

Quality of Questions on Common Tests at Issue

By [Stephen Sawchuk](#)

Most experts in the testing community have presumed that the \$350 million promised by the U.S. Department of Education to support common assessments would promote those that made greater use of open-ended items capable of measuring higher-order critical-thinking skills.

But as measurement experts consider the multitude of possibilities for an assessment system based more heavily on such questions, they also are beginning to reflect on practical obstacles to doing so.

The issues now on the table include the added expense of those items, as well as sensitive questions about who should be charged with the task of scoring them and whether they will prove reliable enough for high-stakes decisions. Also being confronted are matters of governance—the quandary of which entities would actually “own” any new assessments created in common by states and whether working in state consortia would generate savings.

“The reality is that it does cost more to base a system on open-ended items, no question about it,” said Scott Marion, the vice president of the Dover, N.H.-based Center for Assessment, a test-consulting group, who is advising several states. “If the model we’re thinking about has got to be on-demand and high-stakes and used in systems with scores that are returned quickly, then it’s going to cost a lot.”

Higher Costs?

State dependence on multiple-choice testing under the federal No Child Left Behind Act has led to a backlash by those who say the tests, while cheap and technically reliable, come at a cost: not measuring complex cognitive skills.

Using a slice of money from the \$4.35 billion Race to the Top Fund, created last year under the American Recovery and Reinvestment Act, U.S. Secretary of Education Arne Duncan has called for state consortia to craft richer item types aligned to common standards that would include constructed-response questions, extended tasks, and performance-based items in which students would apply knowledge in new ways. (“[Stimulus Seeks Enriched Tests](#),” Aug. 12, 2009.)

[← Back to Story](#)

Socially relevant,
fully interactive
video games for teens

VIDEO GAMES THAT MAKE A DIFFERENCE

WILL
CAMPUS

CLICK HERE
TO LEARN MORE

The advertisement features a green header with the text 'Socially relevant, fully interactive video games for teens'. Below this is a black banner with the text 'VIDEO GAMES THAT MAKE A DIFFERENCE'. The main image shows four young people's faces in a row, each in a small white frame. At the bottom, there is a black box with the 'WILL CAMPUS' logo on the left and the text 'CLICK HERE TO LEARN MORE' on the right.

His department is expected to open up a competition in March for the assessment aid from the stimulus law. Work is under way, meanwhile, on a national project to produce a "common core" of academic standards for adoption by states.

Assessment experts caution that open-ended test items carry with them a number of practical challenges. For one, items that measure higher-order skills are generally more expensive to devise, depending on how extensive the item is and how much of the total test such items make up.

For instance, with their detailed prompts and scenarios, questions that require students to engage in extensive writing or to defend their answer choices often are "memorable," meaning the items can't be reused for many years and must be replaced in the meantime.

Wes Bruce, the chief assessment officer for the Indiana education department, recalled one prompt that required 5th graders to write about what would happen if a kangaroo bounded into the classroom.

"All across the state, kids were talking about the prompt," he said. "From an assessment perspective, that's not good. Teachers will use it [subsequently] as an example for classroom work."

The scoring process for open-ended items is also far more complicated than sticking a bunch of test papers into a computer scanner. It relies on "hand scorers" who are trained according to a scoring guide for each question and a set of "anchor papers" that give examples of performance on the item at each level. Each open-ended item typically goes through multiple reviews to ensure consistent scoring.

Depending on the complexity of the item and how long it takes to score, the costs can increase dramatically. A short constructed-response item with four possible scores might take one minute to score, said Stuart R. Kahl, the president of Measured Progress, a nonprofit test contractor based in Dover, N.H. But an extended performance-based or portfolio item might take up to an hour, he said. With test scorers paid in the range of \$12 to \$15 an hour, such costs would add up.

For a midsize state with about 500,000 students within the tested grades and subjects, the scoring of tests based even partly on constructed-response items would make up more than a fifth of the total annual contract cost, Mr. Kahl estimated.

Scoring Scenarios

For some, the idea of expanding human-scored items raises issues of reliability: Performance-based items are typically less mathematically reliable than those based entirely on multiple choice.

Todd Farley, a 15-year veteran of the test-scoring business who detailed his experiences in a recent book, *Making the Grades*, is among the skeptics. In the book, Mr. Farley alleges that the scoring guidelines for open-ended items were frequently counterintuitive, and that as a "table leader"—an individual supervising other scorers' work—he occasionally changed other

reviewers' scores.

Though test publishers interviewed for this story dismissed Mr. Farley's account, independent sources do point to areas of concern. At least two reports issued by the Education Department's office of inspector general last year, for instance, found lapses in [Florida's](#) and [Tennessee's](#) oversight of test contractors charged with scoring open-ended items.

In part to ameliorate the errors and costs associated with human scoring, test publishers are investing heavily in automated-response systems that use artificial-intelligence software to "read" student answers.

Such programs are already in use to minimize the number of scorers needed for major tests such as the Graduate Record Examination. A handful of states, including Indiana, have piloted the technology for their own tests, and some experts, like Joan Herman, a director of the National Center for Research on Evaluation, Standards, and Student Testing, or [CRESST](#), a group of assessment researchers headquartered at the University of California, Los Angeles, believe that the technology will be widespread within five years.

"When that happens, it will open up entirely new windows for doing more complex, open-ended items on large-scale assessments," Ms. Herman said.

But such systems are not perfect, and experts including Mr. Bruce, in Indiana, noted that one of their limitations is that they typically don't generate interpretative information about where a student needs to improve.

And in an era in which most teachers are distanced from the assessment process, unions and other stakeholders argue that teachers should have a greater role in the development and scoring of assessments, which can serve an important function for gathering information on student performance.

Teacher scoring of assessments has generally been eschewed under the NCLB law, the current version of the Elementary and Secondary Education Act, in part because of fears that the law's accountability pressure would cause educators to inflate their students' scores. But other countries have tackled the problem by using a blind-scoring process, in which teachers meet in central locations to score exams with names removed.

Hiring substitute teachers and paying the scoring teachers for their time would be costly as well. But there are benefits in the form of professional development, said Marcia Wilbur, the executive director of curriculum and content development for the College Board's Advanced Placement program, which uses such a system to grade open-ended essays on AP exams.

Teachers, Ms. Wilbur said, spend a week looking at sample responses at each level of the scoring guides before they begin to score, and the guides often help them better understand students' learning progressions.

"Teachers will bare their souls about what they do in their classrooms while discussing the samples," she said. "It's not that they go back to their classrooms and teach to the test, but they have a better understanding of the skills and the features of the skills that students

struggle with.”

Another model, drawn from international practice in places such as Australia, England, Hong Kong, and New Zealand, would rely on teachers’ scoring assessments that are embedded within classroom curricular activities. Kentucky used such a system in the days before the 8-year-old NCLB law.

Not only could such assessments provide better feedback on where instruction needs to be improved, but they also could be included in state gauges of achievement if there were a central auditing process to ensure appropriate administration and reliable scoring, CRESST’s Ms. Herman said.

Consortia Challenges

Aside from costs, the problem with systems that rely heavily on teacher scoring, Mr. Kahl argued, is that results cannot be scored as quickly or efficiently as those done by machine. That means it could prove tough to turn results around under the quick timeline envisioned by the NCLB law and current state accountability systems.

Both Ms. Herman and Mr. Kahl said it might be possible to build systems that coupled curriculum-embedded assessments, scored throughout the year, with information from more-typical “on demand” standardized tests. But such a model hasn’t been tested in the United States.

“While most people agree that some amount of analyzing student work is good professional development, every teacher probably doesn’t have to do 100 papers to get the full value of it,” Ms. Herman added.

Also uncertain is which entities would actually own test forms or items developed in common, and which would bear the responsibility for updating them. Under the current system, such decisions are now made through contracts with individual states.

A common test, Mr. Marion of the Center for Assessment said, argues for some kind of open-source setup, but the details are “dicey,” he concedes. “The reality is that if this is paid for out of stimulus funds, the country should own them,” he said, “but at some point, someone has to house these items.”

Mr. Kahl, meanwhile, warns that one of the major selling points about state consortia—cheaper tests—might yield only limited savings for some states. Consortia would probably ease costs more for small states, where test development is a high percentage of overall assessments, but might not help larger states, where most testing costs involve factors such as printing, scoring, and reporting results.

Such questions of governance, finance, and sustainability have drawn the concern of policy experts, too.

Chester E. Finn Jr., the president of the Thomas B. Fordham Institute, warned in a recent article for the Washington think tank’s newsletter that the federal competition could lock in features of an assessment system that would be difficult to change in the future.

He also worries that the Obama administration's ambitious goals for the assessment funding—which include generating information about both school and student performance as well as data about teacher effectiveness—could prove to be irreconcilable.

"If all the glitterati ... remains in the grant competition, anyone that wants to win the competition is going to have to pretend they can do all those things," Mr. Finn said in an interview. "But since we know that they can't all be done by the same assessment, in the same period of time at a finite price, something will get left in the dust."

But for all those challenges, state testing experts hope to see breakthroughs with the federal funding.

"They are expecting that there be some innovation in the assessment area," Mr. Bruce of Indiana said. "As a state that has been committed to using what at one point were innovative item types, and is still looking at ways to innovate in the scoring, it's exciting."

Coverage of the American Recovery and Reinvestment Act is supported in part by grants from the William and Flora Hewlett Foundation, at www.hewlett.org, and the Charles Stewart Mott Foundation, at www.mott.org.

Vol. 29, Issue 19