U.S. Reviews of Standards, Tests Enter New Phase

By Catherine Gewertz
Education Week
August 5, 2014

The U.S. Department of Education is on the verge of releasing the first draft of new guidance on the peer-review process for standards and tests, a document that could exert a powerful influence on how states set academic expectations.

Little known outside the assessment world, the process is wonky and technical. But it is an important tool for the federal agency in reviewing—and shaping—states' academic standards and testing systems.

The draft of updated guidance, expected this month, arrives as most states are trying out or designing new tests to reflect the Common Core State Standards. The testing industry, which crafts those assessments, and state testing directors, who oversee their administration to millions of students, have been waiting anxiously for any sign that the Education Department will change the criteria used to evaluate their systems.

"We’re in this huge transition to a whole new system of tests, and this is one of the only leverage points the department has on what those assessment systems look like," said Anne Hyslop, who has been monitoring the peer-review process as a policy analyst with the New America Foundation, a Washington think tank.

Many in the assessment world are worried, though, because few, if any, of the prominent figures in the field have been asked to help shape the upcoming draft.

At meetings with state schools chiefs and assessment leaders this summer, Education Department officials tried to assuage those worries by repeating that the draft is only a "straw man," intended to prompt input from the field. Once reaction is gathered from experts and the public, the document will be revised and a final version released in early 2015.

Valid and Reliable

States have been undergoing peer review of their standards and assessments since the late 1990s because of requirements in the two most recent incarnations of the federal Elementary and Secondary Education Act: the Improving America's Schools Act of 1994 and the No Child Left Behind Act, signed into law in 2002. Among other criteria, states must show that their tests are aligned with their standards and are valid and reliable for their intended purposes.

The Education Department under President Barack Obama has articulated a vision of testing that goes beyond such provisions, however. The department suspended the peer-review process in December 2012, telling states in a letter that the criteria needed updating in light of assessment capabilities the agency articulated in its Race to the Top assessment competition, which funded two state consortia to design tests for the common standards, and in its No Child Left Behind waiver program, which imposed conditions on states in exchange for exemption from certain tenets of that law.
To be part of those projects, states had to have tests that show how well students are progressing toward college and career readiness, measure skills that previously were hard to measure, and produce data that can be used to judge the effectiveness of teachers, principals, and schools. Just how the department will reshape the criteria to reflect those ideas is a subject of intense interest in key corners of the K-12 world.

There has been talk of including other matters in the criteria as well. Federal education officials have been urged to consider requiring states to show that their tests have appropriate security measures. Internally, department officials have discussed whether to require states' tests to assess writing, a pivotal skill in the common standards, which are now in effect in more than 40 states. Many states' current assessments don't probe students' writing skills.

One federal Education Department official told Education Week that a central idea in developing new criteria is ensuring that states' tests reflect a "depth of knowledge" that might well require "going beyond a multiple-choice answer structure."

The department hopes to move the peer-review process "away from minutiae" to "bigger-picture validity that is predictive of college and career readiness," said the official.

**Difficult Terrain**

Even before the new draft criteria are issued, however, the Education Department is in a politically tricky position because of the controversies that have flared in some states around the common core.

Opponents have argued that the common standards and tests represent a federal intrusion into local education decisions because the department funded the two main testing consortia—the Partnership for Assessment of Readiness for College and Careers, or PARCC, and Smarter Balanced—and offered incentives for states to adopt the standards. Such opposition has led some states to back out of the projects.

"The department is between a rock and a hard place" in setting the peer-review criteria, said a former department official who, like most of the experts interviewed by Education Week for this article, agreed to speak only on condition of anonymity to avoid alienating colleagues.
"If they don't take this responsibility seriously, they realize it could all devolve again into where we were with NCLB, with 50 states, 50 different goal posts, and 50 different ideas of what assessment should look like," the former official said. "On the other hand, by wading in at all, even though it's their legal responsibility to do so, the department once again becomes the lightning rod for claims of federal overreach."

That landscape means reaction to new peer-review criteria in high-level state offices could be very different from what it might have been five years ago.

"States have been going through this process for a long time, but the temperature has been turned way up now," said Andy Smarick, a partner at Bellwether Education Partners, a nonprofit Washington consulting firm. "I wouldn't be surprised if most governors, and many state chiefs, especially new ones, won't understand that this has a long legacy. Many will come to this for the first time, and how many will be upset that the feds are involved in this?"

Opinion on the value of the peer-review process is mixed, since a number of problems have hobbled it in the past. Some wonder whether years of reviewing has done anything to improve standards or assessments.

Michael J. Petrilli, the president of the Thomas B. Fordham Institute, a Washington research and advocacy group, noted that while some states have been respected for high standards and good-quality tests, others have had weak ones.

"There doesn't seem to be any evidence that [peer review] has helped improve assessments in the past," he said. "It's been a waste of time."

Even some policy experts who were central to developing the process acknowledge that long-standing legal restrictions limit its usefulness.

Because federal laws bar the Education Department from controlling the content taught in schools, peer reviewers can't pass judgment on the quality of states' standards. That part of the review is little more than a "check-box exercise" of compliance, said Michael Cohen, who helped develop the process as the department's assistant secretary for elementary and secondary education under President Bill Clinton.

"Not all the [states'] standards were great, but the federal criteria were that tests had to be aligned to standards," said Mr. Cohen, who is now the president of Achieve, a Washington group that advocates higher standards and helped develop the common core.

**A Frequent Complaint**

Peer reviewers don't examine states' actual standards or tests. Instead, they examine evidence—typically, multiple boxes of it—of whether those standards and assessments meet specific requirements of federal law. To evaluate whether a state's standards are "challenging," for instance, peer reviewers might look at documentation of the steps a state took to create rigorous standards.

Still, Mr. Cohen and others said, peer review has tremendous value because it makes states focus more intently on aligning tests to standards and on documenting their tests' technical quality.
A frequent complaint about peer review is inconsistency in findings from state to state.

Michael Hock, the assessment director in Vermont, recently told attendees at the Council of Chief State School Officers' annual assessment conference that states in the New England Common Assessment Program all used the same test, but got differing evaluations of that test from peer-review teams.

States' experience going through peer review depended a lot on which set of peer reviewers, and which Education Department staff members, they were assigned to work with, said William J. Erpenbach, who served as a peer reviewer under three presidential administrations and has advised many states as they prepare their materials for submission.

Experts say the process has been undermined, too, by weakness on the key question of "validity"—whether tests are designed appropriately for the way states want to use them. They cited both inadequate proof of validity by states, and insufficient demands by reviewers for stronger evidence.

"They never really attended to validity," a senior-level source in the assessment industry said of the reviews.

To complicate the situation further, federal education officials' concept of validity has evolved to emphasize predictive ability, experts say. It's not enough anymore for a state to show that a test is a valid indicator of a middle school student's math skills; it must show that the test is a good predictor of whether that student is "on track" to be college-ready in a few more years.

The field also increasingly seeks "more sophisticated" evidence of validity, said Ellen Forte, who served as a department peer reviewer and now advises states on their assessment systems as the president, CEO, and chief scientist at edCount, a Washington consulting firm.

"Now [the field is] working at a much finer grain size, going deep into the domain and its skills," said Ms. Forte. She isn't convinced, she added, that the peer-review process demands the kinds of evidence that form "the backbone of validity."

Similarly, peer reviews have often sought to determine a test's alignment to standards based on whether most of the standards are found in the assessment items. That's a lower level of alignment than federal officials seem to be seeking now when they describe high-quality assessments as measuring deeper, more nuanced levels of student achievement, experts say.

"[Peer review] has looked at alignment only on a superficial level," said the senior-level assessment source. "Now, defining what alignment means will be big deal. If a test doesn't reflect the intended depth of knowledge of the standards, it will be found wanting."

**Alignment in Question**

Several people interviewed for this story said that if the new criteria don't require states' tests to reflect the writing skills in the common core, such as citing evidence from text to support an argument, the federal government will, in effect, be allowing states to use tests that aren't aligned to those standards.
"If kids aren't writing and drawing evidence from text, then a test on its face isn't aligned" to the common core, Mr. Cohen of Achieve said. "You don't need an elaborate set of criteria to figure that out."

Some educators question the value of peer review in part because states rarely experience penalties when their tests fall short of full approval. And many do fall short: At one point in 2002, only 19 states' systems met federal criteria. Between 2010 and 2012, 15 to 20 states' systems had not obtained even conditional approval.

And while some states have had to submit to compliance agreements, few have ever paid the ultimate penalty for unapproved standards or tests: forgoing a portion of their federal Title I administrative funds.

**Need For Expertise**

In addition to concerns about a lack of outside input in developing the forthcoming criteria, many in the assessment field are worried that the Education Department currently lacks the right kinds of expertise to craft good criteria for assessments. Key staff members with backgrounds in measurement or large-scale assessment, such as Carlos Martinez and Sue Rigney, who oversaw peer review in recent years, have retired or changed jobs within the department.

"The department is now in a place where it's far less capable" of designing the right criteria and supporting states in building good testing systems, said Ms. Forte, the edCount executive.

Department officials did not respond to a request for comment on that question of capacity, or on concerns in the field as it undertakes revisions of the peer-review process.

Many state and testing-industry officials who were interviewed for this story said they'd like to see the peer-review process evolve into an ongoing relationship of technical support. They'd also like to see it become more open and collaborative.

During some periods of peer review, state officials have been allowed to speak directly with their reviewers. But during other periods, no face-to-face communication was permitted. States simply received decision letters from the federal department.

Mr. Erpenbach said that when he was allowed to sit down and talk with state officials during the process, he was often able to resolve many issues.

Some in the assessment world have worried privately that the Education Department's new criteria might set forth requirements that only its own grantees—PARCC and Smarter Balanced—could meet. That would pose big problems, since nearly half the states plan to use other tests in 2014-15.

"It's really important in this process that we stay open to other solutions that could meet the criteria," said Chris Minnich, the executive director of the CCSSO, which co-led the common-standards initiative.