Many state testing offices suffer from heavy turnover and shortages of skilled staff as a result of underfunding and hiring freezes introduced during the recession of the late 1990s. "The capacity of assessment and accountability offices in state departments of education is very low," says Marion of the Center for Assessment, who was testing director in Wyoming. "And there has been a significant increase in workload since the passage of NCLB, without an increase in qualified staff."

Education Sector found in a survey of state testing directors that it conducted for this report that over half the states have problems recruiting and retaining the testing staff they need to respond to NCLB adequately. Testing companies and large school systems, many say, are luring away scarce psychometricians and other key staff with higher salaries. Ohio's new testing director, Judy Feil, for example, has lost seven of 20 staffers in the past year to burnout and higher-paying jobs in school systems and testing companies. She has been forced to hire replacements without testing backgrounds, she says, because "skilled people are difficult to find." Many state testing offices operate with skeleton staffs. Indiana, a relatively large state, has five professional testing employees.

Testing directors themselves turn over at an alarming rate. Matt Gandal, executive vice president of Achieve Inc., a Washington, D.C.-based organization that promotes high education standards, says the testing directors of Ohio, Florida, Texas, Maryland, and Rhode Island have left in the past few years for the private sector and an opportunity to earn higher salaries. Marion describes standing with six or seven state testing directors at a national conference, only half of whom had been in their jobs for more than eight months.

The result of the understaffing and lack of expertise in many state testing offices, says Marion, is that, "for the testing companies, it's like being auditors of their own work." Many state testing agencies simply don't have the capacity to scrutinize the work of their testing contractors closely.

Troubling Consequences

The mounting scoring errors and reporting delays that have resulted from the many challenges confronting the testing industry and state testing agencies as they struggle to respond to NCLB's testing mandates have tarnished NCLB's testing-based system of school accountability. They have created a public-relations problem.

But the lack of state oversight of testing contractors, the industry wide shortage of testing experts, and the many other problems that have plagued the spread of statewide testing under NCLB are also damaging the cause of standards-based reform in ways that don't make many headlines but are arguably more fundamental.

Testing experts say many statewide tests are not getting sufficient psychometric scrutiny to ensure that they accurately measure student and school performance under NCLB. "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 Marion of the Center for Assessment, who has taught test-making at the University of Maine. "But they aren't doing it." "In many cases," says Cook, "they are putting [test] items on the street they shouldn't."

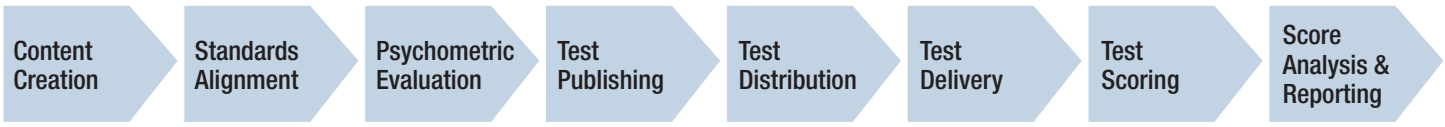
That's particularly true of test questions that require students to write a response rather than fill in a bubble on an answer sheet. The reason is that so-called open-response questions are more costly to field-test because they must be scored by people rather than machines. "You are paying a fortune on an individual item just to try it out," says Hoover, the onetime ITBS author, "so, frequently, companies never try them out and they are bad items." University of Iowa psychometrician Stephen Dunbar, Hoover's successor at the ITBS, refers to NCLB as "No Item Left Behind," because the law has led to such a shortage of quality test questions.

In another example of the consequences of psychometric failings, the Ohio Department of Education announced in the fall of 2005 that Measurement Inc. had failed to correctly translate raw scores on the state's high school test into scores on a publicly reportable scale. The "scaling" mishap resulted in new scores for 5,000 of the 5,400 students who had taken the test the previous summer, including 900 students who had been told they could not graduate because they had failed the test, when they hadn't.[11]

And while NCLB's strength as a source of standards-based school reform rests on its requirement that states measure students' grasp of statewide standards and then take steps to improve schools and school systems where students don't measure up, lack of time, money, and skilled staff have led a substantial number of states to introduce tests that many testing experts say are not fully aligned with state standards—tests that don't test what states expect their students to know. This is happening in part, experts say, because rather than building tests from scratch, states are hiring 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. "When you ask publishers if they align the tests with state standards, you'll rarely get an answer of less than '85 percent,' " says Marion. "But our studies show it's a lot lower, 50 percent. As a result, teachers are teaching stuff that they can't be sure is on the tests," because the tests don't necessarily measure the skills that states say teachers should teach.

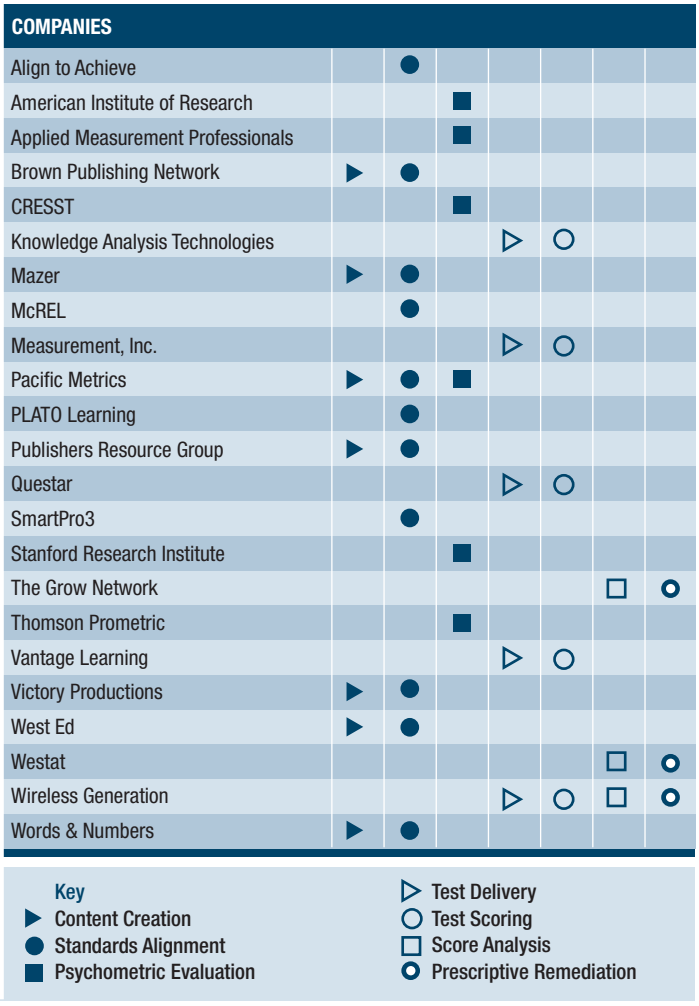
Nor is the quality of the many practice tests that students are taking in increasing numbers to prepare for NCLB testing what it should be. Formative testing has the potential to help students by giving teachers frequent and, in theory, useful, information on student performance. But so far, say industry analysts, schools and school systems have been unwilling to pay for high-quality test items for these new tests, leading testing companies to focus resources on supplying the new market with banks of test questions that are less fully field-tested and thus less expensive—but are also less accurate measures

The Work of the Testing Industry



Source: The State of the K-12 State Assessment Market, Eduventures, Inc., 2005

Niche Players



Source: Staying Ahead of the Curve, Eduventures, Inc., 2004

of student performance. "Certain firms claim to offer tens of thousands of exam items," Eduventures writes in an industry report, *Testing in Flux*. But many of the items, says Eduventures, "have not undergone rigorous psychometric evaluation."[12] Says Marion, "The items that end up on most of these formative tests are ones that get rejected from state tests." As a result, many formative test questions don't accurately measure what students know.

But the use of such test items is increasingly widespread. In 2003, in one of a number of moves by major testing companies to tap into the formative-testing market, Houghton Mifflin bought Edusoft, a then-three-year-old company that permits teachers to give tests, scan student answer sheets, upload them to Edusoft servers, and receive detailed score reports. School systems began asking Houghton to supply banks of practice test items. But Houghton couldn't deliver fully field-tested items for what many school systems were willing to pay for them. "School districts don't appreciate, or can't afford, high-quality items and tests," says Alvaro Fernandez, a former Edusoft executive. "They have an insatiable hunger for inexpensive item banks that their teachers can use to help them do better on the NCLB tests." So Houghton struck a deal with FS Creations, an Ohio-based company, to market low-end questions alongside its higher-quality Riverside products, says Fernandez.

Simple Questions

Perhaps the most troubling classroom consequence of the tumult in the testing industry is the strong incentive the problems have created for states and their testing contractors to build tests that measure primarily low-level skills.

NCLB has sought to lift the level of teaching in the nation's classrooms by requiring states to set challenging standards for what students should know and be able to do. But testing experts say that many of the tests that states are introducing under NCLB contain many questions that require students to merely recall and restate facts rather than do more demanding tasks like applying or evaluating information, largely because it's easier and cheaper to test the simpler tasks.

Such test questions do have a role; it's important that students' grasp the most basic skills. But because teachers have so much riding on their students' results, tests that stress such skills encourage teachers to emphasize them in their classrooms at the expense of the high standards that NCLB has sought to promote. They strip teachers of the incentive to teach higher-level skills. "Tests are focusing more and more on rote skills because it's difficult, given the demand that they be constructed quickly and cheaply, for anything else to happen," says Hoover. "Writing items that tap higher levels of comprehension is really difficult. The problem is that tests of rote skills encourage rote teaching. It's not a good model for instruction." As Marion puts it, "The further away we get from testing the types of things we want kids to do in school, the less likely we are to improve education."

Such tests also give a skewed sense of student achievement. Scores on reading tests that measure mainly literal comprehension are going to be higher than those on tests with a lot of questions that require students to

evaluate what they've read by, say, reading two passages and identifying themes common to both. The same is true in math. In a study by Lorrie Shepard, a testing expert and the dean of the school of education at the University of Colorado–Boulder, 85 percent of third-graders who had been drilled in computation for a standardized test picked the right answer to 3 x 4, but only 55 percent answered correctly when presented with three rows of four Xs.[13] Many of the new NCLB tests, as a result, are likely to suggest that students are achieving at higher levels than they really are. The tests have the potential to create glass ceilings for higher-achieving students, who have less of an opportunity to demonstrate the extent of their abilities. And when the scores of low-achieving students rise, this achievement ceiling could create the sense that performance gaps among groups of students are closing, when in fact they may not be.

It is possible to write multiple-choice questions that measure higher-level skills. But doing so is difficult and time consuming. "It's almost always easier to pick out factoids," says Marion. And when they're given a choice, most testing experts would measure students' grasp of more advanced abilities through open-ended or constructed-response questions that require students to produce their own answers rather than select answers from among those supplied by test-writers, the format of multiple-choice questions. "Constructed-response questions give you more measurement depth," says John Olson, director of psychometric and research services at Harcourt Assessment and director of assessment at the Council of Chief State School Officers from 1998 to 2003. "They give you a better sense of what students can do. And as a result, teachers get more out of them."

But such questions are more expensive and slower to process than their multiple-choice counterparts. Multiple-choice answer sheets, with their rows of "bubbled-in" circles, are scored in seconds by optical scanners. Harcourt's Galt says the company has scored 1 million Florida multiple-choice tests in a day. But grading constructed-response questions, where students write out answers, is more complicated and time consuming.

States and their testing contractors must first establish "rubrics," or standards, for judging students' responses, since there are often no "right" answers to such questions. 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 a week training their test graders to ensure that answers of comparable quality receive the same scores from different graders, most of whom are moonlighting or retired educators earning anywhere from $9 to $25 an hour. 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 in Dover, N.H.; Iowa City; Minneapolis; Durham, N.C.; Monterey, Calif.; and elsewhere, working their way through hundreds of answers at a rate of 20 to 30 items per hour. The result is that it costs anywhere from 50 cents to $5 to score a constructed-response question, compared with pennies per multiple-choice question, says Cook.

States Reporting Testing Problems

52%	**Capacity**—Difficulty recruiting and retaining qualified staff for testing-related positions
35%	**Accuracy**—Experienced a significant error by a contractor in scoring a state test since 2000
20%	**Timeliness**—Did not receive test results from a contractor in a timely fashion since 2000

Source: Survey of State Testing Offices, Education Sector, 2006, based on responses from 23 states

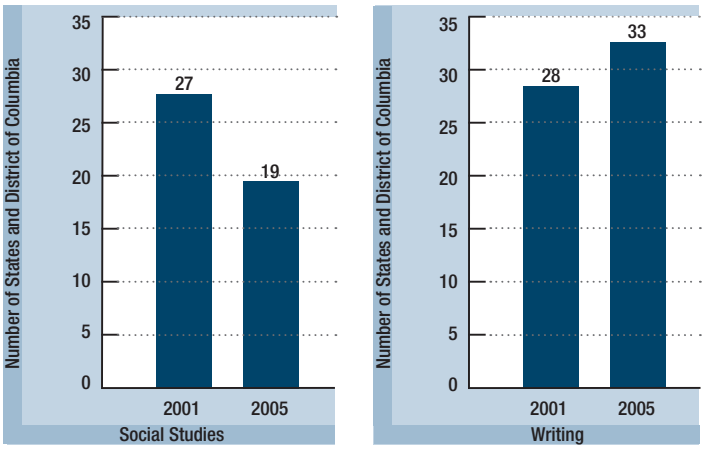
And the cost differential is not lost on the state legislators who control state education department budgets. In 2004, Pearson brought the membership of the Michigan House and Senate education committees to Iowa City to tour the company's high-tech facility for scoring multiple-choice answer sheets.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

The "efficiency" of multiple-choice questions also works against open-ended questions in another way. The reliability of a test, its ability to accurately gauge that a strong student is a strong student, increases with the number of questions it has, and students are often able to answer multiple-choice questions faster because it's quicker to fill in a bubble than write out an answer. So there can be many more multiple-choice questions than open-ended questions on a two-hour test.

As a result, there are very few open-ended questions on many of the new NCLB tests, say testing experts. "There's

pressure to use fewer of them," says Pearson's Twing. "There are hardly any in grades three through eight in many states and there are just a handful at the high school level—three or four out of 40, and they tend to be short answers that can be quickly answered and easily scored." "States are shifting from constructed response to multiple choice due the cost and time of scoring constructed-response questions; we are seeing more of that," agrees Olson, Twing's counterpart at Harcourt Assessment, who spent a decade working on the federally funded National Assessment of Educational Progress (NAEP), considered to be one of the nation's most sophisticated tests with many questions measuring higher-level skills. "During the

Number of States Testing Subjects Not Required by NCLB



Source: The Impact of NCLB on Non-Tested subjects, Patricia V. Pederson, Harcourt Assessment, Inc., 2005

1990s states had 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 eliminated non-multiple-choice questions from its state tests in 2005-06. So did Kansas. In all, 15 states serving 42 percent of the nation's students are using NCLB reading and math tests in 2005-06 that have no open-ended questions, writes *Education Week*.[14] NCLB's requirements for more testing in reading, math, and, by 2006-07, science have also led states to cut back the use of testing to drive improvements in other subject areas. "There has been so much focus on math and reading that states are cutting back on other subjects, social studies in particular," says Olson. "They can't afford to do both," says Feil, Ohio's testing director.

Research by Patricia Pederson, a former director of science and social studies testing at Harcourt Assessment, reveals that the number of states testing in science rose from 34 in 2001, prior to the enactment of NCLB, to 40 in 2005, in anticipation of the NCLB's 2006-07 science-testing requirement. The number of states testing writing rose from 28 to 33. But the number testing social studies declined from 27 to 19 during the same period.[15] Because schools tend to teach what's tested when test scores have consequences for teachers and principals, the cutbacks in testing in such subjects as social studies are encouraging schools to focus their energies on reading, math, and, increasingly, science, in the same way that the large number of multiple-choice questions requiring only rote responses on the new state tests is leading them to focus on the lowest-level skills within reading and math.

There's clearly value in pressing educators to ensure that students have a strong grounding in the building-block subjects of math and reading. Such grounding is one of NCLB's primary goals. But the cost has been a narrowing of the curriculum in many schools and classrooms. And Hoover, for one, calls this marginalization of history, art, and music a huge downside of NCLB's testing requirements and the overwhelming demands they have placed on state testing agencies and testing companies.

The Low-Cost Track

Congress has sought since the implementation of NCLB to help states meet the financial burden of the law's testing mandates. It is giving them $412 million for statewide tests in 2005-06 and $408 million in 2006-07, about $9 a test. A provision in NCLB championed by the late Minnesota Sen. Paul Wellstone suspends the law's student testing requirements if federal testing aid falls below specified levels—$400 million in 2005-06.

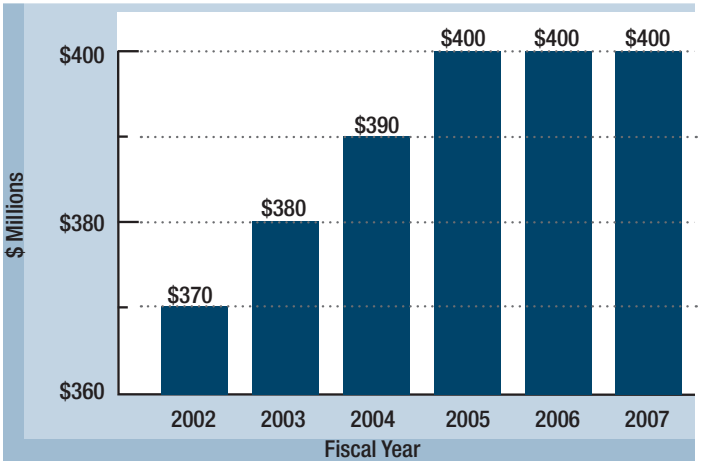
But states can spend the money on a range of tasks unrelated to test-building, such as "improving the dissemination of information on student achievement and school performance," and Eduventures reports that "states in many cases have opted to allocate most of those [federal] dollars to the development of state standards or initiatives aimed at aligning instruction to state standards"—important activities, but not the work of state testing agencies.[16] In some instances, testing experts say, state lawmakers are using the federal testing money to supplant rather than supplement their state

testing budgets. "If I'm a Senate education committee chairman and you tell me you've got $10 million in federal testing money, my response is, 'That's real simple, I'll knock $10 million from your testing budget,' " says Cook, the former Wisconsin testing director. There is supplement-not-supplant language in NCLB regarding testing funding, but such provisions are notoriously difficult to enforce.

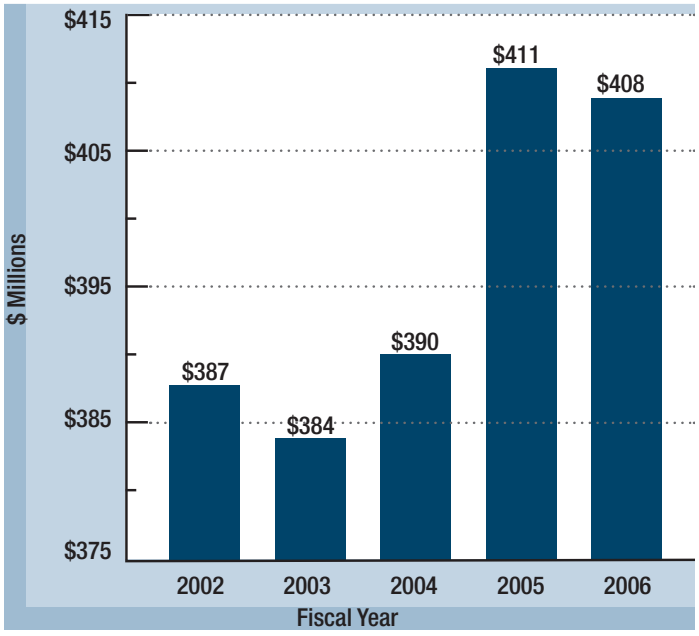
Connecticut has had high-quality tests with many open-ended questions since the 1980s, including math questions that require students to write explanations of their answers. In 2005, the state sued the U.S. Department of Education over the cost of NCLB testing, saying that Connecticut's share of the congressional appropriation is inadequate to fund tests of the same caliber under NCLB. U.S. Secretary of Education Margaret Spellings seemed to suggest as much in a letter to Connecticut's commissioner of education, Betty Sternberg, in response to the state's legal action. "Some of the costs of Connecticut's testing system are attributable to state decisions [regarding the types of tests it uses]," Spellings wrote. "While these decisions are educationally sound, they go beyond what was contemplated by NCLB." [17]

Federal Spending on State Testing

Minimum Required by NCLB



Actual Spending



Source: The State of the K-12 State Assessment Market, Eduventures, Inc., 2005

The then-General Accounting Office, a research arm of the U.S. Congress, more or less predicted the confrontation between Connecticut and the U.S. Department of Education when, in a 2003 report, it produced three widely varying estimates of what it would cost states to comply with NCLB's testing mandates. It would cost them $1.9 billion between 2003 and 2008 if they used tests with only multiple-choice questions, the agency suggested. The price would rise to $3.9 billion if their tests used multiple-choice questions and some open-ended items. And it would reach $5.3 billion to build tests with a larger percentage of open-ended questions.[18]

But the agency didn't predict the consequences on the testing industry, state testing agencies, and the nation's classrooms of the low-cost track that federal and many state appropriators have followed and that Secretary Spellings has sought to defend. Given the tremendous influence of statewide standardized testing on public schools today, NCLB's pursuit of significantly higher levels of student achievement requires 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 teachers and principals with timely information on students strengths and weaknesses.

That means investing resources commensurate with testing's central role in school reform today. Together, steps such as increased federal spending and targeted research and development activities, the establishment of a national testing oversight body, and incentives for greater inter-state collaboration would strengthen the national testing infrastructure dramatically, allowing it to support the full weight of standards-based reform.

RECOMMENDATIONS

State and federal policymakers can address the problems raised in this report by enhancing federal leadership on testing issues, creating an independent national oversight body to promote test quality, giving states incentives to collaborate on test development and, ultimately, developing voluntary national assessments.

Federal Leadership

The federal government should take several steps to improve the nation's education testing infrastructure. First, it should greatly increase the supply of well-trained psychometricians and other testing experts. The U.S. Department of Education should fund the training of 1,000 such specialists over the next five years, through grants to support students who commit to working in the field after they've completed their degrees. The Education Department should also fund university-based research on test quality. Doing so would help build university capacity to train more testing experts and give professors in psychometrics and related fields incentives to recruit students to serve as research assistants, while promoting needed research and development in testing.

Second, federal funding for testing under NCLB should be increased from its current level of $408 million to $860 million annually. This would help give all states the resources necessary to develop tests on par with those in states that currently have the strongest testing programs. Massachusetts's standards and tests, for example, have been widely praised.[19]

The state's custom-designed statewide tests include a healthy mix of high-quality multiple-choice and open-ended questions that require students to construct answers. The state has also developed a large network of in-state teachers to write enough new test questions that every test can be made public after students have taken it to help parents, educators, and the public understand the process. Massachusetts currently supports a testing staff of 30 that monitors the scoring of the state's tests at contractor's centers across the nation, ensures that open-ended questions are appropriately benchmarked and scored, and conducts training sessions for teachers throughout Massachusetts to ensure that test administration goes smoothly.

But Massachusetts's $8 million share of the federal government's current funding for statewide testing under NCLB covers only 30 percent of the state's NCLB testing costs, says Jeff Nellhaus, associate commissioner of the Massachusetts State Department of Education and the state's testing director from 1994 to 2002. Most states don't fund their testing programs as generously as Massachusetts does. Increasing federal funding for statewide testing to $860 million would allow the federal government to provide every state two-thirds of the funding necessary to reach Massachusetts's level of per-student spending on NCLB-required testing.[20] This support would balance federal and state educational responsibilities and help ensure that NCLB fosters a race to create high-quality assessments, not a race to the bottom.

Such an investment is only a small fraction of the approximately $500 billion spent on public education in 2005-06 and is actually a minimal investment, given the key role that statewide testing plays in NCLB and standards-based reform generally. In fact, most industries would clamor for the opportunity to have a quality-control and feedback mechanism that costs less than 1 percent of their overall operating revenue. "Given the [state testing] systems' low costs," says Hoxby, the Harvard University economist, "we ought not hesitate to improve them."[21] Because statewide standardized testing has such a strong influence on teaching and learning in the nation's classrooms today, there should be sufficient funding to create and administer very high quality tests. Given the stakes, even pretty good tests are not good enough, much less the tests in place in many states today.

At the same time that the federal government invests more in testing, it should demand more, too. The federal government should make certain that strict supplement-not-supplant provisions ensure that new investments support the development of high-quality tests rather than displace existing state funding. In addition, Washington should ensure that states are using assessments that are well aligned with state standards and based on sound and clearly articulated definitions of what students should know and be able to do.

Third, the federal government should fund new research and development on testing. Though competition in the testing industry has led companies such as Harcourt Assessment to invest in equipment to improve the efficiency of their test processing, the pressure on profit margins has made investments in testing research and development difficult.

The federal government could make an important contribution by stepping in to support research and development on testing, catalyzing new ideas and technologies. It could fund research into ways to make the use of open-ended questions less expensive or ways to more accurately test students with disabilities who are able to participate in the general curriculum. Similarly, more sophisticated assessments for English-language learners would help educators more accurately pinpoint the needs of those students.

Another area needing research is online testing, which offers the promise of more customized assessments, a shorter and simpler testing and scoring process, and more comprehensive reports on student achievement quickly. Online testing also offers the possibility of states incorporating new types of test questions in their statewide exams, including "drag and drop" items and simulations. And it creates the potential to identify students' skills and weaknesses using fewer test items through so-called computer-adaptive testing, which adjusts the difficulty of test questions to students' performance on previous questions.

There have been significant advances in these areas, but a host of technical and financial challenges remain to be solved before the technology can be successfully applied to statewide testing programs, says Kahl of Measured Progress, which has worked with Maine and other states in implementing online testing. "Half the states are asking for it," he says, "but there are bugs galore." More research and development would help to address the problems facing technology-based testing.

Effective Oversight

Improving the nation's fragmented testing system will require strong, effective federal leadership. **President Bush should begin by inviting the leaders of both parties in both branches of Congress to work with him to name leading experts to a bipartisan presidential commission on standardized testing.** That group, comprising state testing officials, testing industry representatives, and independent testing experts, would study a range of testing-related issues, including funding, testing quality, industry capacity, and human capital.

As part of its work, the commission should establish an independent national testing oversight agency. Despite the enormous importance of statewide standardized testing in public education today, the tremendous potential for problems in test development and test administration under NCLB testing, and the inability of many states to supervise their testing contractors effectively, there is no entity to independently audit state testing programs and the testing industry. Such a body would perhaps be called the National Testing Quality Commission and would operate in the spirit of the Consumer Product Safety Commission and other federal consumer-protection agencies.

The U.S. Department of Education requires states to submit their NCLB accountability plans for approval through "peer review," a process that evaluates the way in which states propose to establish their academic standards, align their tests to the standards, and ensure the technical quality of their tests. But the department's peer review system does not audit the quality of test items or the performance of state testing contractors.

While professional associations give guidance on the technical quality of standardized tests, they do not play an auditing role. An independent National Commission on Testing and Public Policy over a decade ago wrote that, "Although the American Psychological Association, the American Education Research Association, and the National Council on Measurement in Education have formulated professional standards for test development and use in education … they lack any effective enforcement mechanism." [22]

The U.S. Department of Education's Office of Inspector General seems to agree. Part of its study of scoring problems on state tests will be to determine "whether there is a need for federal oversight to help ensure [that] errors in scoring high-stakes tests are prevented, detected, and disclosed publicly." [23]

Interstate Collaboration

The inefficiency of states and the District of Columbia administering 51 separate testing programs is obvious. **States could create higher-quality tests at lower cost if they worked together to develop common tests.**

Three states—New Hampshire, Rhode Island, and Vermont—have done this. Following the enactment of NCLB, they formed the New England Common Assessment Program. The states' testing officials worked with Measured Progress and the Center for Assessment to build reading and math tests in grades three through eight that students in the three states took for the first time in fall 2005. The consortium produced tests with an ample number of open-ended questions at two-thirds of what the states would have paid had they developed the tests individually, says Kahl. Each state spent $2 million on the project rather than the $3 million they would have each spent had they not collaborated. There is a particularly strong incentive for smaller states to create testing consortia, because it costs nearly the same to develop a test in a small state like South Dakota as it does in a large state like Texas. The savings for small states are likely to be greater per student because with lower enrollments their development costs are higher on a per-student basis. But larger states would also benefit financially from such collaborations, and even the largest would benefit from the greater efficiencies that would result from multi-state contracts for test administration, scoring, and reporting.

States entering testing consortia would have to agree on test content, as Vermont, New Hampshire, and Rhode Island have. But despite the political mantra of local control, state standards in reading and math do not vary that much from state to state now. This means partnerships like the New England Common Assessment Program are both politically and practically possible. **The federal government should encourage the creation of state testing consortia by offering states that elect to work together additional funding to support their assessment programs under NCLB.**[24]

In the long run, the logic of regional consortia leads to a solution that already exists in many of the industrialized nations of Europe and Asia: a single national testing system. By encouraging states to arrive at such a system through voluntary collaboration the federal government could support these efforts while sidestepping some of the thornier political questions about national testing. The closer we get to such a system, the more the nation's overextended testing infrastructure could focus on creating much smaller numbers of much higher quality assessments and test items. Moreover, it would free up resources to build strong national tests that states could voluntarily adopt in important disciplines such as social studies and writing, subjects that don't benefit from consistent measurement under NCLB.[25]

The concept of national testing is, of course, highly controversial. President George H.W. Bush proposed the creation of voluntary national standards in elementary and secondary education in the early 1990s, and the Clinton administration sought voluntary national tests in the mid-1990s. Partisan politics played a role in the defeat of both initiatives. But there was also strong philosophical opposition from both the left and the right. Liberals argued that the tests would be unfair to students in impoverished communities, while conservatives argued that national standards and testing would amount to a federalization of public schooling.

But although there is certainly not consensus support for national tests, the politics of the issue have changed dramatically since it was last seriously considered in Washington. By mandating statewide testing throughout the country, NCLB, a law proposed by a conservative Republican president and passed by a Republican-controlled Congress, has imposed a much greater degree of centralization in standards-setting and testing than existed previously. NCLB also expanded federal authority in education. Meanwhile, prominent liberal advocacy groups, including the Citizens' Commission on Civil Rights and the Education Trust, have been vocal supporters of statewide testing under NCLB. And both conservative voices in education such as the Thomas B. Fordham Foundation and liberal organizations such as the Center for American Progress have recently endorsed national standards and testing in public education. The early success of the New England Common Assessment Program is likely to further soften opposition to common standards and common tests.

ENDNOTES

Report Endnotes

[1] Editorial Projects in Education Research Center, as cited in *Education Week*, Nov. 30, 2005; Education Sector calculations based on data from U.S. Department of Education, National Center for Statistics.

[2] For further reference please see the following: Robert A. Frahm, "State Fires Testing Contractor," *Hartford Courant*, Sept. 13, 2005; Doug Guthrie and Christine MacDonald, "Some Schools Await MEAP Tests," *The Detroit News*, Oct. 5, 2005; Paul Tosto, "Analysis of State Education: Reliance on Testing Poses Many Pitfalls," *St. Paul Pioneer Press,* March 13, 2004; Frank Rich, "Testing Co. Mistakenly Fails Ohio Students," Associated Press, Dec. 12, 2005; Kathleen Rhoades and George Madus, *Errors in Standardized Tests: A Systemic Problem* (Boston: National Board on Educational Testing and Public Policy, 2003).

[3] *Office of Inspector General FY 2006 Work Plan* (document available by search on U.S. Department of Education Web site: www.ed.gov).

[4] J. Mark Jackson and Eric Bassett, *The State of the K-12 State Assessment Market* (Boston: Eduventures, 2005).

[5] President's FY 2006 Budget Request for the U.S. Department of Education, Appendix 3: Total Expenditures for Elementary and Secondary Education in U.S. (Published online at http://www.ed.gov/about/overview/budget/budget06/summary/edlite-appendix3.html).

[6] Editorial Projects in Education, "Quality Counts 2001: A Better Balance," *Editorial Projects in Education* Vol. 20: No. 17, January 2001 (Published online at http://counts.edweek.org/sreports/qc01/articles/qc01story.cfm?slug=17toc.h20).

[7] *Doctorate Recipients from United States Universities:Summary Report 2004* (sponsored by the National Science Foundation, the National Institutes of Health, the U.S. Department of Education, the National Endowment for the Humanities, the U.S. Department of Agriculture, and the National Aeronautics and Space Administration; published online at http://www.norc.uchicago.edu/issues/sed-2004.pdf).

[8] *From the Capital to the Classroom: Year Three of the No Child Left Behind Act* (Washington, D.C.: Center on Education Policy, March 2005).

[9] Caroline M. Hoxby, "The Cost of Accountability," *NBER Working Papers 8855*, (Cambridge, Mass.: National Bureau of Economic Research Inc, 2002).

[10] Jackson and Bassett, *The State of the K-12 State Assessment Market*.

[11] Vicki Lee Parker, "Measurement Inc. Slips Down a Notch; After a Goof on Ohio Students' Tests, Company Loses No-errors Bragging Rights," *The News & Observer* (Raleigh, N.C.), Dec. 21, 2005.

[12] Matt Stein, *Testing in Flux: Future Directions in the Pre-K-12 Assessment Market* (Boston: Eduventures, 2004): 12.

[13] L.A. Shepard, *Measuring Achievement: What Does it Mean to Test for Robust Understandings?* (Princeton, N.J.: Policy Information Center, Educational Testing Service, 1997).

[14] *Education Week*, Nov. 30, 2005.

[15] Patricia V. Pederson, *The Impact of NCLB on Non-Tested Subjects: An Assessment Developer's Perspective* (Session 1108T, presented at Voices of Education: Unleashing the Power, Passion and Promise, ASCD Annual conference and Exhibit Show 2005, Saturday, April 2, 2005).

[16] Jackson and Bassett, 18.

[17] Margaret Spellings, the secretary of education, to the Honorable Betty J. Sternberg, commissioner of education for the State of Connecticut, May 3, 2005 (Published online at http://www.state.ct.us/sde/nclb/Correspondence/SpellingsLettertoBetty5-3-05.pdf).

[18] United States General Accounting Office, *TITLE I: Characteristics of Tests Will Influence Expenses; Information Sharing May Help states Realize Efficiencies* (USGAO, GAO-03-389, May, 2003): 15.

Recommendation Endnotes

[19] *Measuring Up: A Report on Education Standards and Assessments for Massachusetts*, Published online at http://www.achieve.org/achieve.nsf/StatePro-Massachusetts?OpenForm. (Achieve, Inc., 2001).

[20] The proposed $860 million amount was calculated by multiplying Massachusetts' estimated per-student test cost of $25 by the twice the number of public elementary and secondary students in grades 3-8 and 10 (to account for the cost of two tests, one in reading and one in math), and then multiplying that result by two-thirds.

[21] Caroline M. Hoxby, "The Cost of Accountability," *NBER Working Papers 8855*, (Cambridge, MA: National Bureau of Economic Research, Inc, 2002).

[22] Kathleen Rhoades and George Madus, *Errors in Standardized Tests: A Systemic Problem* (Boston: National Board on Educational Testing and Public Policy, 2003): 8.

[23] *Office of Inspector General FY 2006 Work Plan* (document available by search on USDOE website: www.ed.gov).

[24] Andrew J. Rotherham, "Asking The Wrong Test Questions," *The Washington Post*, May 29, 2001, A15.

[25] Thomas Toch, "Bush's Big Test," *The Washington Monthly*, November, 2001.

ABOUT THE AUTHOR

Thomas Toch is co-founder and co-director of Education Sector. He can be reached at ttoch@educationsector.org.

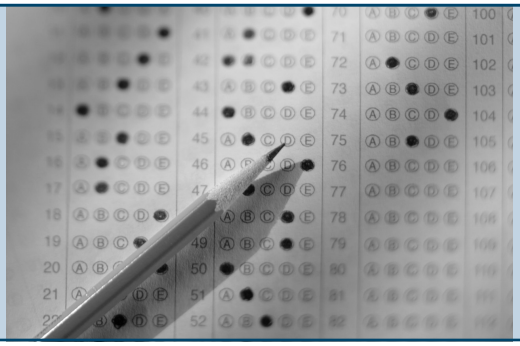
ABOUT EDUCATION SECTOR

Education Sector is an independent education think tank based in Washington, D.C. It is a non-profit and non-partisan organization devoted to developing innovative solutions to the nation's most pressing educational problems. The organization seeks to be a dependable source of sound thinking on education policy and an honest broker of evidence in key education debates in Washington and nationally.

ACKNOWLEDGEMENTS

Thanks to my Education Sector colleagues Andrew Rotherham, Kevin Carey, Sara Mead, and Bill Tucker for their thoughtful comments on drafts of the report. Kevin and Sara also surveyed state testing offices for the report. Sharon Cannon, Renée Rybak, Ethan Gray, and Lisa Guido, provided valuable research support, and Sharon managed the report's production. Susan Vavrick copy-edited the report, Alison Dacher designed it, and Molly Norton managed our communications effort. Thanks also to Scott Marion of the Center for Assessment for reviewing a draft of the report and contributing to the glossary of testing terms and the chart describing how a test item is created. Gary Cook of the Wisconsin Center for Education Research was also kind enough to read a draft of the report. Eric Bassett of Eduventures, Inc., graciously permitted Education Sector to reproduce charts and graphs from Eduventures research reports. David Hackensen of Pearson Educational Measurement, Mark Slitt of Harcourt Assessment, and Tom Ewing of the Educational Testing Service helped set up interviews and track down information. I am particularly grateful to the nearly three dozen testing industry executives, state testing officials, and other testing experts who shared their candid assessments of the testing industry for this report.

EDUCATION**SECTOR** REPORTS



MARGINS OF ERROR:

The Education Testing Industry in the No Child Left Behind Era

by Thomas Toch

Table of Contents

1101 Pennsylvania Ave, N.W., Fifth Floor, Washington, DC 20004
202.756.4944 • www.educationsector.org

State standards and standardized tests have become dominant forces in American public schooling. For most of its history, public education in the U.S. was a local matter, with local schools and school systems setting their own educational priorities. But in the wake of mounting evidence that the preparation most students received from public schools wouldn't suffice in a postindustrial economy, and with the conscience of the nation having been transformed by the civil rights movement, policymakers began to pursue a new paradigm, one that sought to establish statewide public school standards and hold local educators accountable if their students fell short of these standards. Standardized tests, used to measure student performance against the new state expectations, are the linchpin of this strategy of standards-based reform.

The No Child Left Behind Act of 2001 (NCLB) solidified standards-based reform as a national priority, part of a bold attempt by federal policymakers to force state and local educators to improve the education of minorities and other students that public schools traditionally hadn't served very well. The legislation required that by the spring of 2006 states test nearly every public school student in grades three through eight and in one high school grade to gauge whether students have met standards in reading and math—a task requiring some 45 million standardized tests annually.

To comply with NCLB, 23 states that have not yet fully implemented the law's testing requirements will administer some 11.4 million new tests during the 2005-06 school year alone, half in reading, half in math. Within two years, states must begin testing students in a minimum of one elementary, middle, and high school grade in science under NCLB, requiring at least another 11 million tests.[1]

Standardized test scores form the basis of NCLB's accountability mechanisms—school report cards, tutoring and school-choice options for students, and serious consequences for low-performing schools. Increasingly, as a result, the content of statewide tests has become the focus of teaching and learning in public school classrooms throughout the nation, to the point where many schools have begun to do much more testing than is required by NCLB, in an effort to prepare their students for the high-profile NCLB-mandated exams.

But this surge in testing has created immense challenges for both the industry that writes, scores, and reports the vast majority of the new statewide tests and the state agencies charged with carrying out NCLB's requirements. NCLB's test-based accountability system has given local educators powerful incentives to help students whom public education has long neglected. But the scale of the NCLB testing requirements, competitive pressures in the testing industry, a shortage of testing experts, insufficient state resources, tight regulatory deadlines, and a lack of meaningful oversight of the sprawling NCLB testing enterprise are undermining NCLB's pursuit of higher academic standards.

Symptoms of the turmoil in the testing industry aren't difficult to find: Newspapers carry accounts of testing companies giving students college scholarships to atone for the fact that scoring errors deprived them of their high school diplomas; of scoring errors sending thousands of students to summer school when they had in fact passed their tests; of months-long scoring delays; of administrators losing their jobs for low scores on tests that, had they been scored correctly, would have shown improvements in student achievement.[2]

These problems have damaged the credibility of standards-based reform in the eyes of many educators and parents, and they have attracted the attention of the Office of Inspector General at the U.S. Department of Education, which has announced plans to examine the extent of test-scoring and reporting mistakes under NCLB.[3]

But there are deeper, more structural problems stemming from the tremendous expansion of statewide standardized testing that haven't made headlines in the buildup to the full implementation of NCLB's testing requirements in spring 2006. Many states are constructing tests that don't fully measure student and school performance against state standards. And they are using tests that measure mostly low-level skills, a move that encourages teachers to make the same low-level skills the priority in their classrooms at the expense of the higher standards that NCLB has sought to promote.

The testing infrastructure that undergirds NCLB's accountability system must be improved, as this report makes clear, and if steps aren't taken to do so, teachers and principals will lose valuable tools to improve instruction, and both NCLB's work on behalf of public education's neediest students and standards-based reform itself will be increasingly at risk. Statewide testing, envisioned under NCLB as a key part of the solution to what ails public schools, is fast becoming part of the problem in public education.

States Adding Reading and Math Tests in 2005-06

STATE	TOTAL TESTS	STATE	TOTAL TESTS
Connecticut	311,286	New Jersey	619,588
Illinois	974,160	New York	1,704,592
Kansas	315,138	Ohio	850,850
Kentucky	443,828	Oklahoma	190,066
Maine	126,474	Pennsylvania	864,686
Massachusetts	456,750	Rhode Island	150,796
Michigan	1,062,907	Vermont	89,104
Minnesota	383,214	Virginia	565,096
Missouri	566,802	Washington	629,156
Montana	90,602	Wisconsin	512,240
Nevada	188,358	Wyoming	59,147
New Hampshire	198,032		
TOTAL			11,352,872

Source: Editorial Projects in Education Research Center; National Center for Education Statistics; Education Sector calculations

The Industry

The testing industry is surprisingly small, given its outsize role in public education today. Eduventures Inc., a Boston-based research firm, estimates that the value of tests, testing services, and test-prep materials purchased in 2006 will be $2.3 billion. But that includes purchases by school systems and schools, state-level testing, and college-admissions testing and test prep. Total expenditures for developing, publishing, administering, grading, and reporting NCLB-required statewide tests, Eduventures estimates, will be $517 million in the 2005-06 school year.[4] Some testing company executives peg the number somewhat higher, at $700 million to $750 million, still a small portion of the approximately $500 billion the United States spends on public elementary and secondary education annually.[5]

A handful of companies capture some 90 percent of the statewide testing revenue, Eduventures estimates. They include Pearson Educational Measurement, a subsidiary of London-based publisher Pearson PLC; CTB/McGraw-Hill, a division of the New York-based publishing and information conglomerate McGraw-Hill Cos.; Harcourt Assessment Inc., owned by Anglo-Dutch publishing giant Reed Elsevier; Riverside Publishing, a division of the privately owned publisher Houghton Mifflin Co.; and the nonprofit, Princeton-based Educational Testing Service (ETS), best known as maker of the SAT college-admissions test. They are "full-service" companies that create tests; align them with state standards; ensure they are technically sound; publish, distribute, and score them; and analyze results.

Other, smaller full-service companies have entered the statewide testing business more recently, including Measurement Inc., Questar Educational Systems, Data Recognition Corp., and non-profits Measured Progress, Northwest Evaluation Association, and American Institutes of Research. And there is a growing number of niche companies that focus on aspects of the testing enterprise such as test-question writing or test scoring.

The Major Players

CTB/McGraw-Hill
Major Test: TerraNova
Activities: Development, Administration, Scoring, Reporting
State Testing Contracts: 23
K-12 Tests Administered, 2005: 16.5 million
Test Items Written, 2005: 167,000
Percent of Nation's Students Taking a CTB test: 35
K-12 Testing Revenue: N/A
Corporate Parent: The McGraw-Hill Companies

Educational Testing Service
Major Test: None. Creates Custom-designed Statewide Exams
Activities: Test Development, Reporting
State Testing Contracts: 15
K-12 Tests Administered, 2005-06: 7 million
K-12 Testing Revenue: $150 million
Key Fact: Best Known as Creator of the SAT; New Player in K-12 Testing
Corporate Parent: Non-profit

Harcourt Assessment
Major Test: Stanford 10
Testing Activities: Development, Administration, Scoring, Reporting
State Testing Contracts: 22
K-12 Tests Administered, 2005-06: 9.5 Million
Test Items Written, 2004: Nearly 85,000
Open-ended Questions Scored, 2005-06: 40 Million
K-12 Testing Revenue: N/A
Corporate Parent: Reed Elsevier

Pearson Educational Measurement
Major Test: None. Creates Custom-designed Statewide Exams
Activities: Development, Administration, Scoring, Reporting
State Testing Contracts: 20
K-12 Tests Administered, 2005-06: 40 million
K-12 Testing Revenue: N/A
Key Fact: Nation's Largest Test Scorer
Corporate Parent: Pearson PLC

Riverside Publishing
Major Tests: Iowa Test of Basic Skills
Activities: Design, Development, Scoring, Assessment Management
State Testing Contracts: 4
K-12 Tests Administered, 2005-06: N/A
K-12 Testing Revenue: N/A
Key Fact: Top Player in Formative-assessment Market
Corporate Parent: Houghton Mifflin Company

Source: Testing companies

The Path to NCLB

The major players have been around for a long time. Harcourt's Stanford 10 test and CTB-McGraw Hill's TerraNova tests date to the 1920s, Riverside Publishing's Iowa Test of Basic Skills to the 1930s.

But in keeping with the tradition of local control in public education, publishers for decades sold their elementary- and secondary-school achievement tests only to schools and school systems, where they were used to compare local student performance with that of representative national samples of students—so-called norm groups.

The publishers' local sales staff sold the tests at the same time they sold textbooks, because the major publishers were in both businesses. There was thus a patchwork of different tests in place in every state rather than a single, statewide testing system. There were typically no consequences for local educators for their students' performance on the tests. And there wasn't any attempt to measure student performance against state standards, because with local school systems establishing their own educational agendas, there weren't any statewide standards.

All that began to change in the late 1960s. The Elementary and Secondary Education Act of 1965 (ESEA), of which NCLB is the latest reauthorization, called for evaluation of federal programs for disadvantaged students and set aside funding for the task. A nascent accountability movement also took shape in the 1970s, as state lawmakers, in the face of reports that many students weren't learning and demands that state officials address the problem, started to require statewide testing programs as a way of ensuring that students had "minimum competencies" in core subjects; they wanted to know, for example, that sixth-graders were performing at least as well as typical fourth- or fifth-graders. Michigan created the first statewide standardized testing program, in 1969, and Florida created the second, in 1971.

The minimum-competency movement expanded to other states during the 1970s, with lawmakers typically requiring testing at two or three grade levels each year. And the movement spread more rapidly in the early 1980s with the publication of *A Nation At Risk* and several other national studies that laid bare the troubled state of the nation's public schools. New funding for school reforms began to flow from state coffers in the wake of the reports, and

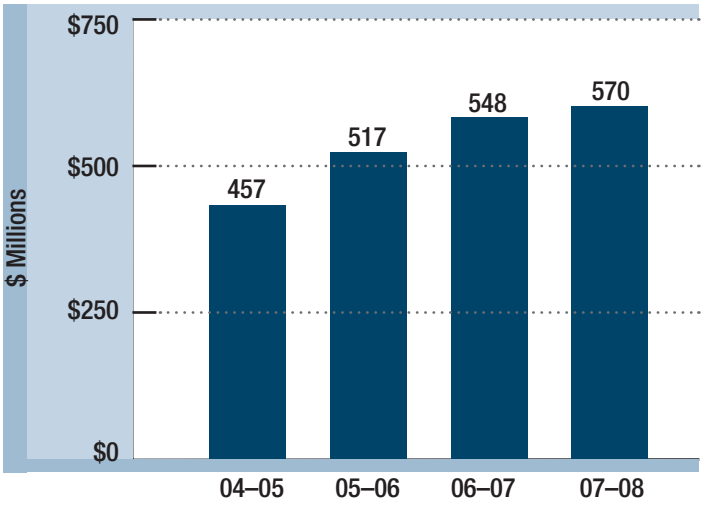
lawmakers wanted evidence that their investments were paying dividends, so they mandated more testing.

By the end of the 1980s, frustration with the pace of local reforms had led President George H.W. Bush to convene a national summit with the nation's governors in Charlottesville, Va., to explore ways to promote reform on a larger scale. Bush and the governors, led by then-Arkansas Gov. Bill Clinton, helped win bipartisan support for standards-based reform by establishing as a national goal that students demonstrate "competency" in core subjects such as English and math in grades four, eight, and 12. By the time Congress reauthorized the Elementary and Secondary Education Act in 1994, Clinton was in the White House and standards and accountability were the watchwords of reform.

Known as the Improving America's Schools Act (IASA), the Clinton administration's ESEA-reauthorization legislation required every state to put in place both standards and tests in reading and math at three grade levels, and about two-thirds of the states had done so by the end of the second Clinton presidency.[6] Most of the tests were designed to measure whether students mastered states' standards, rather than how they compared with students nationally; they were "criterion-referenced" tests rather than the "norm-referenced" tests that most states had introduced previously.

State Spending on Standardized Testing

Millions spent per academic year on the development, publishing, administration, grading and organizing of state exams.



Source: The State of the K-12 State Assessment Market, Eduventures, Inc., 2005

NCLB built on the Clinton-era accountability measures. It more than doubled the amount of testing required of the states, from three grade levels to seven; it established much tighter deadlines for introducing new tests; it required that results be broken down by a range of subgroups of students in every school; and, most significant, it linked serious consequences for schools to student test scores. Today, under NCLB, more students are tested more often than at any time in the nation's history, and the stakes are far higher.

'Harder Than the Dickens'

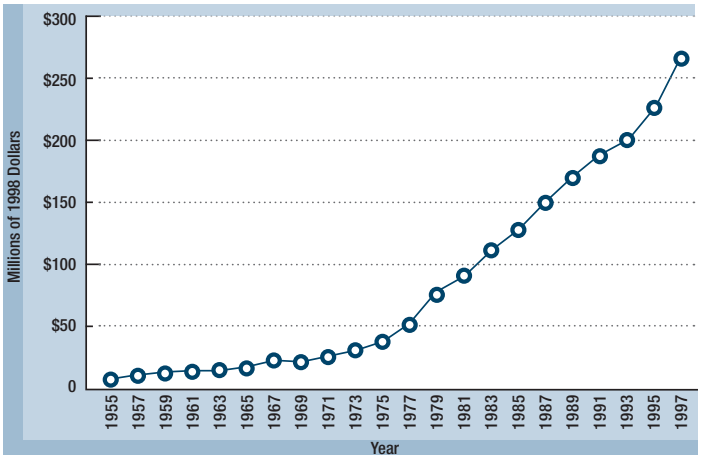
Creating high-quality tests is difficult and labor intensive. The process involves determining the length and content of a test, hiring curriculum experts to write questions, and ensuring that the questions align with state standards so the questions test what students are supposed to know. Then questions are field-tested on thousands of students to ensure that they don't discriminate against groups of students but do discriminate between strong and weak students, a complex mathematical task that requires comparing how students do on other questions with how they perform on the questions being trial-tested. Test-makers also have to ensure that every multiple-choice question has one, and only one, correct (or clearly best) answer and that the questions on a test reflect an appropriate range of difficulty. Another complex statistical computation has to be performed to ensure that the same scores on different tests represent the same level of performance. Then tests have to be edited, printed, and distributed to every public school in the country.

It's a demanding process under the best of circumstances. "Hundreds of people have to touch every item," says Gary Cook, a research scientist at the University of Wisconsin's Center for Education Research, who served as Wisconsin's testing director and as vice president of state accounts at Harcourt Assessment. The difficulty of writing questions that can clear testing's many hurdles has typically resulted in a majority of them being dropped—even those drafted by the most experienced item writers, says H.D. Hoover, who was the principal author of the Iowa Test of Basic Skills (ITBS) for nearly two decades. "Building a good test looks easy, but it's harder than the dickens," says Hoover. Ohio now administers 400 different forms of its statewide tests in order to field-test enough questions to keep its bank of test items stocked. As a result, it costs anywhere from $300 to $1,000 to develop a simple multiple-choice question, the least

expensive type of test item. State tests typically have 50 to 100 questions per subject per grade.

This complex test-making infrastructure is buckling under the weight of NCLB's testing demands. "There's way too much demand and not enough supply," says Hoover.

The Rise of Standardized Testing: Test Sales, 1955 to 1997



Source: The Bowker Annual of Library and Book Trade Information, 1970-98; Association of American Publishers, 1970-1998

When tests were purchased by schools and school systems, and in the early years of the standards movement, when most states used the Stanford and other major norm-referenced tests as their statewide exams, test publishers produced a new battery of tests only every six to eight years, because they didn't release test items and were thus able to recycle their tests. The result was a manageable demand for test items and ample time to vet them. NCLB has changed that.

The law's requirement that states align their tests to challenging state standards is an important step toward clarifying classroom expectations. But it is forcing the testing industry to custom-build the majority of the tests that must be in place at seven grade levels in every state this spring. And because a growing number of states release at least portions of their tests once they have been administered each year in order to give educators and parents more-detailed reports of student performance, testing companies have to generate vastly larger pools of credible test questions and do so within far shorter timelines. In the view of many in the industry, they can't find enough qualified people to do the work.

Testing companies "are desperately looking for people to write test items," says Hoover. "It's hard to do well, and it's hard to recruit people to do it." Adds Daniel Koretz, a testing expert at the Harvard Graduate School of Education: "Testing company executives tell me, 'We're having a hell of a time finding the caliber of people we need.' "

The surge in state testing under NCLB has created a severe shortage of the specialists who do the analyses of how test items perform in field trials and the other heavy statistical lifting in test-making. Though the work of these experts, who are trained in measurement theory and statistics and are known as psychometricians, is crucial to creating high-quality tests, only a handful of them enter the workforce each year from the University of Iowa, Michigan State, the University of Massachusetts, and the dozen or so other campuses that train them—under a dozen a year, reports a survey of doctoral degrees earned nationwide between 1995 and 2003 by the National Opinion Research Corp. An additional 35 Ph.D.s were awarded annually in the related field of "statistics, testing and education measurement." Cook, the University of Wisconsin testing expert, refers to NCLB as the "No Psychometrician Left Unemployed Act." [7]

The dearth of testing experts isn't hard to explain. Psychometrics is a highly technical, mathematics-based discipline that doesn't pay particularly well by private-sector standards (about $120,000 a year in top industry slots and much less in state testing agencies). And many potential recruits, undergraduates studying educational and quantitative psychology at colleges of education, are discouraged from entering the field by education school professors, many of whom are opposed to the rise of standardized testing in public education, says Hoover. "They say we're the bad guys."

To make matters worse, a growing number of psychometricians are working on non-educational tests, says James Impara, president of the National Council on Measurement in Education, a professional organization for education measurement experts. Over 1,000 occupations from accounting to firefighting now require licensure or certification, Impara says, and many of them give tests to applicants.

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-

long windows in which 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 computers or, as in Michigan, back to school systems and from there to the state agencies.

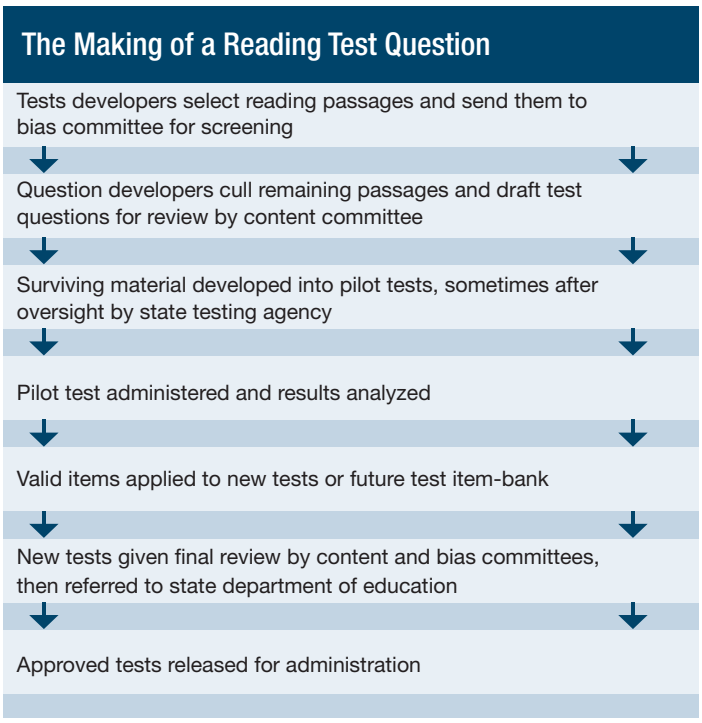
States' testing staffs calculate the percentages of students meeting state standards in reading and math. Once they do this for every NCLB student subgroup (students are grouped by race/ethnicity, family income, disability, and language proficiency) in every tested grade in every public school and school system, they 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, what NCLB calls "adequate yearly progress." Then the state agencies 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—in time for parents to place their children in tutoring or in different public schools prior to 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 up in August, that means that the entire testing and state rating process must be completed within six weeks from the end of the typical public school year.

This would be difficult enough to do successfully with long timelines. But many state policymakers, under pressure from public educators to give students as much time as possible to prepare for NCLB's high-stakes tests, are demanding that tests be administered late in the school year and that testing companies nonetheless complete their scoring and reporting in time to place underachieving students in summer school and in time for states to do their adequate yearly progress calculations ahead of the midsummer deadline for public reporting, says Jeff Galt, chief executive officer of Harcourt Assessment.

Lobbying by local educators led 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, Washington, D.C.-based American Institutes of Research and North Carolina-based Measurement Inc., both small companies, report scores on the tests by June 15—two weeks faster than in the past. And some states want even quicker turnarounds. Michigan ended its contract with Measurement Inc. in 2005 in the wake of months-long delays in scoring the state's tests. Pearson, the state's new contractor, has to get test results to local school systems within 30 days.

Despite the fact that testing companies have sought to upgrade their test-processing infrastructure in recent years, the pressure put on them by the volume of testing and the new scoring deadlines is "immense," says Scott Marion, vice president of the New Hampshire-based Center for Assessment, a nonprofit test-consulting firm that advises 15 state testing agencies. And it is intensified, says Stuart Kahl, president and CEO of Measured Progress, a New Hampshire-based testing company, by the fact that companies often must spend weeks after students take tests tracking down test booklets that school systems have failed to forward, resolving discrepancies between enrollment figures and the number of students' tests, and "cleaning up" basic student biographical information required to report test results under NCLB.

Headlines about scoring blunders are one measure of the overwhelming demands of the scale and speed of test processing required by NCLB. Another is that over half of the school systems in a 2005 national survey by the



Source: Scott Marion, Center for Assessment

Glossary of Terms

Validity: The extent to which tests accurately measure the knowledge or skills that the tests are intended to measure.

Constructed response: A test item that requires students to provide the answer to a question, as opposed to a multiple-choice question, where students choose among possible answers that the test creator provides. Constructed-response questions can be as simple as fill in the blank (e.g., 9 x 9 = ?) or require more complex answers, such as a written essay.

Rubric: A tool used to score answers to a test question. Most commonly used for scoring answers to more complicated, constructed-response questions, rubrics help ensure consistent grading from different graders by describing the specific elements of the answer needed for students to receive various score levels.

Criterion-referenced test: An assessment that measures the extent to which students have mastered a specific body of knowledge and skills, such as standards-based tests that determine if students have reached certain predefined levels of proficiency in a subject.

Norm-referenced test: An assessment that measures a student's performance relative to that of a representative national sample of students called a norming sample. Results on norm-referenced tests are often expressed in percentiles. A student scoring at the 60th percentile on a norm-referenced test, for example, scored better on the test than 60 percent of the students in the norming sample.

Scaling: A process of converting raw test results, such as the number of correct answers, into a score that can be used to compare results from different students or different versions of a test. Student scores on the SAT, for example, are converted to a scale where the minimum score for each section is 200 and the maximum score is 800.

Equating: The process of placing scores from different versions of the same test on a common scale, so that student results on those tests can compared on a fair, "apples to apples" basis. Equating, for example, ensures that a score of 700 on the SAT in 2005 is comparable to a score of 700 in 2006.

Reliability: A measure of the consistency and dependability of a test score's representation of a student's knowledge and skills. There are various dimensions of reliability, such as the consistency of test results when comparing the administration of the same test at different points in time, or comparing the results of different questions that measure the same skill, or comparing the scores given by different graders to the same answers.

Source: Scott Marion, Center for Assessment; Pearson Educational Measurement

Center on Education Policy said that late reports by state departments of education created "serious or moderate" problems in meeting NCLB's start-of-the-school-year deadline for informing parents of their children's eligibility to attend higher-performing public schools.[8]

Market Pressures

The testing industry is facing these challenges in a time of tight budgets and thin margins. A study by Harvard economist Caroline Hoxby revealed that states typically spend less than one-quarter of 1 percent of public school revenues on their statewide testing programs.[9] In 2005-06, combined federal, state, and local per-student spending in public education averaged over $8,000. Despite testing's tremendous influence on what students are taught and how teachers teach in the nation's public schools, and despite the importance of testing to school reform under NCLB, states spend between $10 and $30 per student on their testing programs, says Harcourt's Galt and other industry experts. Eduventures estimates that schools and school systems spend twice that amount on test-prep materials.[10]

The major testing companies weren't fazed by state testing budgets when they were selling large quantities of the ITBS, Stanford, and other national norm-referenced tests to schools and school systems. These "catalogue sales," as they are known in the testing industry, were lucrative, says Cook, who coordinated bids for state testing contracts at Harcourt. Publishers would invest $3 million to $6 million in a testing series and earn $15 million to $20 million over the five- to eight-year life of the tests, he says, because they were able to keep the development cost of every test booklet they sold low by using the same tests for a number of years.

But schools and school systems are buying far fewer of the major norm-referenced tests in the NCLB era of statewide testing. Sales of such tests are down 30 percent to 70 percent, says Cook. In the new NCLB marketplace, the publishers must make customized criterion-referenced tests that measure students' grasp of each state's unique academic expectations. Such tests, Galt says, can be five times as expensive to construct as the ITBS and other norm-referenced tests. And the fact that many states own the copyrights to their tests and release them to the public after the tests have been administered has wiped out the economies that publishers enjoyed when they were able to use their norm-referenced tests for several years

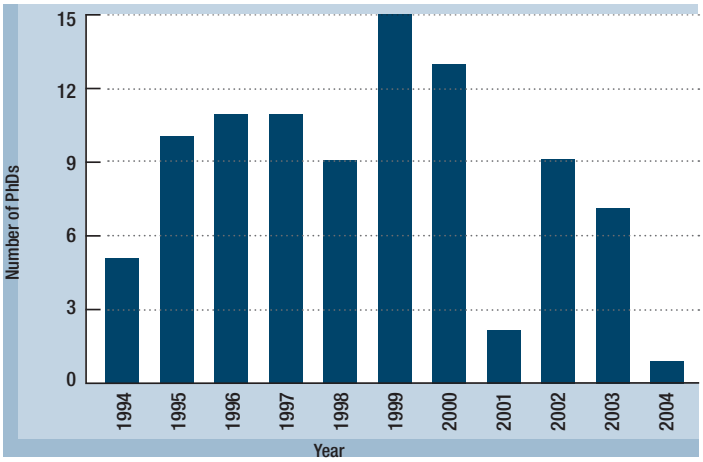
running. The result, predictably, has been much lower profit margins. “When you are building a new test every year,” says Hoover, the ITBS author, “it’s difficult to get your money back.”

To compensate, the major testing companies are vying aggressively—both with one another and with Measured Progress, Data Recognition Corp., and the other new players—for as many state NCLB testing contracts as possible, in an attempt to achieve efficiencies through scale.

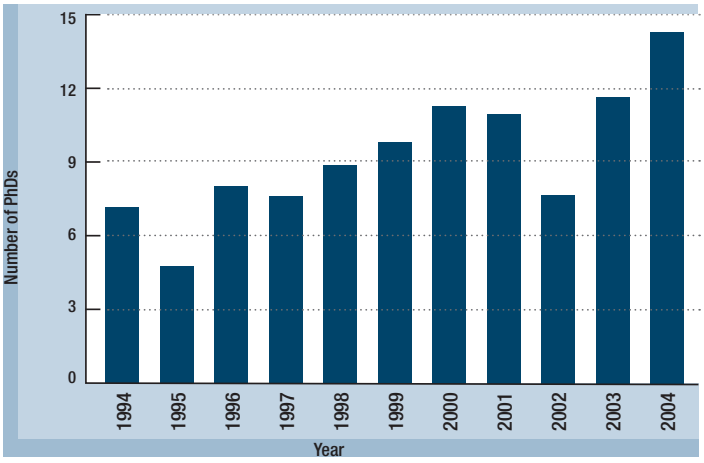
But this intense competition has led to more pressure on profit margins. Princeton-based ETS lost $18 million on its first NCLB testing deal, a three-year, $175 million contract with California, the nation’s largest market, says Anthony Carnevale, a former ETS vice president. Nonetheless, California tentatively approved a new three-year deal with ETS in late 2005, for even less money. John Oswald, an ETS senior vice president and general manager of the company’s elementary- and secondary-education division, says ETS lost money on its first California contract, made a slight profit during a one-year extension to the deal, and expects the new contract to be profitable.

Talent Search

Ph.D.s granted in Psychometrics, 1994 to 2004



Ph.D.s granted in Educational Assessments/Testing/ Measurement, 1994 to 2004



Source: National Opinion Reasearch Corporation, 2004

“With three or four companies bidding, one always has a reason to low-ball to get into the state,” says Kahl of Measured Progress. Pearson won a three-way competition in 2005 for Michigan’s testing business with a $48 million bid on a three-year contract. The second company, Data Recognition, bid $84 million, and the third, Measurement Inc., $114 million. Says Jon Twing, senior vice president for test and measurement services at Pearson, where he heads the company’s 200-person division that develops state tests: “We are reducing our rates in bids because our competitors are doing the same thing. It cuts profit margins.”

“There’s a tremendous amount of competition to get volume to cover fixed costs,” says Galt, Harcourt’s president. In response to higher testing volume and tighter scoring deadlines, Harcourt is spending $50 million over three years on printers, scanners, software, and other test-processing infrastructure—“most of which,” Galt says, “remain idle for 10 or 11 months of the year.” Doug Kubach, president and CEO of Pearson Educational Measurement, says there’s “hyper-competition in the industry.”

Penalty clauses in state testing contracts that have become more common and more prescriptive since NCLB became law are also squeezing testing company profit margins. Pearson’s 2005 deal with Michigan, for instance, stipulates that the company must pay the state four cents per student per day for every NCLB test it fails to score and return to school systems within four weeks—a potential maximum fine of $100,000 a day. Pearson’s penalty clock began ticking in early December of that year, and the company quickly racked up two weeks’ worth of fines in several parts of the state, says Roeber, adding that, “We’ll probably save some money on our testing program this year.”

Meanwhile, NCLB has spawned a secondary testing market that is further taxing the testing industry’s capacity. The tremendous pressure on schools and

school systems to have their students to do well on NCLB tests and the advent of technology that permits companies to very rapidly give school superintendents, principals, and teachers detailed breakdowns of student test results have produced a burgeoning market among school systems for so-called formative tests—short tests that are administered throughout the school year to help educators respond quickly to student weaknesses. The rise of formative testing has increased demand for banks of test questions dramatically, industry experts say. Sales increased 50 percent between 2003 and 2006, to $323 million, Eduventures estimates.

Industry Leaders

Nine companies capture more than 95 percent of the expenditures from states for tests and testing services.

COMPANY & PRIMARY STATE CONTRACTS

Education Testing Service
California
New Jersey

Harcourt Assesment
Arizona
Deleware
Hawaii
Idaho
Illnois
Mississippi
New Mexico
Rhode Island
South Dakota
Virginia
Wyoming

Riverside Publishing
Arkansas
Iowa
Louisiana

CTB/McGraw-Hill
Alabama
Alaska
California
Colorado
Connecticut
District of Columbia
Florida
Indiana
Kentucky
Mississippi
Missouri
New Mexico
New York
North Dakota
South Dakota
Tennessee
West Virginia
Wisconsin

Measured Progress
Maine
Massachusetts
Montana
Nevada
New Hampshire
Utah
Vermont

Questar Education System
Arkansas
Minnesota

Pearson Education Measurement
Minnesota
New Jersey
Texas
Washington

Northwest Evaluation Association
Idaho

Data Recognition Corp.
Alaska
Pennsylvania

Source: The State of the K-12 State Assessment Market, April 2005, Eduventures Inc.

Uneducated Consumers

State departments of education are ultimately responsible for carrying out NCLB's testing mandates, and they are, if anything, in a weaker position than the testing industry to respond to the surge in testing under NCLB. Many state testing offices suffer from heavy turnover and shortages of skilled staff as a result of underfunding and hiring freezes introduced during the recession of the late 1990s. "The capacity of assessment and accountability offices in state departments of education is very low," says Marion of the Center for Assessment, who was testing director in Wyoming. "And there has been a significant increase in workload since the passage of NCLB, without an increase in qualified staff."

Education Sector found in a survey of state testing directors that it conducted for this report that over half the states have problems recruiting and retaining the testing staff they need to respond to NCLB adequately. Testing companies and large school systems, many say, are luring away scarce psychometricians and other key staff with higher salaries. Ohio's new testing director, Judy Feil, for example, has lost seven of 20 staffers in the past year to burnout and higher-paying jobs in school systems and testing companies. She has been forced to hire replacements without testing backgrounds, she says, because "skilled people are difficult to find." Many state testing offices operate with skeleton staffs. Indiana, a relatively large state, has five professional testing employees.

Testing directors themselves turn over at an alarming rate. Matt Gandal, executive vice president of Achieve Inc., a Washington, D.C.-based organization that promotes high education standards, says the testing directors of Ohio, Florida, Texas, Maryland, and Rhode Island have left in the past few years for the private sector and an opportunity to earn higher salaries. Marion describes standing with six or seven state testing directors at a national conference, only half of whom had been in their jobs for more than eight months.

The result of the understaffing and lack of expertise in many state testing offices, says Marion, is that, "for the testing companies, it's like being auditors of their own work." Many state testing agencies simply don't have the capacity to scrutinize the work of their testing contractors closely.

Troubling Consequences

The mounting scoring errors and reporting delays that have resulted from the many challenges confronting the testing industry and state testing agencies as they struggle to respond to NCLB's testing mandates have tarnished NCLB's testing-based system of school accountability. They have created a public-relations problem.

But the lack of state oversight of testing contractors, the industry wide shortage of testing experts, and the many other problems that have plagued the spread of statewide testing under NCLB are also damaging the cause of standards-based reform in ways that don't make many headlines but are arguably more fundamental.

Testing experts say many statewide tests are not getting sufficient psychometric scrutiny to ensure that they accurately measure student and school performance under NCLB. "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 Marion of the Center for Assessment, who has taught test-making at the University of Maine. "But they aren't doing it." "In many cases," says Cook, "they are putting [test] items on the street they shouldn't."

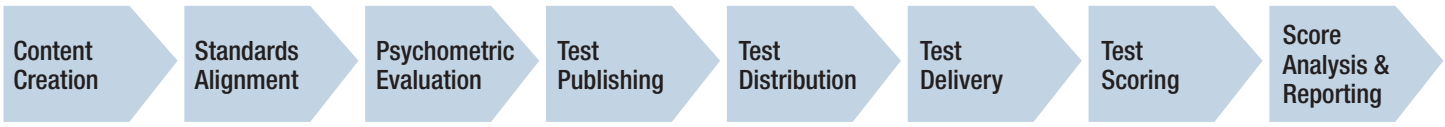
That's particularly true of test questions that require students to write a response rather than fill in a bubble on an answer sheet. The reason is that so-called open-response questions are more costly to field-test because they must be scored by people rather than machines. "You are paying a fortune on an individual item just to try it out," says Hoover, the onetime ITBS author, "so, frequently, companies never try them out and they are bad items." University of Iowa psychometrician Stephen Dunbar, Hoover's successor at the ITBS, refers to NCLB as "No Item Left Behind," because the law has led to such a shortage of quality test questions.

In another example of the consequences of psychometric failings, the Ohio Department of Education announced in the fall of 2005 that Measurement Inc. had failed to correctly translate raw scores on the state's high school test into scores on a publicly reportable scale. The "scaling" mishap resulted in new scores for 5,000 of the 5,400 students who had taken the test the previous summer, including 900 students who had been told they could not graduate because they had failed the test, when they hadn't.[11]

And while NCLB's strength as a source of standards-based school reform rests on its requirement that states measure students' grasp of statewide standards and then take steps to improve schools and school systems where students don't measure up, lack of time, money, and skilled staff have led a substantial number of states to introduce tests that many testing experts say are not fully aligned with state standards—tests that don't test what states expect their students to know. This is happening in part, experts say, because rather than building tests from scratch, states are hiring 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. "When you ask publishers if they align the tests with state standards, you'll rarely get an answer of less than '85 percent,' " says Marion. "But our studies show it's a lot lower, 50 percent. As a result, teachers are teaching stuff that they can't be sure is on the tests," because the tests don't necessarily measure the skills that states say teachers should teach.

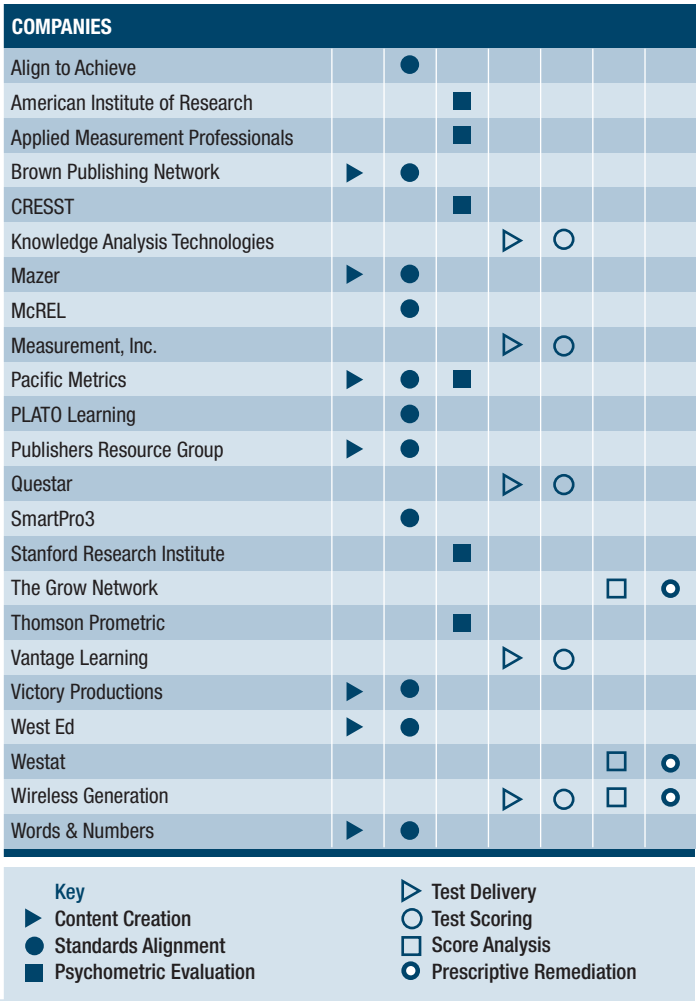
Nor is the quality of the many practice tests that students are taking in increasing numbers to prepare for NCLB testing what it should be. Formative testing has the potential to help students by giving teachers frequent and, in theory, useful, information on student performance. But so far, say industry analysts, schools and school systems have been unwilling to pay for high-quality test items for these new tests, leading testing companies to focus resources on supplying the new market with banks of test questions that are less fully field-tested and thus less expensive—but are also less accurate measures

The Work of the Testing Industry



Source: The State of the K-12 State Assessment Market, Eduventures, Inc., 2005

Niche Players



Source: Staying Ahead of the Curve, Eduventures, Inc., 2004

of student performance. "Certain firms claim to offer tens of thousands of exam items," Eduventures writes in an industry report, *Testing in Flux*. But many of the items, says Eduventures, "have not undergone rigorous psychometric evaluation."[12] Says Marion, "The items that end up on most of these formative tests are ones that get rejected from state tests." As a result, many formative test questions don't accurately measure what students know.

But the use of such test items is increasingly widespread. In 2003, in one of a number of moves by major testing companies to tap into the formative-testing market, Houghton Mifflin bought Edusoft, a then-three-year-old company that permits teachers to give tests, scan student answer sheets, upload them to Edusoft servers, and receive detailed score reports. School systems began asking Houghton to supply banks of practice test items. But Houghton couldn't deliver fully field-tested items for what many school systems were willing to pay for them. "School districts don't appreciate, or can't afford, high-quality items and tests," says Alvaro Fernandez, a former Edusoft executive. "They have an insatiable hunger for inexpensive item banks that their teachers can use to help them do better on the NCLB tests." So Houghton struck a deal with FS Creations, an Ohio-based company, to market low-end questions alongside its higher-quality Riverside products, says Fernandez.

Simple Questions

Perhaps the most troubling classroom consequence of the tumult in the testing industry is the strong incentive the problems have created for states and their testing contractors to build tests that measure primarily low-level skills.

NCLB has sought to lift the level of teaching in the nation's classrooms by requiring states to set challenging standards for what students should know and be able to do. But testing experts say that many of the tests that states are introducing under NCLB contain many questions that require students to merely recall and restate facts rather than do more demanding tasks like applying or evaluating information, largely because it's easier and cheaper to test the simpler tasks.

Such test questions do have a role; it's important that students' grasp the most basic skills. But because teachers have so much riding on their students' results, tests that stress such skills encourage teachers to emphasize them in their classrooms at the expense of the high standards that NCLB has sought to promote. They strip teachers of the incentive to teach higher-level skills. "Tests are focusing more and more on rote skills because it's difficult, given the demand that they be constructed quickly and cheaply, for anything else to happen," says Hoover. "Writing items that tap higher levels of comprehension is really difficult. The problem is that tests of rote skills encourage rote teaching. It's not a good model for instruction." As Marion puts it, "The further away we get from testing the types of things we want kids to do in school, the less likely we are to improve education."

Such tests also give a skewed sense of student achievement. Scores on reading tests that measure mainly literal comprehension are going to be higher than those on tests with a lot of questions that require students to

evaluate what they've read by, say, reading two passages and identifying themes common to both. The same is true in math. In a study by Lorrie Shepard, a testing expert and the dean of the school of education at the University of Colorado–Boulder, 85 percent of third-graders who had been drilled in computation for a standardized test picked the right answer to 3 x 4, but only 55 percent answered correctly when presented with three rows of four Xs.[13] Many of the new NCLB tests, as a result, are likely to suggest that students are achieving at higher levels than they really are. The tests have the potential to create glass ceilings for higher-achieving students, who have less of an opportunity to demonstrate the extent of their abilities. And when the scores of low-achieving students rise, this achievement ceiling could create the sense that performance gaps among groups of students are closing, when in fact they may not be.

It is possible to write multiple-choice questions that measure higher-level skills. But doing so is difficult and time consuming. "It's almost always easier to pick out factoids," says Marion. And when they're given a choice, most testing experts would measure students' grasp of more advanced abilities through open-ended or constructed-response questions that require students to produce their own answers rather than select answers from among those supplied by test-writers, the format of multiple-choice questions. "Constructed-response questions give you more measurement depth," says John Olson, director of psychometric and research services at Harcourt Assessment and director of assessment at the Council of Chief State School Officers from 1998 to 2003. "They give you a better sense of what students can do. And as a result, teachers get more out of them."

But such questions are more expensive and slower to process than their multiple-choice counterparts. Multiple-choice answer sheets, with their rows of "bubbled-in" circles, are scored in seconds by optical scanners. Harcourt's Galt says the company has scored 1 million Florida multiple-choice tests in a day. But grading constructed-response questions, where students write out answers, is more complicated and time consuming.

States and their testing contractors must first establish "rubrics," or standards, for judging students' responses, since there are often no "right" answers to such questions. 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 a week training their test graders to ensure that answers of comparable quality receive the same scores from different graders, most of whom are moonlighting or retired educators earning anywhere from $9 to $25 an hour. 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 in Dover, N.H.; Iowa City; Minneapolis; Durham, N.C.; Monterey, Calif.; and elsewhere, working their way through hundreds of answers at a rate of 20 to 30 items per hour. The result is that it costs anywhere from 50 cents to $5 to score a constructed-response question, compared with pennies per multiple-choice question, says Cook.

States Reporting Testing Problems

52%	**Capacity**—Difficulty recruiting and retaining qualified staff for testing-related positions
35%	**Accuracy**—Experienced a significant error by a contractor in scoring a state test since 2000
20%	**Timeliness**—Did not receive test results from a contractor in a timely fashion since 2000

Source: Survey of State Testing Offices, Education Sector, 2006, based on responses from 23 states

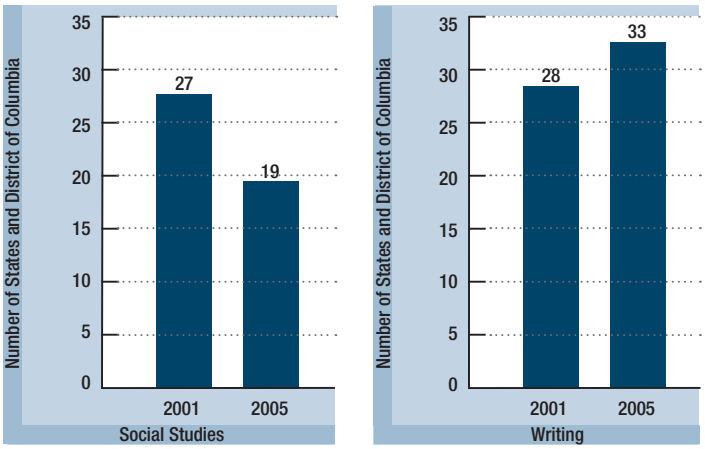
And the cost differential is not lost on the state legislators who control state education department budgets. In 2004, Pearson brought the membership of the Michigan House and Senate education committees to Iowa City to tour the company's high-tech facility for scoring multiple-choice answer sheets.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

The "efficiency" of multiple-choice questions also works against open-ended questions in another way. The reliability of a test, its ability to accurately gauge that a strong student is a strong student, increases with the number of questions it has, and students are often able to answer multiple-choice questions faster because it's quicker to fill in a bubble than write out an answer. So there can be many more multiple-choice questions than open-ended questions on a two-hour test.

As a result, there are very few open-ended questions on many of the new NCLB tests, say testing experts. "There's

pressure to use fewer of them," says Pearson's Twing. "There are hardly any in grades three through eight in many states and there are just a handful at the high school level—three or four out of 40, and they tend to be short answers that can be quickly answered and easily scored." "States are shifting from constructed response to multiple choice due the cost and time of scoring constructed-response questions; we are seeing more of that," agrees Olson, Twing's counterpart at Harcourt Assessment, who spent a decade working on the federally funded National Assessment of Educational Progress (NAEP), considered to be one of the nation's most sophisticated tests with many questions measuring higher-level skills. "During the

Number of States Testing Subjects Not Required by NCLB



Source: The Impact of NCLB on Non-Tested subjects, Patricia V. Pederson, Harcourt Assessment, Inc., 2005

1990s states had 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 eliminated non-multiple-choice questions from its state tests in 2005-06. So did Kansas. In all, 15 states serving 42 percent of the nation's students are using NCLB reading and math tests in 2005-06 that have no open-ended questions, writes *Education Week*.[14] NCLB's requirements for more testing in reading, math, and, by 2006-07, science have also led states to cut back the use of testing to drive improvements in other subject areas. "There has been so much focus on math and reading that states are cutting back on other subjects, social studies in particular," says Olson. "They can't afford to do both," says Feil, Ohio's testing director.

Research by Patricia Pederson, a former director of science and social studies testing at Harcourt Assessment, reveals that the number of states testing in science rose from 34 in 2001, prior to the enactment of NCLB, to 40 in 2005, in anticipation of the NCLB's 2006-07 science-testing requirement. The number of states testing writing rose from 28 to 33. But the number testing social studies declined from 27 to 19 during the same period.[15] Because schools tend to teach what's tested when test scores have consequences for teachers and principals, the cutbacks in testing in such subjects as social studies are encouraging schools to focus their energies on reading, math, and, increasingly, science, in the same way that the large number of multiple-choice questions requiring only rote responses on the new state tests is leading them to focus on the lowest-level skills within reading and math.

There's clearly value in pressing educators to ensure that students have a strong grounding in the building-block subjects of math and reading. Such grounding is one of NCLB's primary goals. But the cost has been a narrowing of the curriculum in many schools and classrooms. And Hoover, for one, calls this marginalization of history, art, and music a huge downside of NCLB's testing requirements and the overwhelming demands they have placed on state testing agencies and testing companies.

The Low-Cost Track

Congress has sought since the implementation of NCLB to help states meet the financial burden of the law's testing mandates. It is giving them $412 million for statewide tests in 2005-06 and $408 million in 2006-07, about $9 a test. A provision in NCLB championed by the late Minnesota Sen. Paul Wellstone suspends the law's student testing requirements if federal testing aid falls below specified levels—$400 million in 2005-06.

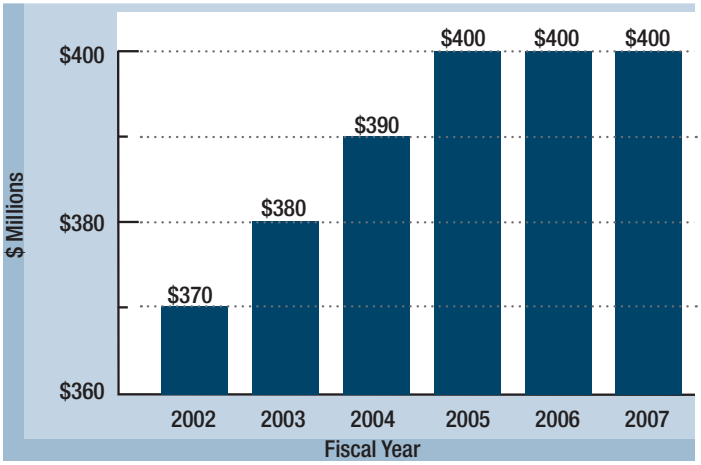
But states can spend the money on a range of tasks unrelated to test-building, such as "improving the dissemination of information on student achievement and school performance," and Eduventures reports that "states in many cases have opted to allocate most of those [federal] dollars to the development of state standards or initiatives aimed at aligning instruction to state standards"—important activities, but not the work of state testing agencies.[16] In some instances, testing experts say, state lawmakers are using the federal testing money to supplant rather than supplement their state

testing budgets. "If I'm a Senate education committee chairman and you tell me you've got $10 million in federal testing money, my response is, 'That's real simple, I'll knock $10 million from your testing budget,' " says Cook, the former Wisconsin testing director. There is supplement-not-supplant language in NCLB regarding testing funding, but such provisions are notoriously difficult to enforce.

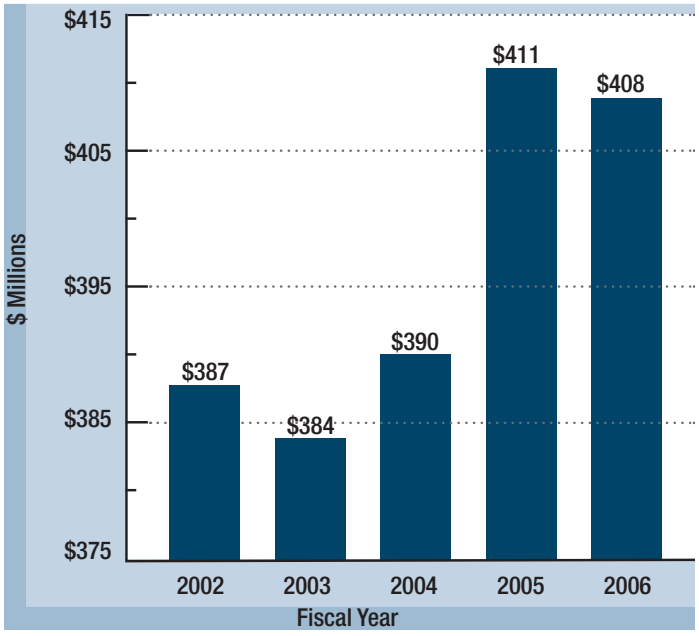
Connecticut has had high-quality tests with many open-ended questions since the 1980s, including math questions that require students to write explanations of their answers. In 2005, the state sued the U.S. Department of Education over the cost of NCLB testing, saying that Connecticut's share of the congressional appropriation is inadequate to fund tests of the same caliber under NCLB. U.S. Secretary of Education Margaret Spellings seemed to suggest as much in a letter to Connecticut's commissioner of education, Betty Sternberg, in response to the state's legal action. "Some of the costs of Connecticut's testing system are attributable to state decisions [regarding the types of tests it uses]," Spellings wrote. "While these decisions are educationally sound, they go beyond what was contemplated by NCLB." [17]

Federal Spending on State Testing

Minimum Required by NCLB



Actual Spending



Source: The State of the K-12 State Assessment Market, Eduventures, Inc., 2005

The then-General Accounting Office, a research arm of the U.S. Congress, more or less predicted the confrontation between Connecticut and the U.S. Department of Education when, in a 2003 report, it produced three widely varying estimates of what it would cost states to comply with NCLB's testing mandates. It would cost them $1.9 billion between 2003 and 2008 if they used tests with only multiple-choice questions, the agency suggested. The price would rise to $3.9 billion if their tests used multiple-choice questions and some open-ended items. And it would reach $5.3 billion to build tests with a larger percentage of open-ended questions.[18]

But the agency didn't predict the consequences on the testing industry, state testing agencies, and the nation's classrooms of the low-cost track that federal and many state appropriators have followed and that Secretary Spellings has sought to defend. Given the tremendous influence of statewide standardized testing on public schools today, NCLB's pursuit of significantly higher levels of student achievement requires 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 teachers and principals with timely information on students strengths and weaknesses.

That means investing resources commensurate with testing's central role in school reform today. Together, steps such as increased federal spending and targeted research and development activities, the establishment of a national testing oversight body, and incentives for greater inter-state collaboration would strengthen the national testing infrastructure dramatically, allowing it to support the full weight of standards-based reform.

RECOMMENDATIONS

State and federal policymakers can address the problems raised in this report by enhancing federal leadership on testing issues, creating an independent national oversight body to promote test quality, giving states incentives to collaborate on test development and, ultimately, developing voluntary national assessments.

Federal Leadership

The federal government should take several steps to improve the nation's education testing infrastructure. First, it should greatly increase the supply of well-trained psychometricians and other testing experts. The U.S. Department of Education should fund the training of 1,000 such specialists over the next five years, through grants to support students who commit to working in the field after they've completed their degrees. The Education Department should also fund university-based research on test quality. Doing so would help build university capacity to train more testing experts and give professors in psychometrics and related fields incentives to recruit students to serve as research assistants, while promoting needed research and development in testing.

Second, federal funding for testing under NCLB should be increased from its current level of $408 million to $860 million annually. This would help give all states the resources necessary to develop tests on par with those in states that currently have the strongest testing programs. Massachusetts's standards and tests, for example, have been widely praised.[19]

The state's custom-designed statewide tests include a healthy mix of high-quality multiple-choice and open-ended questions that require students to construct answers. The state has also developed a large network of in-state teachers to write enough new test questions that every test can be made public after students have taken it to help parents, educators, and the public understand the process. Massachusetts currently supports a testing staff of 30 that monitors the scoring of the state's tests at contractor's centers across the nation, ensures that open-ended questions are appropriately benchmarked and scored, and conducts training sessions for teachers throughout Massachusetts to ensure that test administration goes smoothly.

But Massachusetts's $8 million share of the federal government's current funding for statewide testing under NCLB covers only 30 percent of the state's NCLB testing costs, says Jeff Nellhaus, associate commissioner of the Massachusetts State Department of Education and the state's testing director from 1994 to 2002. Most states don't fund their testing programs as generously as Massachusetts does. Increasing federal funding for statewide testing to $860 million would allow the federal government to provide every state two-thirds of the funding necessary to reach Massachusetts's level of per-student spending on NCLB-required testing.[20] This support would balance federal and state educational responsibilities and help ensure that NCLB fosters a race to create high-quality assessments, not a race to the bottom.

Such an investment is only a small fraction of the approximately $500 billion spent on public education in 2005-06 and is actually a minimal investment, given the key role that statewide testing plays in NCLB and standards-based reform generally. In fact, most industries would clamor for the opportunity to have a quality-control and feedback mechanism that costs less than 1 percent of their overall operating revenue. "Given the [state testing] systems' low costs," says Hoxby, the Harvard University economist, "we ought not hesitate to improve them."[21] Because statewide standardized testing has such a strong influence on teaching and learning in the nation's classrooms today, there should be sufficient funding to create and administer very high quality tests. Given the stakes, even pretty good tests are not good enough, much less the tests in place in many states today.

At the same time that the federal government invests more in testing, it should demand more, too. The federal government should make certain that strict supplement-not-supplant provisions ensure that new investments support the development of high-quality tests rather than displace existing state funding. In addition, Washington should ensure that states are using assessments that are well aligned with state standards and based on sound and clearly articulated definitions of what students should know and be able to do.

Third, the federal government should fund new research and development on testing. Though competition in the testing industry has led companies such as Harcourt Assessment to invest in equipment to improve the efficiency of their test processing, the pressure on profit margins has made investments in testing research and development difficult.

The federal government could make an important contribution by stepping in to support research and development on testing, catalyzing new ideas and technologies. It could fund research into ways to make the use of open-ended questions less expensive or ways to more accurately test students with disabilities who are able to participate in the general curriculum. Similarly, more sophisticated assessments for English-language learners would help educators more accurately pinpoint the needs of those students.

Another area needing research is online testing, which offers the promise of more customized assessments, a shorter and simpler testing and scoring process, and more comprehensive reports on student achievement quickly. Online testing also offers the possibility of states incorporating new types of test questions in their statewide exams, including "drag and drop" items and simulations. And it creates the potential to identify students' skills and weaknesses using fewer test items through so-called computer-adaptive testing, which adjusts the difficulty of test questions to students' performance on previous questions.

There have been significant advances in these areas, but a host of technical and financial challenges remain to be solved before the technology can be successfully applied to statewide testing programs, says Kahl of Measured Progress, which has worked with Maine and other states in implementing online testing. "Half the states are asking for it," he says, "but there are bugs galore." More research and development would help to address the problems facing technology-based testing.

Effective Oversight

Improving the nation's fragmented testing system will require strong, effective federal leadership. **President Bush should begin by inviting the leaders of both parties in both branches of Congress to work with him to name leading experts to a bipartisan presidential commission on standardized testing.** That group, comprising state testing officials, testing industry representatives, and independent testing experts, would study a range of testing-related issues, including funding, testing quality, industry capacity, and human capital.

As part of its work, the commission should establish an independent national testing oversight agency. Despite the enormous importance of statewide standardized testing in public education today, the tremendous potential for problems in test development and test administration under NCLB testing, and the inability of many states to supervise their testing contractors effectively, there is no entity to independently audit state testing programs and the testing industry. Such a body would perhaps be called the National Testing Quality Commission and would operate in the spirit of the Consumer Product Safety Commission and other federal consumer-protection agencies.

The U.S. Department of Education requires states to submit their NCLB accountability plans for approval through "peer review," a process that evaluates the way in which states propose to establish their academic standards, align their tests to the standards, and ensure the technical quality of their tests. But the department's peer review system does not audit the quality of test items or the performance of state testing contractors.

While professional associations give guidance on the technical quality of standardized tests, they do not play an auditing role. An independent National Commission on Testing and Public Policy over a decade ago wrote that, "Although the American Psychological Association, the American Education Research Association, and the National Council on Measurement in Education have formulated professional standards for test development and use in education … they lack any effective enforcement mechanism." [22]

The U.S. Department of Education's Office of Inspector General seems to agree. Part of its study of scoring problems on state tests will be to determine "whether there is a need for federal oversight to help ensure [that] errors in scoring high-stakes tests are prevented, detected, and disclosed publicly." [23]

Interstate Collaboration

The inefficiency of states and the District of Columbia administering 51 separate testing programs is obvious. **States could create higher-quality tests at lower cost if they worked together to develop common tests.**

Three states—New Hampshire, Rhode Island, and Vermont—have done this. Following the enactment of NCLB, they formed the New England Common Assessment Program. The states' testing officials worked with Measured Progress and the Center for Assessment to build reading and math tests in grades three through eight that students in the three states took for the first time in fall 2005. The consortium produced tests with an ample number of open-ended questions at two-thirds of what the states would have paid had they developed the tests individually, says Kahl. Each state spent $2 million on the project rather than the $3 million they would have each spent had they not collaborated. There is a particularly strong incentive for smaller states to create testing consortia, because it costs nearly the same to develop a test in a small state like South Dakota as it does in a large state like Texas. The savings for small states are likely to be greater per student because with lower enrollments their development costs are higher on a per-student basis. But larger states would also benefit financially from such collaborations, and even the largest would benefit from the greater efficiencies that would result from multi-state contracts for test administration, scoring, and reporting.

States entering testing consortia would have to agree on test content, as Vermont, New Hampshire, and Rhode Island have. But despite the political mantra of local control, state standards in reading and math do not vary that much from state to state now. This means partnerships like the New England Common Assessment Program are both politically and practically possible. **The federal government should encourage the creation of state testing consortia by offering states that elect to work together additional funding to support their assessment programs under NCLB.**[24]

In the long run, the logic of regional consortia leads to a solution that already exists in many of the industrialized nations of Europe and Asia: a single national testing system. By encouraging states to arrive at such a system through voluntary collaboration the federal government could support these efforts while sidestepping some of the thornier political questions about national testing. The closer we get to such a system, the more the nation's overextended testing infrastructure could focus on creating much smaller numbers of much higher quality assessments and test items. Moreover, it would free up resources to build strong national tests that states could voluntarily adopt in important disciplines such as social studies and writing, subjects that don't benefit from consistent measurement under NCLB.[25]

The concept of national testing is, of course, highly controversial. President George H.W. Bush proposed the creation of voluntary national standards in elementary and secondary education in the early 1990s, and the Clinton administration sought voluntary national tests in the mid-1990s. Partisan politics played a role in the defeat of both initiatives. But there was also strong philosophical opposition from both the left and the right. Liberals argued that the tests would be unfair to students in impoverished communities, while conservatives argued that national standards and testing would amount to a federalization of public schooling.

But although there is certainly not consensus support for national tests, the politics of the issue have changed dramatically since it was last seriously considered in Washington. By mandating statewide testing throughout the country, NCLB, a law proposed by a conservative Republican president and passed by a Republican-controlled Congress, has imposed a much greater degree of centralization in standards-setting and testing than existed previously. NCLB also expanded federal authority in education. Meanwhile, prominent liberal advocacy groups, including the Citizens' Commission on Civil Rights and the Education Trust, have been vocal supporters of statewide testing under NCLB. And both conservative voices in education such as the Thomas B. Fordham Foundation and liberal organizations such as the Center for American Progress have recently endorsed national standards and testing in public education. The early success of the New England Common Assessment Program is likely to further soften opposition to common standards and common tests.

ENDNOTES

Report Endnotes

[1] Editorial Projects in Education Research Center, as cited in *Education Week*, Nov. 30, 2005; Education Sector calculations based on data from U.S. Department of Education, National Center for Statistics.

[2] For further reference please see the following: Robert A. Frahm, "State Fires Testing Contractor," *Hartford Courant*, Sept. 13, 2005; Doug Guthrie and Christine MacDonald, "Some Schools Await MEAP Tests," *The Detroit News*, Oct. 5, 2005; Paul Tosto, "Analysis of State Education: Reliance on Testing Poses Many Pitfalls," *St. Paul Pioneer Press,* March 13, 2004; Frank Rich, "Testing Co. Mistakenly Fails Ohio Students," Associated Press, Dec. 12, 2005; Kathleen Rhoades and George Madus, *Errors in Standardized Tests: A Systemic Problem* (Boston: National Board on Educational Testing and Public Policy, 2003).

[3] *Office of Inspector General FY 2006 Work Plan* (document available by search on U.S. Department of Education Web site: www.ed.gov).

[4] J. Mark Jackson and Eric Bassett, *The State of the K-12 State Assessment Market* (Boston: Eduventures, 2005).

[5] President's FY 2006 Budget Request for the U.S. Department of Education, Appendix 3: Total Expenditures for Elementary and Secondary Education in U.S. (Published online at http://www.ed.gov/about/overview/budget/budget06/summary/edlite-appendix3.html).

[6] Editorial Projects in Education, "Quality Counts 2001: A Better Balance," *Editorial Projects in Education* Vol. 20: No. 17, January 2001 (Published online at http://counts.edweek.org/sreports/qc01/articles/qc01story.cfm?slug=17toc.h20).

[7] *Doctorate Recipients from United States Universities:Summary Report 2004* (sponsored by the National Science Foundation, the National Institutes of Health, the U.S. Department of Education, the National Endowment for the Humanities, the U.S. Department of Agriculture, and the National Aeronautics and Space Administration; published online at http://www.norc.uchicago.edu/issues/sed-2004.pdf).

[8] *From the Capital to the Classroom: Year Three of the No Child Left Behind Act* (Washington, D.C.: Center on Education Policy, March 2005).

[9] Caroline M. Hoxby, "The Cost of Accountability," *NBER Working Papers 8855*, (Cambridge, Mass.: National Bureau of Economic Research Inc, 2002).

[10] Jackson and Bassett, *The State of the K-12 State Assessment Market*.

[11] Vicki Lee Parker, "Measurement Inc. Slips Down a Notch; After a Goof on Ohio Students' Tests, Company Loses No-errors Bragging Rights," *The News & Observer* (Raleigh, N.C.), Dec. 21, 2005.

[12] Matt Stein, *Testing in Flux: Future Directions in the Pre-K-12 Assessment Market* (Boston: Eduventures, 2004): 12.

[13] L.A. Shepard, *Measuring Achievement: What Does it Mean to Test for Robust Understandings?* (Princeton, N.J.: Policy Information Center, Educational Testing Service, 1997).

[14] *Education Week*, Nov. 30, 2005.

[15] Patricia V. Pederson, *The Impact of NCLB on Non-Tested Subjects: An Assessment Developer's Perspective* (Session 1108T, presented at Voices of Education: Unleashing the Power, Passion and Promise, ASCD Annual conference and Exhibit Show 2005, Saturday, April 2, 2005).

[16] Jackson and Bassett, 18.

[17] Margaret Spellings, the secretary of education, to the Honorable Betty J. Sternberg, commissioner of education for the State of Connecticut, May 3, 2005 (Published online at http://www.state.ct.us/sde/nclb/Correspondence/SpellingsLettertoBetty5-3-05.pdf).

[18] United States General Accounting Office, *TITLE I: Characteristics of Tests Will Influence Expenses; Information Sharing May Help states Realize Efficiencies* (USGAO, GAO-03-389, May, 2003): 15.

Recommendation Endnotes

[19] *Measuring Up: A Report on Education Standards and Assessments for Massachusetts*, Published online at http://www.achieve.org/achieve.nsf/StatePro-Massachusetts?OpenForm. (Achieve, Inc., 2001).

[20] The proposed $860 million amount was calculated by multiplying Massachusetts' estimated per-student test cost of $25 by the twice the number of public elementary and secondary students in grades 3-8 and 10 (to account for the cost of two tests, one in reading and one in math), and then multiplying that result by two-thirds.

[21] Caroline M. Hoxby, "The Cost of Accountability," *NBER Working Papers 8855*, (Cambridge, MA: National Bureau of Economic Research, Inc, 2002).

[22] Kathleen Rhoades and George Madus, *Errors in Standardized Tests: A Systemic Problem* (Boston: National Board on Educational Testing and Public Policy, 2003): 8.

[23] *Office of Inspector General FY 2006 Work Plan* (document available by search on USDOE website: www.ed.gov).

[24] Andrew J. Rotherham, "Asking The Wrong Test Questions," *The Washington Post*, May 29, 2001, A15.

[25] Thomas Toch, "Bush's Big Test," *The Washington Monthly*, November, 2001.

ABOUT THE AUTHOR

Thomas Toch is co-founder and co-director of Education Sector. He can be reached at ttoch@educationsector.org.

ABOUT EDUCATION SECTOR

Education Sector is an independent education think tank based in Washington, D.C. It is a non-profit and non-partisan organization devoted to developing innovative solutions to the nation's most pressing educational problems. The organization seeks to be a dependable source of sound thinking on education policy and an honest broker of evidence in key education debates in Washington and nationally.

ACKNOWLEDGEMENTS

Thanks to my Education Sector colleagues Andrew Rotherham, Kevin Carey, Sara Mead, and Bill Tucker for their thoughtful comments on drafts of the report. Kevin and Sara also surveyed state testing offices for the report. Sharon Cannon, Renée Rybak, Ethan Gray, and Lisa Guido, provided valuable research support, and Sharon managed the report's production. Susan Vavrick copy-edited the report, Alison Dacher designed it, and Molly Norton managed our communications effort. Thanks also to Scott Marion of the Center for Assessment for reviewing a draft of the report and contributing to the glossary of testing terms and the chart describing how a test item is created. Gary Cook of the Wisconsin Center for Education Research was also kind enough to read a draft of the report. Eric Bassett of Eduventures, Inc., graciously permitted Education Sector to reproduce charts and graphs from Eduventures research reports. David Hackensen of Pearson Educational Measurement, Mark Slitt of Harcourt Assessment, and Tom Ewing of the Educational Testing Service helped set up interviews and track down information. I am particularly grateful to the nearly three dozen testing industry executives, state testing officials, and other testing experts who shared their candid assessments of the testing industry for this report.

Margins of Error: The Testing Industry in the No Child Left Behind Era