Special Report: Testing Takes Off

State Test Programs Mushroom as NCLB Mandate Kicks In

Nearly half of states are expanding tests into more grades in 2005-06 school year

BY LYNN OLSON

You could describe it as the year of the tests. Twenty-three states are expanding their testing programs to additional grades this school year to comply with the federal No Child Left Behind Act.

"Every group I've been talking to, I've just said, 'Be patient with us this year,' " said Alexa E. Posny, the deputy commissioner of education in Kansas, which is adding reading and math tests in four grades in each subject.

"First, it's the sheer volume," she said. "In the past, we would develop 4,000 test items; we're developing 18,000 items. Second is the number of teachers who have never participated in state assessments, so it's a whole new ballgame for them. And then there's the overwhelming amount of data that will be available because there are so many more grades."

Forty-eight states and the District of Columbia will give standards-based tests in reading and mathematics in grades 3-8 and at least once in high school this school year, as required by the nearly 4-year-old federal law, according to a survey by the Editorial Projects in Education Research Center.

The holdouts are Iowa and Nebraska. Districts in Iowa give the Iowa Tests of Basic Skills, a national test not designed to measure state or local content standards, while districts in Nebraska craft their own tests, except for a state writing exam.

In devising the new tests, most states have defied predictions and chosen to go beyond multiple-choice items, by including questions that ask students to construct their own responses.

But many state officials are worried that, over the long run, such decisions could push their assessment programs into the red. In addition, despite efforts aimed at getting test results back faster, they fear that the sheer volume of assessments could generate delays and errors in releasing scores.

In general, states have filled in the gaps in their testing programs with assessments that mirror those in other grades. Many analysts had predicted that, because of costs, states would rely solely on multiple-choice tests for grades and subjects in which they had not tested previously.

A 2003 study by the investigative arm of Congress estimated that it would cost states $1.9 billion to meet the testing requirements over the six years of the federal law's authorization if they relied solely on multiple-choice questions that could be machine-scored. But it would cost states $3.9 billion if they used a mix of multiple-choice and open-ended items, and up to $5.3 billion if the tests required hand-scored, written responses, according to the agency, now called the Government Accountability Office.

Amy Root-Donle for Education Week

Empty testing envelopes sit at a contractor's warehouse in Dover, N.H.

States have generally filled in the gaps in their testing programs with assessments that mirror those in other grades. Many analysts had predicted that, because of costs, states would rely solely on multiple-choice tests for grades and subjects in which they had not tested previously.

"We're developing tests that have the same format, blueprint, rigor, as the tests that we already have in place," said Jeffrey Nellhaus, the deputy commissioner in the Massachusetts Department of Education. The reading and math tests require students to construct their own responses to some items, in addition to answering multiple-choice questions.

The Massachusetts program is expected to cost "somewhere in the order of $10 million to $12 million a year," Mr. Nellhaus estimated, compared with the just over $7.6 million the state is receiving from the federal government this year to cover such costs.

Beyond Multiple Choice

In Nevada, which added reading and math tests in grades 4, 6, and 7 this year, the money has so far been adequate to create tests with a mix of multiple-choice and constructed-response questions, said Paul M. La Marca, the state's assistant deputy superintendent.

"We know that tests are driving curriculum, to some extent, so we think it's important to have items that stretch the cognitive demand of the students," he said. "You can do that with multiple-choice items, but you have a better chance of doing it with other types of items."

Still, Mr. La Marca said, the state doesn't have as many constructed-response items as it might want. "The balance of the tests is skewed toward multiple-choice," he said. "It's more than just a cost issue. In our state, we have significant pressure for quick turnaround time, so that almost hamstrings us a little bit."

New Jersey officials announced Nov. 16 that, in addition to the tests the state already gives in grades 3, 4, 8, and high school, it would add a commercial test aligned to its reading and math standards in grades 5, 6, and 7 for the 2005-06 school year as an interim measure, while it works to revise its entire testing program to provide better diagnostic information for educators.

"Ultimately, we want to build a more robust, more rigorous state system, hopefully to incorporate a performance assessment which we have been piloting for the last three years," said Acting Commissioner of Education Lucille E. Davy. She said the state worked closely with the U.S. Department of Education to ensure the plan could meet federal requirements, although the system still has to go through a peer-review process before it can be approved, as is true for all states.

In Mississippi, meanwhile, officials decided to drop all short-answer questions in the 2005-06 school year to facilitate speedier scoring of test results. It now has only multiple-choice items, except for a state writing test.

Kansas has suspended the use of all items that require an extended response from students and can't be machine-scored.

"It doesn't mean in the future we may not add those items," Ms. Posny said. "We want to figure out how we can do that and still … score [the test] online."

Fifteen states in total will rely solely on multiple-choice items to measure student knowledge, with the exception of their writing

Federal Review Puts State Tests Under Scrutiny

BY LYNN OLSON

By the end of this month, 26 states will have undergone a "peer review" to determine whether their standards and tests meet the requirements of the federal No Child Left Behind Act.

The reviews, conducted by a team of at least three experts in the fields of standards and assessment, are required under the law. The reviewers do not look at the standards and tests themselves, but at documents showing that the assessments meet the law's requirements.

As of Nov. 21, the U.S. Department of Education had posted letters to six states—Maryland, North Carolina, South Carolina, South Dakota, Texas, and West Virginia—on its Web site granting them "deferred" or "final review pending" status under the law.

To receive deferred approval, a state must be able to fully implement its standards and tests this school year; "final review pending" indicates the state still has not met the preponderance of NCLB testing requirements and must submit more evidence. Such documentation can range from technical reports or test manuals, to state statutes and regulations, to memos summarizing the testing program.

One of the issues giving states trouble is a requirement to provide "performance descriptors" that explain the competencies a student must master in mathematics or reading to reach a particular performance level, such as "proficient."

Those descriptions must pertain to specific academic content, said Sue Rigney, an education specialist at the Education Department. "What's not acceptable is to see these very generic descriptors that are the same across grade levels and content areas," she said.

States also are struggling to prove the quality of their alternate assessments for students with disabilities who cannot take the regular exams, and those tests' link to state standards.

States were required to have alternate assessments in place by 2001 under the prior reauthorization of the Elementary and Secondary Education Act. But it wasn't until last summer that the department provided guidance about the criteria for such tests if they are pegged to other than a grade-level standard.

As a result, said Rachel Quenemoen, a senior research fellow at the National Center on Education Outcomes, based at the University of Minnesota-Twin Cities, states have been designing alternate assessments during a period of constantly changing policy and emerging research.

"The states have come a long, long way," she said, "but the depth of the research and the attention that it's gotten is very, very slim. The conditions under which states were working with alternate assessments have changed dramatically."

'Really Encouraged'

As they add tests this school year to comply with the nearly 4-year-old NCLB law, states must design performance standards for those new tests that mesh with those already set for other grades.

In many cases, that will require states to revisit their existing cut-off scores, so that students who perform well in one grade can reasonably be expected to perform well in the next.

Amy Root-Donle for Education Week

Boxes of tests arrive in the scanning room a

Much of that standards-setting will occur over the summer of 2006, after the new tests are given for the first time this coming spring. That will require the federal Education Department to gather additional evidence from most states before it can give full approval to their systems.

Even so, said Kerri L. Briggs, a senior policy adviser in the office of the deputy U.S. secretary of education, "I think we're really encouraged at this point about where states are."

Fewer than a dozen states, she

Special Report: Testing Takes Off

States Adding Tests in 2005-06

	READING						MATH					
Grades:	3	4	5	6	7	8	3	4	5	6	7	8
Connecticut	•		•		•		•		•		•	
Illinois		•		•	•			•		•	•	
Kansas	•	•		•	•		•		•	•		•
Kentucky	•		•	•		•	•	•		•	•	
Maine	•		•	•	•		•		•	•	•	
Massachusetts			•	•		•	•		•		•	
Michigan	•		•	•		•	•		•	•	•	
Minnesota		•		•		•		•		•		•
Missouri		•	•	•		•	•		•	•	•	
Montana	•		•	•	•		•		•	•	•	
Nevada		•		•	•			•		•	•	
New Hampshire	•	•	•	•	•	•	•	•	•	•	•	•
New Jersey			•	•	•				•	•	•	
New York	•		•	•	•		•		•	•	•	
Ohio		•		•	•			•	•	•		
Oklahoma				•	•					•	•	
Pennsylvania		•		•	•				•		•	•
Rhode Island	•	•	•	•	•	•	•	•	•	•	•	•
Vermont	•	•	•	•	•	•	•	•	•	•	•	•
Virginia		•		•	•			•		•	•	
Washington	•		•	•		•	•		•	•		•
Wisconsin	•		•	•	•		•		•	•	•	
Wyoming	•		•	•	•		•		•	•	•	

SOURCE: Editorial Projects in Education Research Center

FOOTNOTES:

1 The District of Columbia previously used the SAT-9 in grades 1-11. In 2005-06, it will use a standards-based exam in grades 3-8 and 10.

2 Louisiana previously administered the Iowa Test of Basic Skills. In 2005-06, it will introduce an assessment program that combines the ITBS and criterion-referenced items, in grades 3, 5, 6, 7, and 9.

3 New Hampshire, Rhode Island, and Vermont suspended testing for 2004-05. These states have implemented jointly developed assessments in grades 3-8 in 2005-06.

Testing Changes At a Glance

	NUMBER OF STATES
States testing in grades 3-8, high school	48*
States adding tests in reading or math	23
States with all multiple-choice except for writing	15

*Does not include the District of Columbia.
SOURCE: EPE Research Center

tests, according to the EPE Research Center survey: Arizona, California, Georgia, Idaho, Iowa, Kansas, Mississippi, North Carolina, Oklahoma, Oregon, South Dakota, Tennessee, Texas, Virginia, and Utah.

The District of Columbia this year switched from a multiple-choice test that was not aligned with its academic-content standards to a new set of standards and tests based on those given in Massachusetts.

Costs a Concern

Although many states said the federal government has provided enough money to cover the development of new tests, they worry about the costs in future years.

Washington state's tests include a 50-50 mix of multiple-choice and open-ended questions. While federal funding has so far been sufficient to expand the tests to grades 3, 5, 6, and 8, said Greg B. Hall, the state's assistant superintendent for assessment and research, starting next fiscal year the state will run a deficit in its testing budget that is expected to increase over time.

CONTINUED ON PAGE 12

As states undergo federal "peer review" of their standards and testing systems under the No Child Left Behind Act, the U.S. Department of Education says it will insist that all states comply with the law.

at Dover, N.H.-based Measured Progress.

noted, had received final approval of their testing systems under the previous reauthorization of the federal education law, in 1994. This time around, the department has made it clear that waivers of the law's testing requirements will not be acceptable.

"From the get-go, we've been really serious about this provision, in particular," said Ms. Briggs, "and we have every intention of implementing it."

Special Report: Testing Takes Off

As States Add Tests, Concerns Rise Over Costs and Logistical Problems

CONTINUED FROM PAGE 11

The reason, he said, is the cost involved in scoring so many additional open-ended items, "which we haven't had to do yet."

Rhode Island, New Hampshire, and Vermont have jointly developed grade 3-8 reading and math tests to meet the federal law's requirements. While federal aid has covered those costs so far, "we're all worried about what will happen when this money goes away," said Mary Ann Snider, the director of assessment and accountability for the Rhode Island Department of Education.

In August, Connecticut became the first and, so far, only state to sue the federal government over the No Child Left Behind law, charging that federal funding falls short of what is needed to meet the law's requirements.

Connecticut officials have sought unsuccessfully to get out of expanding their testing in core subjects beyond grades 4, 6, and 8. An estimate by the state education department pegs the cost of providing tests in the additional grades required under the NCLB law at $41.6 million by 2008, compared with $33.6 million that the state is slated to receive from the federal government by then for testing. *(See Education Week, Aug. 31, 2005.)*

The annual testing requirement is a linchpin of the federal law. Schools and districts are required to meet annual performance targets for their student populations as a whole and for certain subgroups of students. Those that receive federal Title I money and that fail to meet their targets for two or more years face penalties.

Earlier Testing Dates

At least some states are shifting their testing dates to try to get results back sooner. New Hampshire, Rhode Island, and Vermont started giving their new jointly developed tests in the fall, rather than in the spring.

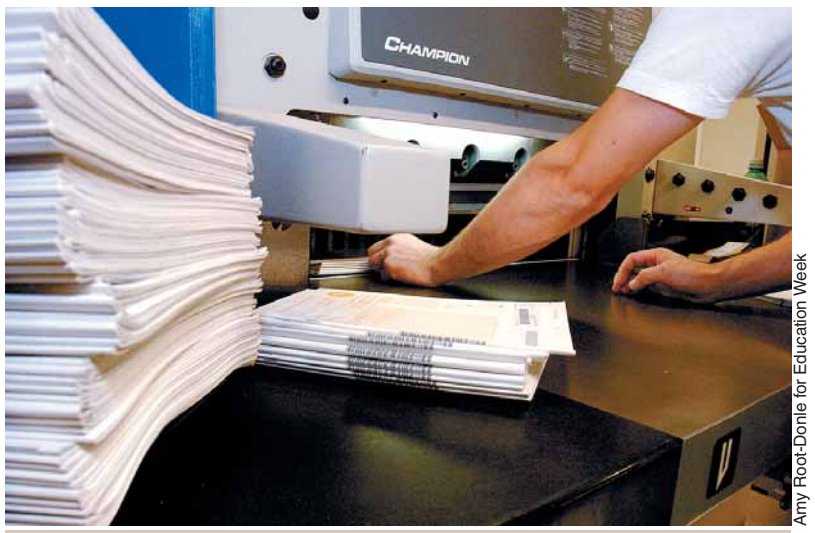
That timing was intended, in part, to ensure that schools would know by the winter whether they have met their performance targets under the federal law and would be subject to any of the law's sanctions, Ms. Snider said.

Michigan shifted from January to October testing, but not without some mishaps. "The idea is to give data back to teachers while they still have the students, and they still have the energy to do something about the results," said Edward D. Roeber, who directs the state testing program.

Still, the change "cut out four months in the test-preparation schedule, which made the summer very challenging," he said. And when the state's test contractor, Pearson Education Inc., failed to get enough tests delivered to school districts on time, the state was forced to extend the testing window by two weeks.

Other states have made more minor adjustments. Kansas, for instance, moved its testing dates up by one week, to March 1. It also has put its writing, science, history, and government tests on hold for a year, while it gets the new reading and math exams in place.

Massachusetts consolidated all of its reading tests in April, so that they would not compete with tests in other subjects during May. New York state adopted a flexible schedule for the statewide administration of its grades 3-8 exams for this school year, to permit schools to give one session per day at each grade level. Maine has opted to replace an existing state test with the SAT college-admissions exam at the high school level, starting next spring.



Amy Root-Donle for Education Week

A worker inserts bundles of tests into a binding-removal machine.

State officials anticipate that it will take longer to report results this school year, as they try to craft new performance standards in additional grades and cope with the extensive amounts of data. Another concern is that the sheer volume of tests will increase the likelihood of errors in everything from the production of test booklets to the processing and reporting of results.

Even so, state officials generally anticipate that it will take longer to report results this school year, as they try to craft new performance standards in the additional grades and cope with the extensive amounts of data.

"It will be delayed, there is no doubt in my mind," said Ms. Posny of Kansas.

Potential for Error

One concern is that the sheer volume of tests will, inevitably, create logistical problems for schools and an increase in administration and scoring errors.

"The more you add to the test contractors' plate and the quicker they have to report the results back, the probability goes up that errors are going to be made," said George Madaus, a professor emeritus of education at Boston College who co-wrote a May 2003 report on the widespread errors in standardized tests. "You're stretching the capacity of a limited number of companies that do this work."

Among the states that have suffered from scoring glitches is Nevada. In 2002, the state board of education required the San Antonio-based Harcourt Educational Measurement to pay penalties totaling $425,000 because of a mistake that threw off the scores of nearly 31,000 students who had taken the state's high school exit exam in math.

"We're very worried about accuracy from the vendor because we have, unfortunately, been snakebitten," Mr. La Marca said.

In Michigan, which is giving 216 different test forms in grades 3-8 this year, state officials are doing "a lot of extra checking," said Mr. Roeber.

"There are just more chances to screw up," he said. "My staff and I have been working 18 hours a day and some weekends."

And state officials aren't alone. "School districts are kind of reeling," Mr. Roeber said.

One issue for schools is simply having enough staff members to administer the tests and to provide accommodations, such as more time, for all the students who need them because of disabilities or limited English skills.

"As with most states, we have an extensive list of accommodations, which is great for students but puts the school staff in a difficult position, because many of the accommodations require alternate settings and additional staff," said Tim Kurtz, the director of assessments for the New Hampshire Department of Education. "Every single adult human being is involved in testing somehow."

At least one thing is clear: With more tests, in more grades, soon many more teachers will be focused on test results.

"The existing evidence suggests that when a grade level is tested, teachers pay a lot more attention to what's on the test," said Joan L. Herman, the co-director of the Center for Research on Evaluation, Standards, and Student Testing, or CRESST, at the University of California, Los Angeles.

"So as more grade levels are being tested," she said, "you can now expect that every teacher will be paying attention to what's on the test and, in the best case, aligning their instruction with standards and, in the worst case, engaging kids in a curriculum of test preparation."

Special Report: Testing Takes Off

Benchmark Assessments Offer Regular Checkups On Student Achievement

BY LYNN OLSON

School districts worried about how students will perform on end-of-the-year state tests are increasingly administering "benchmark assessments" throughout the year to measure students' progress and provide teachers with data about how to adjust instruction.

Nearly seven in 10 superintendents surveyed for *Education Week* this past summer said they periodically give districtwide tests, and another 10 percent said they planned to do so this school year. Such tests typically are aligned to state or district standards for academic content and given three to five times during the year. Some are given as often as monthly.

Most benchmark assessments take one hour each for reading and mathematics, but may include other subjects. Extensive reporting systems break down test results by the same student categories required under the federal No Child Left Behind Act, such as by race, income, disability, and English proficiency, in addition to providing individual progress reports at the district, school, classroom, and student levels.

"I do believe that three years from now, certainly five years from now, no one will remember a time when there weren't benchmarks," said Robert E. Slavin, the director of the Center for Data-Driven Reform in Education, at Johns Hopkins University.

Burgeoning Market

That's certainly what test vendors hope. Last year, Eduventures Inc., a market-research firm based in Boston, identified benchmark assessments as one of two high-growth areas in the assessment industry, alongside state exams, with a compound annual growth rate of greater than 15 percent. The company predicted that by 2006, what it called "the formative-assessment market"—using a term sometimes treated as a synonym for benchmark assessment—would generate $323 million in annual revenues for vendors.

But while many assessment experts agree that the idea of frequent testing of students to monitor their learning and adjust instruction is sound, some also warn that districts should take a close look at what they're getting for their money and how they are using such exams.

"You might say that the message here is, 'Get a second opinion,'" said Grant Wiggins, the president of Authentic Education, a Hopewell, N.J.-based consulting service that works with districts.

It's no secret why districts are turning to benchmark tests. The No Child Left Behind Act, signed into law by President Bush in January 2002, and states' own accountability systems have created a high-stakes environment in which both districts and schools can face penalties for failing to meet performance targets.

In this standards-based environment, the feeling is that the sooner and more often schools have information about how they're doing against the standards, the better.

"The reason that there is a boom in benchmark assessments is that most states and school systems are providing nothing more than autopsy reports right now," said Douglas B. Reeves, the founder of the Center for Performance Assessment, a private consulting organization based in Denver that works with districts to design fair and rigorous assessments and classroom activities. "They tell you why the patient died at the end of the year, and then marveled that the patient didn't get any better."

Studies by the Washington-based Council of the Great City Schools, the Austin, Texas-based National Center for Educational Accountability, and others have found that one feature of high-achieving districts is their use of periodic, benchmark assessments to track student achievement and make adjustments.

"Good formative assessments, good benchmark assessments," Mr. Reeves said, "provide feedback throughout the year, and that is far more fair to principals and teachers, provided they are used wisely."

Vendors Vary

In the past few years, according to Eduventures' 2004 report, "Testing in Flux," new competitors have flooded the formative-assessment market, including:

- Major test publishers, such as the New York City-based CTB/McGraw-Hill and the San Antonio-based Harcourt Assessment;
- Test-preparation companies, including the New York City-based Princeton Review;
- For-profit providers that specialize in linking assessment results with prescribed remediation plans and curricula, such as the San Diego-based Compass Learning and the New York City-based Kaplan K-12 Learning Services;
- Nonprofit organizations, such as the Portland, Ore.-based Northwest Evaluation Association; and
- Suppliers of "whole-school-reform models," such as the New York City-based Edison Schools Inc. and Mr. Slavin's Baltimore-based Success for All Foundation,

CONTINUED ON PAGE 14

Christopher Powers/Education Week

Silas Bender, a 3rd grader at London Towne Elementary School in Centreville, Va., takes a benchmark test.

Not All Teachers Keen on Periodic Tests

BY LYNN OLSON

John W. Hutcheson now teaches in a private Montessori school in Sammamish, Wash., after spending 25 years teaching in the Dallas school district. Looking back, he says the Texas district's thrice-yearly benchmark assessments helped drive him out.

"The benchmarks themselves are a reflection of the standardized exams," Mr. Hutcheson said, "which are only a small piece of learning. You progressively keep narrowing the curriculum down, so we end up preparing students for a world that doesn't exist."

Across the country, school districts are adopting benchmark assessments to help teachers modify instruction over the course of a school year. Yet many teachers remain wary. Like Mr. Hutcheson, they say their experience with such tests has been anything but positive.

In Philadelphia, a social studies teacher who asked not to be named said he found the use of benchmark assessments there "incredibly restricting and unrealistic."

As part of a core high school curriculum, the 214,000-student school system uses a program involving multiple-choice tests given every six weeks, with immediate feedback to teachers and schools via a Web-based system of data analysis and reporting. The district describes the new standardized, college-preparatory curriculum and the related system of assessments as a critical element of its plan to improve secondary education. *(See Education Week, Feb. 9, 2005.)*

"Students found them totally meaningless and very intrusive, because it was another interruption, in addition to all the other testing," he said.

Mr. Hutcheson also complained of the time and stress associated with the tests used in Dallas. "We would spend entire afternoons analyzing benchmark results," he said. "The district, every time the kids took the test, would print up a thorough record of how many answers they missed, the answers they put down, a list of subskills to be worked on, and a complete analysis of each test."

Dallas school officials were unable to comment by press time.

Some districts have reported impressive results using similar methods.

Some classroom teachers see benchmark assessments as time-consuming intrusions that are exacerbating pressure to narrow the school curriculum to focus solely on subjects found on states' standardized tests.

When the Norfolk, Va., school district walked away with the $500,000 Broad Prize in Urban Education this year, it was largely on the strength of its gains in reading and math scores and its progress in closing racial and ethnic achievement gaps. Officials there pointed to the strong focus on data-driven instruction as one key to the district's success.

The 36,700-student district requires quarterly benchmark assessments in all grades. Ninety percent of Norfolk's schools also have developed common assessments that teachers give monthly. And teachers regularly meet in "data teams" to review the data, draw up common plans, and adjust instruction.

Over the past several years, the 12,000-student Santa Monica, Calif., school district has used a mix of teacher-designed tests and assessments linked to its adopted textbooks at the elementary school level. This year, secondary school teachers are meeting in departmental teams across sites to develop what the district is calling formative assessments in English, mathematics, science, and social studies that they'll agree to give in common about three times a year.

"These are for teachers to really help guide their instruction," said Maureen L. Bradford, the district's director of educational services. "We feel like there probably isn't something off the shelf that's going to work for us; that teachers really need to come to one mind about what's important to teach, and when to teach it and how to assess it appropriately. It's a tremendous amount of work."

Carol Jago, who chairs the English department at Santa Monica High School, praised the approach the school system is taking to developing the tests. "I hope we're going to end up with essays or something that's really authentic," she said.

Still, Ms. Jago is worried.

"Inevitably, any time you try to institutionalize it, it becomes one more summative assessment that just happens before the state assessment," she said, referring to a test given after teaching in the subject is completed. "So it's right-headed, but I don't think it's something you can actually do properly because of the nature of the beast."

Special Report: Testing Takes Off

Growing Demand for Benchmark Tests Fuels Policy Debate

CONTINUED FROM PAGE 13

which designed the 4Sight assessment series.

The products of such suppliers range from formatted tests linked to the standards in individual states, to item banks that districts and schools can use to develop their own assessments, to online testing, scoring, and reporting systems.

Skimming the Surface?

Lorrie A. Shepard, the dean of the school of education at the University of Colorado at Boulder, voices caution about the trend.

While "not all formal benchmarking systems are bad," she said, she worries about the effects of using 15- or 20-item multiple-choice tests that mirror the format of state exams to drive classroom instruction.

Previous research by Ms. Shepard and others has found that students who do well on one set of standardized tests do not perform as well on other measures of the same content, suggesting that they have not acquired a deep understanding.

"The data-driven-instruction fad means earlier and earlier versions of external tests being administered at quarterly or monthly intervals," Ms. Shepard said. "The result is a long list of discrete skill deficiencies requiring inexperienced teachers to give 1,000 mini-lessons."

Good benchmark assessments, she suggested, should include rich representations of the content students are expected to master, be connected to specific teaching units, provide clear and specific feedback to teachers so that they know how to help students improve, and discourage narrow test-preparation strategies.

Rather than trying to assess everything, added Mr. Reeves, the best benchmark tests focus on the most important state or district content standards. And they provide results almost immediately, in simple, easy-to-use formats, he said.

The National Center for Educational Accountability stresses that good benchmark assessments measure performance "on the entire curriculum at a deep level of understanding." They also begin before grade 3 in both reading and math and provide a process to ensure that data on student performance are reviewed and acted upon by both districts and schools, the center says. In addition to such tests, it adds, districts may provide unit or weekly assessments that principals and teachers can use to monitor student progress.

Approaches Differ

But in talking about benchmark assessments, not everyone means the same thing.

According to Mr. Slavin, some benchmark tests, like 4Sight, are designed primarily to predict students' performance on end-of-the-year state exams. They measure the same set of knowledge and skills at several points during the school year to see if students are making progress and to provide an early warning of potential problems.

Other benchmarks are tied more closely to the curriculum, and to the knowledge and skills students are supposed to have learned by a particular time. For example, a skill-by-skill benchmark series in math might focus on fractions in November, decimals in January, geometry in March, and problem-solving in May, rather than testing all skills at the same time, Mr. Slavin said.

Such benchmarks serve as pacing guides for teachers and schools, providing information on whether students have learned the curriculum they've just been taught. Some companies claim their tests serve both purposes, predicting students' ultimate success on state tests and gauging how they're progressing through the curriculum.

Historically, vendors would design one set of benchmark tests for the entire country. Now they craft tests for each state, starting with the larger ones.

Many companies also work with districts to design the districts' own assessments, tied to state and district standards, or permit districts and schools to modify previously formatted exams. Some vendors provide large, computerized pools of item banks that teachers and schools can use to create their own classroom tests and check students' progress on state standards.

Stuart R. Kahl, the president of Measured Progress, a Dover, N.H.-based testing company, says that while item banks hold great promise, because they permit teachers to design tests that can be used during the ongoing flow of instruction, one issue is whether teachers are prepared to use them appropriately.

"Now we're putting individual items in the hands of teachers," he said, "saying, 'You construct the test; make it as long or as short as you want.' Do we think they have the understanding to know how much stock they can put in the generalizations they make from such exams?"

Some also worry that as vendors have rushed in, quality has not kept pace. The Eduventures report noted that many vendors have marketed formative assessments "on the basis of the quantity of exam items, as opposed to those items' quality." For example, companies may tout having tens of thousands of exam items, it said, although many of the items have not been extensively field-tested or undergone a rigorous psychometric review.

"I think vendors in our space have found it challenging," said Marissa A. Larsen, the senior product manager for assessment at the Bloomington, Minn.-based Plato Learning Inc., whose eduTest online assessment system is now used in more than 3,000 schools.

While districts sometimes apply the same psychometric standards to benchmark tests that are applied to high-stakes state exams, she said, "in many cases, that's not what vendors in this space are trying to do. If we did that, it would be well beyond what districts could afford to buy for formative systems."

Critics also say that even the best benchmark assessments are more accurately described as "early warning" or "mini-summative" tests, rather than as true "formative" assessments, which are meant to help adjust teaching and learning as it's occurring. In contrast, summative tests are designed to measure what students have learned after instruction on a subject is completed.

"Formative assessments are while you're still teaching the topic, providing on-the-spot corrections," said Mr. Kahl. "With benchmark assessments, you're finished. You've moved on. Not that you don't get individual student information, but at that stage, it's remediation."

What Is 'Formative'?

Yet Eric Bassett, the research director for Eduventures, said the terms formative and benchmark assessments are often used interchangeably in the commercial education market.

And that, some critics say, is precisely the problem.

"I recognize that I've lost the battle over the meaning of the term 'formative assessment,' " said Dylan Wiliam, a senior researcher at the Educational Testing Service, based in Princeton, N.J.

In the 1990s, he wrote an influential review that found that improving the formative assessments teachers used dramatically boosted student achievement and motivation. Now that same evidence, he fears, is being used to support claims about the long-term benefits of benchmark assessments that have yet to be proven. "There's a lack of intellectual honesty there," Mr. Wiliam said. "We just don't know if this stuff works."

He and others say the money, time, and energy invested in benchmark assessments could divert attention from the more potent lever of changing what teachers do in classrooms each day, such as the types of questions they ask students and how they comment on students' papers.

"If you're looking, as you should be, at the full range of development that you want kids to engage in, you're going to have to look at their work products, their compositions, their math problem-solving, their science and social-studies performance," said Mr. Slavin of Johns Hopkins.

Mr. Wiggins of Authentic Education said that while some commercially produced benchmark assessments are far from ideal, they're better than nothing. "I would rather see a district mobilizing people to analyze results more frequently," he said. "That's all to the good."

The key point, he and others stress, is what use is made of the data.

"It's only a diagnosis," Mr. Slavin said. "If you don't do anything about it, it's like going to the doctor and getting all the lab tests, and not taking the drug."

What Are Benchmark Assessments?

While not everyone means the same thing by the term, benchmark assessments typically:

- Are given periodically, from three times a year to as often as once a month;
- Focus on reading and mathematics skills, taking about an hour per subject;
- Reflect state or district academic-content standards; and
- Measure students' progress through the curriculum and/or on material in state exams.

SOURCE: *Education Week*

Christopher Powers/Education Week

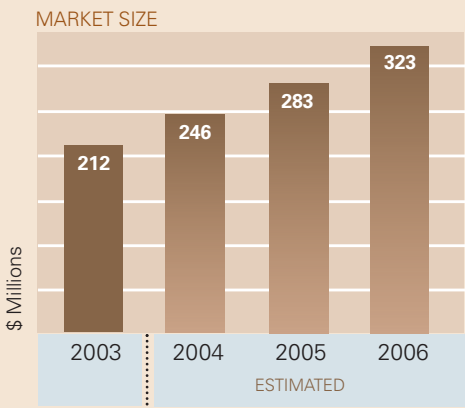
Victoria Todd, a 3rd grader at London Towne Elementary, finishes one of 38 problems on a Benchmark Assessment Resource Tool test.

Analysts see benchmark assessments as one of two high-growth areas in the testing industry, alongside state exams. Experts say good benchmark tests provide valuable feedback on students' mastery of key knowledge and skills. But some worry that as vendors have rushed in, quality has not kept pace.

Benchmark-Test Market Foresees Growth

A 2004 report predicted that the market for benchmark or formative assessments would expand by a compound annual growth rate of more than 15 percent from 2003 to 2006.

SOURCE: Eduventures Inc.



TEST MARKET

New competitors have emerged in recent years to supply school districts with benchmark assessments. They include:

MAJOR TEST PUBLISHERS, such as CTB/McGraw-Hill, based in New York City, and the San Antonio-based Harcourt Assessment;

TEST-PREPARATION COMPANIES, including the Princeton Review, based in New York City;

SUPPLIERS of whole-school-reform models, such as Edison Schools Inc., of New York, and the Success for All Foundation, of Baltimore.

FOR-PROFIT PROVIDERS that specialize in linking assessment results with prescribed remediation plans and curricula, such as the San Diego-based Compass Learning and the New York City-based Kaplan K-12 Learning Services;

NONPROFIT ORGANIZATIONS, such as the Northwest Evaluation Association, in Portland, Ore.

SOURCE: Eduventures Inc., *Education Week*