How can state assessments better test deeper learning?

Three models that can work

Stuart Kahl | November 2017

People expect too much of traditional state tests—they're not designed to serve the needs of policy makers as well as the immediate needs of teachers. Furthermore, these traditional assessments don't do justice to the assessment of deeper learning called for in college and career readiness standards. But if we have reasonable expectations of state assessments, and we think creatively, we can indeed find ways to deliver value to multiple stakeholders and assess deeper learning. Read on to see three potential test designs that accomplish this.

What should we expect?

There would be a lot less frustration with state tests if educators and the general public recognized what a state's summative test can and cannot do well. An end-of-year general achievement measure **can** provide excellent information for use in the evaluation of an instructional program, but **cannot** provide diagnostic information on individual students for immediate formative use by teachers. Formative practices gather evidence of students' understanding of specific learning targets that are the focus of current instruction, but a test addressing a sampling of a full year's curricular content cannot be expected to adequately diagnose gaps in a student's learning relative to any particular target or targets. Besides, it would indeed be unfortunate if a teacher needed the state's annual test to identify a student's specific academic needs.

Rather than providing direction to teachers on individual students' instruction, the results of a state accountability assessment should lead to more general questions about program strengths and weaknesses. Typical questions might be "Why are we performing worse in geometry than in the algebra section of the math test?" and "Why did this subgroup of students perform worse than the rest of our students in reading comprehension?" Investigating to find the answers to such questions is the next step on the path to program improvement. With the recognition that an end-of-year state test cannot be all things for all stakeholders, it can make its contribution to a district's balanced assessment system more efficiently. And the district can then rely on other components of the system to do what they do best.

> "Performance tasks that are part of regular instructional units could yield student work that would be evaluated and counted for state summative purposes."

Assessment of deeper learning

One criticism leveled at state summative assessment programs is that they don't measure the right stuff. That's only a partial truth. State programs call for tests that address the states' curriculum standards—college and career readiness standards that states have adopted and on which instruction is to be focused. State tests that emphasize efficiency do indeed cover important foundational knowledge and skills embodied by the standards. A fair criticism of these instruments, however, is that they do not adequately assess deeper learning—the use of higher order thinking skills in the **application** of the foundational knowledge and skills.

Even the performance tasks included in some on-demand accountability tests do not do justice to deeper learning or to the intent of performance assessment.

> "Adding the performance component to the accountability assessment system allows significant shortening of the end-of-year tests."

Some educators have been considering counting scores from interim assessments toward accountability results. The Every Student Succeeds Act (ESSA) approves of this. However, multiple interim or benchmark tests from the state (on-demand and secure) carry the same disadvantages as the end-of-year state tests and would cause greater disruption to school routines and to curricular scopes and sequences. Of course, districts do make good use of interim or benchmark assessments—of their own or from external resources—to meet their own needs. These tests are important parts of a balanced system.

From a program evaluation perspective, I value measures of students' capabilities at the end of a course of study more than I value measures of short-term memory that are administered immediately after instruction in curricular topics. I believe a much more valuable component of accountability assessment to be conducted on an interim basis would be curriculum-embedded performance assessment (CEPA).

Performance tasks that are part of regular instructional units could yield student work that would be evaluated and counted for state summative purposes. Local teacher scoring would yield the kind of immediate, usable results that teachers need—and that have been unreasonably expected from end-of-year general achievement measures. (For more on CEPA, see Re-balancing Assessment: Placing Formative and Performance Assessment at the Heart of Learning and Accountability (Hofman, Goodwin, Kahl, 2015).

Three models for state accountability assessments using matrix sampling

To meet appropriate expectations of different assessment types and to assess the deeper learning outlined in college and career readiness standards, I present three models for statewide assessments. All three of the models involve matrix sampling— a testing technique by which a whole test is broken down into small non-equivalent subsets of items, with each student taking only one of the small subsets. As demonstrated by successful large-scale programs in the past, the results on the item subsets can be aggregated to produce very reliable group results because the aggregate test provides excellent coverage of the target subject area domain.

All three models include a variety of item types: multiple-choice (MC), constructed-response (CR), and technologically-enhanced items (TEIs) delivered online. Where the designs differ is in their use of common and/or matrix-sampled items and in the use of data from interim curriculum-embedded performance tasks. The high-level differences are shown in Table 1. Details follow.

Table 1. Differences among the models

	Model 1	Model 2	Model 3
Common and/or matrix-sampled items?	Both	Both	Matrix-sampled only
Interim performance component?	No	Yes	Yes

Model 1: Common and matrix-sampled items, no interim performance component

Model 1, depicted in the table below, combines common items and matrix-sampled items. All students take the common items. With MC items and TEIs each counting one point and CR items worth up to four points each, one can see that the total common test is worth 56 points, sufficient to produce reliable student scores. It is important that constructed-response questions count significantly in the state test (in this case, three-sevenths of the total common test). Sole reliance on selected-response questions in a high-stakes accountability environment has been shown to have a negative impact on instruction and local testing.

Table 2. Model 1: Summative Test Only

End-of-Year Summative Test				
	Common Items		Matrix-Sampled Items (8 forms)	
Item types	MC and TEI	CR	MC and TEI	CR
No. of items	32	6	8	2
Total test time	120 minutes			

MC multiple-choice; CR: 4-point constructed response; TEI: technology-enhanced item

For a program without a separate, curriculum-embedded performance component, rich constructed-response questions are doubly important. A lot can be done with this format, even in an on-demand, secure testing situation. For example, a researchable problem can be described to students who are then asked to design an experiment addressing the problem. Or results on pentathlon events (some in units of time and some in units of distance) can be provided to students who must devise a fair way of combining the results of the different events to determine the winner.

With the exception of the scoring of writing samples for which automated scoring can play a role, I believe human scoring is the method of choice for evaluating the responses to these constructed-response questions. The planned use of computer scoring of constructed-response items in the academic disciplines has affected the nature of the items people develop. Designing questions that permit the computer to “do its thing”—counting, sorting, matching, etc.—lowers the cognitive demands on the students.

The aggregated matrix-sampled items can be thought of as another test, which has multiple purposes: the field testing of new items to be used in future years, use in the statistical equating of tests across years, and expanded coverage of a subject domain providing far better program diagnostic information (average subtest scores) than can

be provided by the common items alone. This approach to field testing items is the best possible. Students aren't aware of the distinction between common and matrix-sampled items—they experience the same testing conditions and have the same level of motivation in responding to both sets of items.

Measured Progress didn't invent matrix sampling or the idea of embedding field-test items in operational tests. However, in the early years of statewide testing programs (then as Advanced Systems), the company introduced the Model 1 design to many states across the country. The approach remains as beneficial today as ever in this age of computer-based and standards-based testing.

Model 2: Common and matrix-sampled items and an interim performance component

This model, depicted in the table below, reduces the end-of-year testing time from the 120 minutes in Model 1 to 90 minutes. The roles of common and matrix-sampled items remain the same as in the previous model. The reduced summative testing time (and related point total) is enabled because of the addition of the performance component. Three curriculum-embedded performance assessments (CEPAs) could easily generate as many or more total points as the common item test, and it is the intent in this design that the results of those two components be combined to produce reliable student scores on a broader measure of performance, addressing both foundational knowledge/skills and deeper learning.

Table 3. Model 2: Summative Test plus Performance Component

	End-of-Year Summative Test				Performance Component
	Common Items		Matrix-Sampled Items (6 forms)		
Item types	MC and TEI	CR	MC and TEI	CR	Curriculum-Embedded Performance Assessments
No. of items	22	4	8	2	3
Total test time	90 minutes				Multi-day instructional units with summative tasks

The CEPAs are multi-day instructional units involving multiple activities, some yielding student work for formative use, and some yielding student work for summative use-all evaluated or scored by teachers for immediate results. To incorporate the summative scores into the state accountability results, states would employ a score audit procedure. To assure high task quality and validity, the CEPAs should be put through the same review and field-testing procedures as the items for the state end-of-year tests. Multiple summative tasks in any one CEPA could yield a total of 20 to 30 independent points. However, score points could be collapsed in such a way that a smaller score range could have similar score distributions across CEPAs. In this way, CEPAs could be considered of equal "difficulty" just like pre-equated writing prompts historically.

In the very beginning of a program involving CEPAs, common ones may be used. But as more pre-equated CEPAs are developed over time, teachers could choose the three they want to use, fitting them into their instructional sequences as they desire. Security of instructional units should not be a concern. When engaging students in the activities generating student work for summative purposes, teachers would be expected to follow administration instructions just as they do for traditional summative tests. The score audit procedure for accountability results could involve the electronic submission of student work from a small sample of students in each school.

Model 3: Totally matrix-sampled summative test and an interim performance component

This model pushes the innovation envelope. ESSA and its predecessors NCLB and IASA require the state accountability assessment programs to "produce individual student interpretive, descriptive, and diagnostic reports…" allowing different audiences to address specific academic needs of students…provided "as soon as practicable after the assessment is given." As explained in the opening paragraphs of this paper, existing state assessment systems, even though they may pass muster with peer review committees, generally do not and cannot satisfy this requirement.

That being the case, the end-of-year summative test in Model 3 is totally matrix-sampled, yielding no reportable student scores since students take very short, non-common tests. The ESSA's student-level reporting requirement is best met by the curriculum-embedded performance assessment component in the model. I don't know whether the U.S. Department of Education will allow its peer reviewers to accept this approach. However, given that the reports the USDOE has accepted already have not been interpretive, descriptive, or diagnostic and that the local reports of student performance on CEPAs can be, I believe the approach should be approved.

Table 4. Model 3: Matrix-Only Summative Test plus Performance Component

	End-of-Year Summative Test		Performance Component
	Matrix-Sampled Items (8 forms)		
Item types	MC and TEI	CR	Curriculum-Embedded Performance Assessments
No. of items	16	2	3
Total test time	36 minutes		Multi-day instructional units with summative tasks

Although student-level scores on matrix-sampled item subsets cannot be reported, they can be computed and converted to performance levels so that percentages of students at different performance levels can be determined and reported for schools, as required by the federal law. The measurement error associated with student-level results on matrix-sampled tests is random, allowing group-level—i.e., school-level—results to be highly reliable. Of course, combining the individual student results on the end-of-year and CEPA components further enhances the quality of the accountability data.

This approach to the end-of-year component is not new or untried. Prior to implementing the Massachusetts Comprehensive Assessment System (MCAS), which uses Model 1, the Commonwealth of Massachusetts operated a totally matrix-sampled program called the Massachusetts Educational Assessment Program (MEAP) for over a decade. All students at three grade levels in the state participated in the testing for that program, which in its later years implemented performance-level reporting. The more reliable school-level results associated with matrix-sampling of many more test items than can be included in common tests led to more accurate indicators of school performance and identification of low-performing schools. MEAP tested students in four subjects in just two hours. Years earlier, the California Assessment Program used many more matrix-sampled forms than Massachusetts, such that each student's test was contained on one sheet of paper. Notice the summative testing time in Model 3 is 36 minutes per subject. Adding the performance component to the accountability assessment system allows significant shortening of the end-of-year tests.

Performance assessment via CEPAs need not be costly to the state or burdensome to teachers. CEPAs are reusable instructional units that teachers would use in place of other units they might have used in the past. Thus, Model 3 reduces the burden of a state assessment on the schools and school personnel considerably. Also, the reusable CEPAs and teacher scoring (with some auditing) can result in significant cost savings for the state.

One can think of Models 1 and 2 described earlier as steps in a three-year phase-in of Model 3. The result would be an accountability assessment program that is efficient in several ways, assesses deeper learning, involves teachers meaningfully in the process, and can positively impact instructional practices. Thus, the third model addresses the major criticisms of traditional programs.

However, the transformation of state assessments could go even farther, if desired. The end-of-year summative component could just be administered to a sample of students in each school and used for verifying the results from the curriculum-embedded performance component. Or, given demonstrated success of performance assessment scoring and auditing procedures, the end-of-year summative test could be eliminated altogether. The federal requirement for individual student reports would still be satisfied by the CEPA component as in Model 3.

Radical? Perhaps. But right.

States have been slow to pick up on ESSA's flexibility and acceptance of innovation. Focusing statewide summative testing on what it can do well and not trying to meet outsized expectations should open the door to innovative approaches. Perhaps soon we will see some bold state departments of education implement new solutions that are better for states, better for schools, and most importantly, better for students.