Assessment Matters:

Constructing Model State Systems to Replace Testing Overkill

FairTest
National Center for Fair & Open Testing

Assessment Matters:

Constructing Model State Systems to Replace Testing Overkill

By Monty Neill

Executive Director, FairTest

A Report by the National Center for Fair & Open Testing

October 2016

Thank You

FairTest expresses our deep appreciation to people who offered their valuable time for interviews and discussions as we prepared this report. They have all contributed greatly to our understanding.

Avram Barlowe, Joe Battaglia, Richard Chang, Sarah Chang, Ann Cook, Shawna Coppola, Kristina Danahy, Kathy D'Andrea, Robin Coyne, Dan French, Ayla Gavins, Matthew Glanville, Paul Leather, Kate Lucas, Dennis Littky, Deborah Meier, Mission Hill School teachers, Rob Riordan, Rollinsford Grade School teachers, Lynne Stewart, Hanna Vaandering, and Elliot Washor.

FairTest also thanks our funders who supported this project: Bay and Paul Foundations, Open Society Foundations, New World Foundation, Schott Foundation, National Education Association and numerous state and local affiliates, and many generous individuals.

FairTest staff who contributed to this report: Lisa Guisbond, David Mirabella, Bob Schaeffer.

Cover photograph credit: New York Performance Standards Consortium students at work. Photo by Roy Reid.

FairTest
P.O. Box 300204
Boston, MA 02130
www.fairtest.org
fairtest@fairtest.org
617-477-9792

Assessment Matters:

Constructing Model State Systems to Replace Testing Overkill

Table of Contents

Note: The full report that this section is excerpted from is available online at:
http://www.fairtest.org/assessment-matters-constructing-model-state-system

Assessment Matters:

Constructing Model State Systems to Replace Testing Overkill

Part I

A Model Assessment System for High-Quality Learning: Local Assessments in a Statewide System

The "Innovative Assessment Demonstration Authority" pilot program in the federal Every Student Achieves Act (ESSA) allows up to seven states to implement new state assessment systems. These will be phased in over time to replace existing standardized tests. This initiative could lead states to fundamentally improve student assessment.

To help states and education reformers take advantage of this opportunity, in this report FairTest proposes a model system to maximize high-quality assessment within ESSA's constraints. The model represents a significant departure from the narrow test-and-punish framework of No Child Left Behind (NCLB), which ESSA replaces. Unlike NCLB, which revolved around standardized test scores, the model begins with classroom-based evidence from ongoing student work. FairTest's model is rooted in exemplary practice and a set of principles derived from decades of assessment reform efforts (see Part IV).

The primary purpose of this innovative system is to support high-quality, individualized student learning. It is guided by teachers but substantially student controlled, thereby encouraging pupils to build on their interests, with multiple ways to demonstrate learning. It also provides the basis for making decisions about how best to improve student outcomes, teaching and schools.

In FairTest's model, states design a "system of systems." In it, districts, or consortia of schools or districts have the flexibility to vary the structure and nature of their local assessment plans to address their particular needs and challenges. This could range from assessments rooted in inquiry- and project-based learning, with extensive student choice, to more traditional curriculum, instruction and tests.

To fulfill ESSA's public reporting and accountability requirements, the model system relies primarily on classroom-based evidence. Teachers and their students gather evidence of learning throughout the school year, including from any major projects. Teachers prepare a summative evaluation of each pupil that includes a determination of the student's level of proficiency in line with state standards, as required by federal law. This data can be aggregated and then broken out by demographic groups to shed light on the success or failure of efforts to close gaps in achievement.

To establish "comparability" across schools and districts, the state employs a set of procedures to ensure that a student deemed proficient in one district would be deemed similarly proficient in another with a different local assessment system. Typically, this involves using state standards as the basis for independently re-scoring samples of classroom-based work. This, in turn, provides the information needed for public reporting and accountability.

FairTest's model is intended to help states design a locally empowering, flexible system that provides accountability while ensuring that accountability structures do not undermine rich, deep teaching and learning. ESSA's requirements can create difficulties in implementing quality assessment for learning. However, the space for progress is large enough to make ESSA's innovation pilot an important step forward, if used well.

The Core of a Model System: Classroom-based Evidence

Classroom-based evidence can include student work gathered and evaluated in portfolios, learning records, work samples, some of which include teacher observations. It can include performance tasks produced as part of ongoing academic activities. It also can incorporate student work done out of school, such as internships, and can include group projects. What differentiates this model from similar proposals that focus on performance assessments is its use of practitioner-designed and student-focused assessments that emerge from ongoing schoolwork. Practitioner-designed means that teachers, individually and collaboratively, create assessments that grow out of the specific curriculum in the classroom or school. Student-focused means they have significant choice, with teacher guidance, of content, such as the specific science or history investigation, or in the mode of presentation, such as an oral report, a written report, a video or a computer game. Allowing student control has been shown to improve student learning (Coleman, 1966).

New York Performance Standards Consortium students. Photo bv Rov Reid.

Performance tasks take various forms, from short pieces of work to extended or group projects. Performance tasks may be completed frequently during the year as part of the regular curriculum, be culminating tasks (e.g., senior projects), or be externally required tests. For example, to graduate from high school, students in *New York Performance Standards Consortium* schools must complete four extended performance-based assessment tasks

developed in collaboration with their teachers. Other nations, such as Australia, use performance tasks as key components of their systems (Darling-Hammond, 2014).

Portfolios are ongoing collections of student work. (ESSA does not list portfolios as an explicit option for states, but they fit within the options listed.) The value of portfolios is that they reflect the curriculum (learning opportunities) and the quality of student work. With guidance and strong scoring procedures, they can give a more accurate and multifaceted indication of learning than standardized test scores. Portfolios can incorporate a wide range of work, from short quizzes to longer tests, lab reports to extended research and performance tasks.

The *Learning Record* (LR) is a precisely constructed tool for gathering and summarizing evidence of student learning over time. Evidence shows it is a rich, valid means of documenting progress. Independently re-scoring samples from classrooms has shown that teachers can evaluate their students reliably. The *Work Sampling System* also provides means for gathering and summarizing evidence; it is used in younger grades and includes non-academic components. (Both are described in Part III.)

In FairTest's model system, students exert significant control over assessment content. They can select books to read, science experiments to conduct, social studies investigations, and extended math problems. This applies to portfolios and performance tasks. Thus, work varies across individuals, and in contrast to computer-adaptive, standardized assessments, supports *authentic* personalized learning. Significant student control does not preclude teacher assignments or use of common readings/materials or tasks. Indeed, a hallmark of this model is practitioner control. Teachers have responsibility for their curriculum, their instructional practices, and their use of assessment. They guide student choice.

Classroom-based assessments differ from performance tests. Classroom-based assessments that emanate from student ongoing work in the curriculum differ from performance tests. The latter are tasks generally designed from outside the classroom (though often by teachers) and administered as summary tests or during the course of the year. To take advantage of student interests and help them learn to control their own ongoing learning, the former comprise the core of FairTest's model system. However, performance testing can be a major improvement over current standardized testing and perhaps a bridge to increased use of classroom-based assessing in local and state systems.

One pilot program is already in effect. New Hampshire sought to move away from standardized tests. It won a waiver from NCLB to pilot a new state assessment system – the Performance Assessment for Competency Education (PACE) – which combines statewide and local assessments. In the New Hampshire system, teachers design common performance tasks to be used across participating PACE districts and ultimately the state. They also design local tasks that are administered when they best fit into the curriculum. (They are therefore a form of performance tests.)

One high school geometry common task is "Water Tower," in which "students are asked to design a tower that will hold approximately 45,000 cubic feet of water, with special attention to using the least amount of construction materials. Student work is scored at four levels of mastery and three areas: models and scale drawings; calculations and mathematical strategy; and communication of the analysis" (Richmond, 2016). Each student gets the same task, and local teachers score it using a task-specific rubric.

Students reportedly found the PACE task engaging. Students did have to consider options, there was not just one right answer, and they participated in a form of "real world" problem solving using geometry. But it is assigned as a form of test rather than being a project or demonstration of learning within the curriculum. (For more on PACE, see Part II.)

In contrast to PACE's performance tests, the sorts of tasks required in the Rollinsford (NH) Grade School (RGS) and other places evaluate students on tasks or projects that emerge from the curriculum and are also learning experiences. At Rollinsford, students from Kindergarten on engage in extensive student-selected project/inquiry work in various subjects.

Students at a woodworking shop. Photo from Big Picture Learning.

For example, several fifth and sixth graders decided to investigate river dolphins. Their display project, shared first with the rest of the school and then at a public open house, included biology and environmental science, writing up the results, assembling a graphic display, and selling tie-dye T-shirts to raise funds to support preservation efforts. They discussed their work, the choices they made, and their findings. This investigation emerged out of their classroom activity. They were in charge of the project, guided by their teacher. The results could have been scored according to state-defined achievement levels (based on SBAC), but the school's goal was for every student to share completed or ongoing work *they* wanted to talk about.

The Rollinsford student work is an extended project involving research rather than a task that is likely to take up no more than one or two class periods and that is based entirely on tapping students' existing knowledge, albeit to solve a realistic problem. (For more on Rollinsford, see Part III.)

While the Rollinsford approach should be the foundation of a new system because it builds on classroom work, the use of teacher-made tasks as the core of the New Hampshire system is a

significant step forward. Designing tasks can provide strong opportunities for teacher collaboration and learning. Teachers are free to use additional performance tasks, including student-initiated ones, in their classrooms. Indeed, the core of PACE is that each teacher determines her/his students' level of proficiency based on the student's work over the year (see Part II). The knowledge and cooperative practice teachers develop can provide the basis for moving toward classroom-based assessing as the foundation of a state system, as ESSA allows.

Caution: Computer-based testing. ESSA allows states to build systems in which students are assessed multiple times per year so that each student gets an aggregated score at the end of the year that establishes his or her proficiency level. This could mean portfolios. Or it could mean repeated multiple-choice/short-answer tests that are part of computerized instructional packages. Far from a valuable innovation, this would further reduce teaching and learning to the regurgitation of facts and procedures and thereby block avenues for deeper learning.

For example, various corporations are marketing online curricula that test students frequently, such as when they finish a curriculum unit. These are at times described as "individualized" or "personalized," though those terms simply mean that students proceed through the computerized curriculum at their own pace, or that a computer algorithm determines the next step for each student.

Proponents argue these continuous tests provide more information to teachers and are fairer than one big end-of-year exam. However, they are mostly multiple-choice and short-answer, with some writing samples often scored by computer, same as current standardized tests. They reduce instruction to what can be measured by these kinds of items. In addition, because they are integrated with curriculum, it is more difficult for parents and students to refuse to take them.

ESSA Innovative Assessment and Accountability Requirements

Pilot state programs will have to meet ESSA's general mandates for state assessments as well as specific criteria. The overall mandates include a requirement to sort students into at least four proficiency levels. A state can introduce new features to its current standardized exams (such as performance tasks) without using the innovative assessment pilot, provided the new elements are administered to all students in a grade. However, a state needs U.S. Department of Education (DoE) approval to build a new system up from pilot districts if, during construction, not all children in the state participate in the new assessments.

A state could pick one or more subjects and one or more grades to start its pilot. Indeed, it could decide to implement a new system only in one subject or one grade level, such as elementary, leaving all else measured by statewide standardized tests.

The DoE will study the first three years of each state project. At that point, it could continue, end or expand the program. States will have five years to build their pilots up to statewide, but extensions are possible.

Within the new system, assessments can vary across districts – provided the results can be shown to be comparable. Civil rights groups and others have insisted on comparability to provide evidence that expectations and learning outcomes are similar across diverse students and districts and to provide tools for addressing inequities. Neither the law (2015) nor draft regulations (2016b) specify how that is to be done in a completed new system.

During the development process, however, results of the new assessments must enable comparability with the state's current system. Here, the draft DoE regulations (2016b) are specific. They give pilot states the option to:

- Administer the state test at least once each in elementary, middle and high school, and give the new assessments in at least the other ESSA assessment grades.
- Administer both the state exam and the new assessments to a demographically representative sample of students in the pilot program, at least once each in elementary, middle and high school.
- Include common items in both the pilot and state tests.
- Or propose an alternative method for demonstrating comparability. The state must show how the method will provide for an equally rigorous and statistically valid comparison between student performance on the innovative assessment and the existing statewide tests.

These tools can also help establish the validity and reliability of the new assessments, another ESSA requirement and a basis for determining comparability. However, the requirement to compare new assessments with old tests risks limiting the new assessments to what the old tests measure. Thus, the ability of students to engage in extended investigations, produce rich work samples, apply deep knowledge to real-world situations, and take charge of their own learning could be ignored in favor of superficial tasks that correlate more closely with the rote and procedural knowledge covered by current tests.

Comparability within a New Assessment System

States participating in the pilot will have to choose tools to compare the new assessments to existing tests. They can use similar tools to compare results from different local assessments. Once the new system is built, the old tests will no longer be needed. States could then streamline procedures for determining comparability.

In order to establish comparability among students participating in the innovative assessment or in a completed new system, there are several options. Each has benefits and drawbacks.

Re-scoring

In re-scoring, also termed "moderation," all or (usually) samples of completed work (portfolios, projects) are re-scored, usually by other teachers. This is done to ensure consistency of grading across teachers, schools or districts. If the results are consistent, then "proficient" in one district likely means "proficient" in another. Moderation requires the use of common scoring guides ("rubrics") and samples of student work that exemplify student work at various proficiency levels ("exemplars").

Establishing comparability by re-scoring classroom-based evidence has been done in the U.S. and internationally. It is part of the toolkit for New Hampshire's PACE program. For the *Learning Record*, comparability rests first on the carefully constructed guide for gathering evidence, then on its developmental reading and writing scales. These describe what students know and can do at various stages. Three samples from each classroom are re-scored by other teachers in a system-wide moderation session. Agreement between re-scores and the originating teacher's score tends to be strong. (For more, see Part III.) Originating teachers quickly improve consistency in how they place students on the scales and in the selection of evidence of learning to back up the placement.

In the NY Consortium, comparability is addressed with guidelines for students and teachers to use in developing the graduation tasks and a scoring guide used across schools. Samples are annually re-scored by new teachers to see if originating teachers are applying them with sufficient consistency.

What if re-scoring detects significant scoring discrepancies? New Hampshire says, "Discrepancies between local and state/consortium assessment results do not mean that the local results are wrong. Rather, it should lead to conversations and inquiries to try to understand the reason for any large differences between the two sets of results" (NH DoE, 2014). In any event, in its first year, independent researchers found no major discrepancies between originating districts and the moderated results (Evans, Lyon and Marion, 2016).

New York Performance Standards Consortium students. Photo by Roy Reid.

The main disadvantage of statewide scoring guides is the risk of lowest-common-denominator rubrics that limit the ways students can demonstrate their understanding. Even good rubrics can be problematic in judging creative work, as Chris Gallagher discusses in his book on Nebraska's innovative assessments of the 1990s (2007, pp. 69-71). State scoring guides used across all work in a given subject could enforce a form of back-door standardization, as tests requiring writing in response to a prompt often do. An

example is the infamous “five paragraph essay” on which teachers drill students in order to produce a response that will get a good score. This often leads to bad writing and, even worse, reduces interest in writing.

On the other hand, high-quality rubrics, combined with exemplars, can focus attention on the most important characteristics of much student learning. The NY Consortium (N.D.) provides a good example. At a minimum, teachers should review scoring guides every two to three years to improve them and select new exemplars if needed.

Anchor tasks and tests

ESSA draft regulations propose anchor or common tasks as the principal means for ensuring comparability between new assessments and old tests. They can also be used to establish comparability across districts. They are a reasonable procedure.

Essentially, the idea is that all participating pilot districts administer the state tests in a few grades, or perhaps only to samples of students in those grades. They also administer common tasks across the districts in grades that do not take the state test, or in all ESSA-required grades (e.g., 3-8). Each district scores its common tasks, as NH PACE does. In that system, districts also design their own local assessment systems that employ teacher-made tasks modeled on the common tasks. Under ESSA, other forms of local assessments could be used. Samples of anchor tasks are independently re-scored to determine whether the districts are scoring them consistently. The common task or state test results can then be compared with local assessment results.

New Hampshire compares anchor tasks with state tests, and local assessment results with the tests and tasks. The central comparability tool is for evaluators to compare by district the results on common tasks with teachers’ holistic “competency determination” of each student’s level of proficiency in each subject. The determination incorporates results from the local assessment tasks but also includes evidence of student learning over the whole year. In general, results in all the various comparability procedures have been reasonably consistent. (See Part II for additional detail.)

While it is time-intensive to produce tasks, and re-scoring adds more time, it is far less expensive than creating a complete set of statewide tasks for each subject plus conducting a statewide scoring process for each of them. Done well – based on shared standards and made by teachers who will use them in their classes – anchor tasks should fit cleanly into the actual curriculum in many schools. Writing and scoring them can provide important learning opportunities for teachers. Still, use of anchor tasks creates complications that could undermine the instructional value of performance assessing.

The main disadvantage is that these tasks do not emerge from student interests within the curriculum. Thus, they may or may not engage students, may or may not connect well to the curriculum. Reviews of performance tasks generally report greater student engagement than

with standardized tests, but student ownership, as in the NY Consortium, can provide deeper levels of interest and enhance students' sense of control over their learning.

Even when teachers collaborate to design tasks, there will be less immediate teacher connection to the tasks by teachers not involved in the design, and thus potentially more distance from a teacher's curriculum. In the end, some students may have studied more closely than others the particular topic covered by the state task. As a result, higher scores could be based on that accident.

Pre-set tasks administered as tests are not strong tools for helping students acquire new knowledge, even if they provide good opportunities to solve problems and apply knowledge. In itself, this is not a major concern, especially if there are only a few common tasks. But the model is lacking when compared with the learning potential of deep investigations.

Validation studies

Another approach to comparability is a "validation study." The idea is to analyze performance assessment results in participating districts to determine if they are comparable to the state's standards-based definition of each academic level (e.g., "proficient"). This relies directly on the standards rather than the state tests. During the process of developing a system under ESSA, a state also has to compare local results with the state exam. Once the system is complete, the old state tests would no longer be needed. At that point, a state could compare results from local systems using standards-based descriptions and exemplars, rather than use a state test or anchor tasks.

Fully developed, if a study of a district shows strong comparability by "express(ing) student results or student competencies in terms consistent with the State's aligned academic achievement standards," as ESSA requires of all state assessments, then no more evidence about the district would be needed until a periodic follow-up, for example, after three years.

Addressing Contradictions in Building a New System

ESSA requires a new assessment system to show its results are comparable with a state's existing standardized tests. Both are supposed to be based on state standards. However, they are likely to measure significantly different knowledge and skills.

For example, both SBAC and PARCC "Common Core" tests include a few fairly short performance tasks and some short-answer ("constructed-response") questions. The tests mostly rely on short items, which include multiple-choice as well as computer-based questions, such as "drop and drag" responses. The advantages are that a state can purchase the test inexpensively and it does incorporate a few performance tasks.

The disadvantages include excessive length, too few and too limited performance tasks, and mostly non-performance components. They include no extended projects. Thus, they are unable to assess student ability to engage in research or any work carried out over more than one or two class periods, produce substantial papers or in-depth products, or to take charge of their own learning. They preclude students from demonstrating their learning by using modern technology, from blogs to videos, graphics and computer games. Thus, they are not useful models for designing local systems or for professional development. And they could exercise too much influence on curriculum and instruction.

If the goal is to ensure comparability across a state based on intellectually substantive standards, using tests focused on retention of facts and basic skills to judge performance assessments can be misleading. The latter measure different and more valuable aspects of knowledge and skills and differ in format. They should not be expected to be closely comparable.

Teachers also may confront the problem of serving two masters: old tests and new performance assessments. They could face pressure to establish consistency between high-quality classroom evidence and low-quality tests, thereby distorting how they design the new assessments and how they evaluate student results.

Performance assessments can improve teaching and learning by engaging students more deeply in their coursework and enabling them to strengthen their knowledge and skills through extended, in-depth projects. As children in disadvantaged communities have suffered the most from teaching to standardized tests, the benefits provided through performance assessments may be especially valuable. Students could become more engaged and learn the kinds of knowledge and skills assessed by performance tasks but not by standardized tests. If that happens, performance task results may diverge from test scores. Deborah Meier, founder of the Central Park East Secondary School, which focused on performance assessment, said their graduates did well in college, but their test results rose only modestly. This is also the case with the NY Performance Standards Consortium. The process of judging performance assessments and their results by standardized tests could lead to dismissing real gains in learning that are not measured by the tests.

Using just the state standards would be significantly better, though they are often developmentally inappropriate or have questionable emphases. However, ESSA requires establishing comparability with existing tests during the period in which the state is creating the new system. States will have to carefully consider how to address this problem. Potential problems could be minimized if districts are not forced to alter their performance assessment scores to produce correspondence with old tests. But so long as state tests are falsely presented as the gold standard, problems will remain.

Finally, another danger from comparability requirements is to the assessments themselves. Using the *Learning Record*, NY Consortium teacher-directed assessments, the Work Sampling

System, or other systems that allow fully individualized content for accountability purposes has not been widely established in practice. The primary danger is not lack of validity, reliability or comparability. It is that the assessments will be corrupted by attaching high-stakes, punitive consequences.

Moderation procedures in the U.S. and other nations have ensured teacher accuracy and fairness. However, in most of these cases accountability pressures on schools and teachers have been low, even if sometimes high for students. One exception is the NY Consortium, in which the performance task results are included in state accountability as well as required for graduation. When states move in the direction of teacher-controlled, student-focused assessments, accountability pressure is a danger that must be monitored. It would be a great loss if high-quality assessments were undermined by accountability requirements.

ESSA allows states to focus on assistance, not punishment, which enhances the opportunity to use high-quality assessments. For this and other reasons, states must change their accountability systems.

Accountability and improvement

The goal of FairTest's model is to improve teaching, learning and school quality through the use of performance assessment. There are other tools to consider, including these two:

ESSA requires each state to include at least one "school quality or student success" indicator in its accountability mix. Examples can include school climate surveys, disciplinary data, and more. Indiana listed dozens of possibilities (Chalkbeat, 2016). The National Education Association (2016) called on states to establish "dashboards" with various forms of school data. California's Community-based Accountability requires districts to include evidence from eight areas (CSBA, 2013). The purpose is for local systems to gather a rich array of information about school quality and student progress for use in reporting and in improving school practices. These are significant but limited steps forward as they widen the scope of attention from just test scores but do not sufficiently end the reign of standardized tests (Cody, 2016).

The Darkroom Photography class gets ready to process film for the first time. Photo by Roy Reid

Unfortunately, the US DoE (2016a) has drafted regulations that limit the value of this option: They say the other measure(s) must predict academic outcomes (which the law does not require) and the academic measures must constitute the "great majority" of the weight given the various indicators. Thus components that are valuable in themselves, such as a positive school climate, could only be used if a state could show that a better climate predicts better test

scores. States that choose to continue to focus on exams will minimize the weight given other indicators.

As states think about overhauling accountability and improving schools, they could consider school quality reviews (SQR), modeled on the British school inspectorate (Rothstein, Jacobsen & Wilder, 2008, Ch. 7). Under this approach, teams of experts periodically review schools to provide them with feedback for improvement and for public reporting. The teams usually conduct multi-day visits that include shadowing students through their classes, interviewing staff, students and parents, and reviewing evidence about the school. Several states in the U.S. have piloted SQRs, but in the face of NCLB and test-based accountability, they have either been dropped or operate on the margins. Rothstein, *et al.,* show SQRs can be used for a modest cost.

Using the Model with Current State Testing Systems

It is not clear how many states will apply for the ESSA Innovative Assessment program. Even if a full complement of seven are approved, most states will not participate in the first wave. The question, then, is how educators, schools and districts can apply the tenets of this model in their local practice, despite the continuing state tests.

In fact, that is what the schools and networks we highlight in Part III are doing. Even the NY Consortium students must pass the state's English Language Arts Regents Exam. Rollinsford students take SBAC in grades 3-6, Big Picture Learning (BPL) schools in the U.S. are subject to testing requirements, and so on. Some like Rollinsford are more middle class and white, but the Consortium and BPL serve primarily low-income students of color, as does Mission Hill School and many others. In short, the examples show that schools can move to high-quality performance and portfolio assessing despite the tests. But it is not easy.

The key will be willingness to bite the bullet and let the test results take care of themselves. ESSA makes this far more feasible. First, states no longer need to judge teachers by student test scores. Second, only a small percentage of schools must be identified as low performing ("priority"). Third, those schools design their own improvement strategies; if they do not lead to improvement on state accountability measures after three years, states are to provide assistance. In short, ESSA allows states to stop punishing and start – or strengthen – helping. These will only happen if states are willing or people pressure states to overhaul accountability.

Unions can help. The Oregon Education Association, for example, is collaborating with other organizations to help teachers and schools focus on formative and performance assessments (Oregon, 2015). Local associations can promote and support teacher, school and district efforts.

Even in states that do make major changes, some schools will be at risk, test scores will still be published in newspapers, and some districts will still push educators to beat other schools in the test game. But such problems are far less dangerous than NCLB. The task, then, is for teachers, administrators, parents and students to unite and fight to replace standardized testing with high-quality, teacher-led assessing.

Conclusion

FairTest's model begins with classroom-based evidence, emphasizing ongoing student work that has instructional value and produces assessable results. Teachers engage in formative assessing – feedback to students – as part of their instructional process. Knowledgeable teachers evaluate student learning in ways that are consistent with how other strong teachers would evaluate it. The rich assessment process also provides valuable professional development, positively influencing both curriculum and instruction.

There are good tools to establish consistency and comparability, but each must be used with caution. Some, such as the mandated reliance on existing standardized tests to determine comparability, are dangerous.

The one existing pilot, New Hampshire, represents a big step forward from the failures of NCLB and provides a valuable starting point for other states. Even better, given wider options under ESSA than NH has under its NCLB waiver, a system could allow greater variation in its local assessment systems.

In the end, ESSA opens a door that NCLB had closed. A new, less test-centric Department of Education under the next administration would allow states flexibility to move in the best possible direction on assessment and accountability. Whether that happens will depend on what states themselves attempt and what testing reform advocates – parents, teachers, administrators, students, school boards, and other advocates – are able to persuade states and districts to do and the DoE to allow.

References

Chalkbeat, 2016. "How to measure school quality beyond test scores? State officials count the ways," July 6. <http://www.chalkbeat.org/posts/in/2016/07/06/how-to-measure-school-quality-beyond-test-scores-state-officials-count-the-ways/#.V7SnkDUgG4E>

Cody, A. 2016. "California's New Model for Accountability," July 14. http://www.livingindialogue.com/californias-new-accountability-system-multiple-measures/

Coleman, J., *et al.,* 1966. *Equality of educational opportunity*. Washington, DC: United States Department of Health, Education, and Welfare, Office of Education. U.S. Government Printing Office.

CSBA. 2013. "State Priorities for Funding: The Need for Local Control and Accountability Plans." Fact Sheet. <https://www.csba.org/GovernanceAndPolicyResources/FairFunding/~/media/CSBA/Files/GovernanceResources/GovernanceBriefs/2013_08_LCFF_Fact_Sheet-funding_priority.ashx%20->

Darling-Hammond, L., Ed. 2014. *Next Generation Assessment*. San Francisco: Jossey-Bass.

Evans, C.M., Lyons, L. and Marion, S. 2016, April. "Comparability in Balanced Assessment Systems for State Accountability." Paper Presented at the National Council for Measurement in Education Coordinated Session, "Advances in Balanced Assessment Systems." Washington, DC.

Gallagher, C.W. 2007. *Reclaiming Assessment*. Portsmouth, NH: Heinemann.

National Education Association. 2016. "'Opportunity Dashboard' Indicator," <http://www.nea.org/assets/docs/Backgrounder-Opportunity%20Dashboard%20Indicator.pdf>

N.H. Department of Education. N.D. Performance Assessment for Competency Education (PACE). http://education.nh.gov/assessment-systems/pace.htm. See specific documents cited at end of NH chapter.

NH DoE. 2014. New Hampshire Performance Assessment of Competency Education: An Accountability Pilot Proposal to The United States Department of Education, November 21. http://education.nh.gov/assessment-systems/documents/pilot-proposal.pdf

Neill, M., *et al.* N.D. *Implementing Performance Assessment.* Cambridge, MA: FairTest.

NY Performance Standards Consortium. N.D. *Educating for the 21st Century Consortium: Data Report on the New York Performance Standards Consortium.* http://performanceassessment.org/articles/DataReport_NY_PSC.pdf

Oregon Education Investment Board, Oregon Education Association, and Oregon Department of Education. 2015, July. *A New Path for Oregon: System of Assessment to Empower Meaningful Student Learning.* http://www.oregoned.org/images/uploads/blog/FINAL_July_2015_Assessment_Document_a.pdf

Richmond, E. 2016. "Building Better Student Assessments," *The Educated Reporter*. Education Writers of America, June 23. http://www.ewa.org/blog-educated-reporter/building-better-student-assessments

Rothstein, R., Jacobsen, R., and Wilder, T. 2007. *Grading Education: Getting Accountability Right.* New York: Teachers College Press.

U.S. Department of Education. 2015. ESSA - <https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf>

U.S. Department of Education. 2016a.ESSA draft accountability regs - <https://www.regulations.gov/document?D=ED-2016-OESE-0032-0001>

U.S. Department of Education. 2016b. ESSA draft innovative assessment regs posted 07/11/2016 to federal register - <https://www.federalregister.gov/articles/2016/07/11/2016-16125/elementary-and-secondary-education-act-of-1965-as-amended-by-the-every-student-succeeds?utm_content=header&utm_edium=slideshow&utm_source=homepage>