Measuring what really matters

Halting testing altogether won't fix our flawed assessment and accountability systems. Instead, improve the system of assessments so they measure what matters and ensure that what matters gets taught in public schools.

By Ruth Chung Wei, Raymond L. Pecheone, and Katherine L. Wilczak

In the past 15 years since the passage of No Child Left Behind, large-scale assessments have come to play a central role in federal and state education accountability systems. Across the country, districts, schools, and now teachers are evaluated on the basis of their ability to raise test scores for all students from year to year.

Opponents of NCLB and high-stakes testing have long argued that testing more students more frequently won't improve instruction or learning. Opponents also argue that high-stakes testing hurts students by causing test-related anxiety, driving instruction in undesirable ways (such as teaching to the test and test-prep), and leading schools to narrow the curriculum to focus on the tested subjects at the expense of subjects like science, social studies, arts, and physical education that are important to well-rounded child development. Teachers and parents have expressed a number of concerns about their state testing programs, such as too much time devoted to testing and the high-stakes use of testing for teacher evaluation.

We don't dispute that federal and state testing and accountability policies have been problematic or that there are ethical issues with high-stakes uses of a single measure. New assessment systems being tried in many states that are transitioning to new content standards, Common Core or otherwise, are a work in progress that need a transition period to fine-tune their designs to fully satisfy standards of technical quality. Nonetheless, annual state testing programs can play an important role in diagnosing gaps in teaching and learning that can be used to improve student outcomes. Rather than doing away with state tests, we should improve them so that they measure what matters, and that what matters gets taught in our public schools. In short, we need a better system of assessments.

RUTH CHUNG WEI (rchung@stanford.edu) is director of assessment research and development at the Stanford Center for Assessment, Learning, and Equity (SCALE) at Stanford University, Stanford, Calif. **RAYMOND L. PECHEONE** is a professor of practice at Stanford University and executive director of SCALE. **KATHERINE L. WILCZAK** is a research and policy associate at SCALE. They are coauthors of *Performance Assessment 2.0: Lessons from Large-Scale Policy and Practice* (Stanford University, 2014).

Image: Thinkstock

What is a performance assessment?

Performance assessments are tasks that ask students to produce work or demonstrate their knowledge, understandings, and skills in ways that are authentic to the discipline and/or the real world.

Performance assessments can tap into students' higher-order thinking skills to perform, create, or produce something with real-world relevance or meaning.

A more balanced system of assessments that includes varied and multiple measures of student learning would be a fairer and more valid representation of what students have learned in school.

State education agencies and assessment developers must communicate more transparently about what the new assessments measure and the benefits of more demanding assessments.

What teachers actually do day-to-day and week-to-week should matter and count as one key measure of student learning.

Involving teachers in state and local systems of assessment increases teacher buy-in and ensures that assessments are truly aligned to what teachers teach and what students learn in the classroom.

The role of performance assessment

In the 1990s, a number of states introduced performance assessments into state testing programs in an effort to improve what they knew about student learning. We define "performance assessment" as tasks that ask students to produce work or demonstrate their knowledge, understandings, and skills in ways that are authentic to the discipline and/or the real world. In the past few years, we've witnessed a renewed interest in performance assessments as a way to counterbalance the dominance of standardized multiple-choice tests. States that adopted the Common Core State Standards and the Next Generation Science Standards have realized that existing tests (with primarily multiple-choice items) are inadequate to assess the full breadth of the standards, which require students to demonstrate more complex skills and understandings.

This is where performance assessments come in. Performance assessments can tap into students' higher-order thinking skills — such as evaluating the reliability of sources of information, explaining or arguing with evidence, or modeling a real-world phenomenon — to perform, create, or produce something with real-world relevance or meaning. Researchers also have found that performance assessments can produce positive instructional changes in classrooms, increase student skill development, increase student engagement and postsecondary suc-

cess, and strengthen complex content and conceptual understandings.

For example, consider one of the English language arts performance tasks publicly released by the Smarter Balanced Assessment Consortium following its 2013-14 field test. A 3rd-grade writing performance task on the topic of astronauts assesses students' ability to research and synthesize information from two authentic sources. Students are directed to use the two sources to answer three research questions that measure their ability to:

- Identify relevant sources;
- Evaluate the usefulness of sources; and
- Integrate information from sources.

Responding to these short-answer questions prepares students for addressing the final prompt: "Using more than one source, develop a main idea about being an astronaut." Because of the open-ended nature of the prompt, there are no correct or incorrect responses. Students are scored on their ability to effectively communicate a main idea about the topic and use evidence from the sources to support and elaborate on that main idea.

English language arts performance tasks such as this one go beyond students' short-term factual recall or reading comprehension to evaluate skills that are transferable to their long-term learning and the real world, such as research and media literacy. They are also more challenging in that they require students to read, think about, and analyze sources, and to communicate their own ideas about what they have read. When performance tasks are combined with multiple-choice items and short-answer questions, large-scale assessments have the potential to measure a broader range of the standards and to assess what matters. We argue that a more balanced system of assessments that includes varied and multiple measures of student learning would be a fairer and more valid representation of what students have learned in school. They comprise a better and more rigorous system of assessments.

Learning from the past: What it takes to sustain high-quality assessment systems

In our recently published report, *Performance Assessment 2.0: Lessons from Large-scale Policy and Practice* (Wei, Pecheone, & Wilczak, 2014), we studied nine state and national assessment initiatives that began in the 1990s up to today. While some of the assessment systems were successful, in most cases, large-scale use of performance assessments was discontinued due to a variety of challenges to those systems. The main lessons from our research can be organized into three categories:

- Technical quality;
- Practical issues; and
- Political contexts (specifically the importance of leadership, communication, and public support).

Many of the technical quality issues for integrating performance assessment into large-scale assessment systems have been overcome, and states have made significant progress tackling issues of implementation, but many political, communication, and leadership challenges will continue.

Technical quality issues

Large-scale assessments used for individual and school-level accountability must meet certain technical criteria to be defensible. These criteria include:

- The reliability and comparability of the scores (in the case of performance assessment, scores produced by human raters);
- The validity of the assessments (with clear and specific learning targets being measured); and
- The comparability of performance tasks.

In the 1990s, state assessment programs using performance assessment struggled to meet these criteria.

We've come a long way in the past 15 to 20 years, and the previous technical limitations that led some policy makers to question or reject performance assessments have been largely overcome. Today, the field of assessment development has evolved to include disciplined frameworks for assessment design (e.g., Evidence Centered Design), detailed item specifications, task models or shells, quality criteria, and review processes, so that assessment systems that include performance assessment formats are valid, reliable, comparable, and unbiased/fair. For example, Advanced Placement exams, which include open-response components that must be hand-scored, are accepted by colleges and the public as reliable and credible assessments. Using performance tasks in combination with other item formats (such as multiple choice or constructed response) to measure overlapping measurement targets supports both greater content validity and reliability. Assessment developers are using these state-of-the-art practices.

Practical issues

Performance assessments cost more to develop, implement, and score, and new assessments require

time to be developed, piloted, field tested, and refined to bring them to a level of technical quality required for high-stakes use. Most important, teachers and students need time to adjust to the new standards and tests. In 2001, NCLB dramatically increased the costs of testing across states due to the requirement to test all students in more grades and report more quickly. States experimenting with performance assessment in the 1990s found that they could not afford to sustain the use of performance assessment in this context.

To get schools and teachers up to speed on the new standards and assessments of the 1990s, states should have sought to create coherent systems of assessment, instructional resources, and professional development. But in many cases, state policies and budgets did not prioritize such comprehensive approaches to instructional change. Then and now, a focus on assessment as a lever for reform has not led to widespread instructional improvement or sustained teacher and parent support.

Today, states have combined resources through several assessment consortia to take advantage of economies of scale and to share the cost of developing assessments. In addition, efforts are underway by both public and privately funded entities to build shared instruction and assessment resources (such as formative assessment libraries, interim assessments, and performance assessment task banks) that support instructional change, though more investments in teacher professional learning are still needed.

Political issues

Last, the political context matters for the sustainability of a new assessment system. Scarce resources, competing visions for the purposes of assessment, and changes in political leadership led to defunding or dismantling many of the new assessment programs of the 1990s. Lack of clear communication by state education agencies about the purposes and benefits of the new standards and assessments resulted in poor teacher and parent buy-in. Some of the assessment programs were subjected to damaging media attacks that supported vocal oppositional groups who wanted to dismantle them.

We continue to see political contexts as the biggest obstacle for including performance assessment in large-scale assessments today. While widespread adoption of the Common Core initially made the policy environment more hospitable to performance assessment, we're beginning to see significant resistance to the Common Core from the right and the left. The new standards and assessments have come under attack largely due to misunderstandings of their content and purposes and overwhelming public

We continue to see political contexts as the biggest obstacle for including performance assessment in large-scale assessments today.

opposition to test-based accountability. It is critically important that state education agencies and assessment developers communicate more transparently about what the new assessments measure and the benefits of more demanding assessments.

We are in a new era where we can do better with our large-scale assessments. What will it take?

The political climate is moving toward more local flexibility for states to determine how they'll measure student achievement and hold schools accountable. With the U.S. Department of Education's approval, most state have applied for waivers from the NCLB accountability and testing requirements. This flexibility may pave the way for including more performance assessments and greater local control of assessment and accountability systems.

New Hampshire recently received permission from the federal government to move forward with a performance assessment pilot including four districts across the state. Students will be evaluated by local and common criteria that include performance assessment, with periodic use of the state assessment at key grade levels (eliminating annual state testing at grades 3-8). Teachers are involved in creating and scoring the performance assessments, and those assessments will be reviewed externally by experts to ensure technical quality before they are administered to students. Local assessment supports the provision of more timely and useful information to teachers to inform their day-to-day instructional decisions. The pilot is still in development and both state leaders and local educators have been careful to take their time scaling up and to ensure that teachers have adequate access to professional development opportunities and instructional supports.

Call to action: What can we do now?

For states and districts looking to develop a learning-centered system of assessment, we offer the following recommendations:

State assessment and accountability systems should be based on multiple measures of student learning, including locally developed assessments. States should shift away from an assessment system that relies completely on standardized, multiple-choice tests that measure discrete skills to a system of assessment that meaningfully incorporates multiple measures, including locally developed performance assessments. What teachers actually do day-to-day and week-to-week should matter and count as one key measure of student learning. Instructionally embedded assessments — assessments that are typically completed as end-of-unit summative assessments following a series of learning activities — can potentially be one measure of student learning. For such local assessments to become a viable and trustworthy component of a multiple-measures assessment system, they require well-designed systems to support technical quality, including design tools — design frameworks, task templates or shells, common rubrics, task specifications, task quality criteria — and an effective system of peer review for validation. Such systems of local assessment could augment or replace interim assessments focusing on test prep with richer local assessments adapted to the local curriculum and student needs. Establishing a system of assessments incorporating locally developed, instructionally embedded performance tasks could be one answer to the growing opposition to external state assessments that usurp classroom time, raise parent and student anxiety, and provide an impoverished and incomplete picture of student learning. This hybrid approach that includes local and large-scale assessments is currently being pursued in New Hampshire's performance assessment pilot described above.

We are in a new era where we can do better with our large-scale assessments. What will it take?

Assessment systems should be coherent. If we expect better assessments to drive richer and deeper learning, we need more coherent systems of assessment, curriculum, instruction, and professional learning. To establish system coherence, states must invest in the local capacity of teachers to be integral players in developing and implementing a system of assessments. We need to move away from a one-size-fits-all approach to professional development to a customizable system that supports deeper learning. Teacher professional development should be grounded in authentic work and leadership experience at the local level that privileges practitioner knowledge and supports reciprocal, generative learning. Involving teachers in state and local systems of assessment increases teacher buy-in and ensures that assessments are truly aligned to what teachers teach and what students learn in the classroom. Teachers and students also benefit when we support teachers in implementing instructional strategies that give students opportunities to

learn the transferable college and career skills and understandings assessed by richer assessments.

Systems of assessment should support shared accountability and whole-system improvement. We acknowledge that improving state assessments alone will do little to improve teaching and learning if the accountability systems in place remain unchanged. Not only should states move away from tests that are designed to measure a narrow set of knowledge and skills, they also should move toward accountability policies that support learning and foster continuous improvement at all levels of the system. One alternative to current top-down accountability policies is reciprocal accountability. Accountability systems should not only raise expectations for learning, they should include strategies that support instructional change in ways that ensure that all teachers and students have the opportunity to be successful. In reciprocal accountability, all levels of the system — state, local, school, teacher, and student — are responsible for and must be actively engaged in building the capacity of educational systems to be responsive to the learning needs of all students. The system is anchored in a cycle of continuous improvement that specifies desired outcomes (standards) to measure success and identifies areas for growth. This more balanced system of accountability steers states away from punitive measures and toward support for improvement at the school and local level and helps ensure that all students are provided an equitable opportunity to learn.

Conclusion

Our country already has entered a period of tremendous education policy flux with the reauthorization of the Elementary and Secondary Education Act up in the air, anticipated leadership changes in 2016, and many states transitioning to new standards and assessment systems. This period of transition gives states an opportunity to make bold changes in their assessment and accountability systems. We know that testing and accountability aren't going away. But clearly, parents, teachers, and other stakeholders are telling us we need a change. We can build better assessments and systems of accountability that are designed to support teaching and learning, while also doing a much better job of measuring what really matters.

Reference

Wei, R.C., Pecheone, R.L., & Wilczak, K.L. (2014). *Performance assessment 2.0: Lessons from large-scale policy and practice.* Stanford, CA: Stanford University.