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 is extracted from is available online at:
http://www.fairtest.org/assessment-matters-constructing-model-state-system

Preface

The way we measure students' academic progress sends powerful messages about what kinds of learning we value. When measurement systems are used to evaluate schools, the factors they emphasize can control classroom practices, for good or ill.

The test-and-punish approach embodied in the federal No Child Left Behind (NCLB) law undermined educational quality for many and inhibited school improvement. With these harmful consequences, it also delivered a message that deep learning and supportive, healthy school environments do not matter.

The damage has been most severe in the most under-resourced communities. There, the fixation on boosting test scores not only undermined teaching and learning. It also led to mass firings, school closings, and deteriorating educational climates that fed the school-to-prison pipeline. The Every Student Succeeds Act (ESSA), which replaces NCLB, creates the possibility for states to shift the focus of accountability from punishment of schools and teachers to policies that genuinely help improve educational quality and equity.

ESSA includes an "Innovative Assessment" pilot project, which opens the door to significantly better assessments. This report describes a model system that could be built under ESSA. We share it to empower educators, parents, students and other assessment reformers, as well as public officials, to use the option to reshape state systems. States that take advantage of this provision should focus on measurement practices that support rich, deep learning for all children. That will liberate classroom assessment from the confines of standardized tests, as well as provide useful accountability data.

Unfortunately, the law requires states to maintain some standardized testing in pilot districts. The tests are meant to ensure comparability between the new and the old. This requirement seeks to use testing and accountability to identify continuing educational inequities and correct them. But NCLB showed that test-driven "reform" has failed to improve educational opportunities and outcomes. States must ensure they do not trap new assessments within the limitations of standardized tests.

High-quality assessment is necessary for ensuring a strong and vibrant education for all. But it is not sufficient. In most places, attaining that goal also requires a significant increase in resources – for teachers, counselors, librarians, nurses, professional learning, wraparound services, community schools, libraries, technology, art supplies, and buildings. Often, major improvements in school culture and climate, student discipline and parent engagement are also needed.

In schools dominated by standardized testing, teaching, learning and a healthy climate are endangered. Schools that serve primarily low-income students, black and brown youth and recent immigrants, as well as those with disabilities, most need a major infusion of resources

and high-quality assessment. By improving student assessment and school evaluation, the nation can help ensure that schools meet the needs of every child. Without those changes, they will continue to be pressured to focus on a narrow conception of human potential.

The goal of this report is to contribute to high-quality education through sweeping changes in assessment. States can use the ESSA pilot to develop assessment systems that minimize standardized testing; place classroom-based, teacher-controlled, student-focused assessing at the center; diminish state and federal micro-control of education; provide tools to markedly improve learning outcomes; and produce sufficient data for evaluating schools in order to provide extra support and interventions where needed.

This report begins by describing the core components of a model assessment system under ESSA. It explains what the law requires of such a system and analyzes various ways a state can ensure comparability across districts that use classroom-based evidence as it builds a "system of systems."

Part II examines New Hampshire's new Performance Assessments for Competency Education (PACE) system as one model. PACE combines limited state testing and teacher-made Common Tasks used across districts to establish comparability. School assessment systems include teachers evaluating each student based on local tasks and a complete review of the student's work throughout the year.

Part III summarizes several other models that show the potential of classroom-based assessment. They demonstrate that performance assessments can obtain comparability and have long-term success. Their use significantly improves the chances that disadvantaged students will overcome obstacles and reach their potential. They show the critical role that assessment plays in high-quality schooling. They also show that districts and schools can implement performance assessments despite the state tests. The diminished accountability requirements in ESSA will make that option easier.

As this report rests on key understandings of what assessment should be, it concludes with a statement of principles to guide high-quality assessment and a discussion of its different uses.

Executive Summary

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 that will replace existing standardized tests. This initiative could lead states to fundamentally improve student assessment. ESSA replaces No Child Left Behind (NCLB).

To help states and education reformers take advantage of this opportunity, FairTest proposes a model system to maximize high-quality assessment within ESSA's constraints. The model described in this report represents a significant departure from NCLB's narrow test-and-punish framework. Unlike NCLB, which revolved around standardized test scores, the model begins with classroom-based evidence that emanates from ongoing student work. FairTest's model is rooted in exemplary practice and a set of principles derived from decades of assessment reform efforts.

Fish. Photo from Rollinsford Grade School.

The primary purpose of this innovative system is to support high-quality, individualized student learning. It is guided by teachers but substantially student controlled, with multiple ways to demonstrate learning. This encourages pupils to build on their interests. 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." Districts, or consortia of schools or districts, have the flexibility to ensure the structure and nature of their assessment systems address their local 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 examples of learning throughout the school year, including from any major projects. Teachers prepare a summative evaluation of each pupil. This includes a determination of the student's level of proficiency in line with state standards, as required by federal law. This data is 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, as ESSA requires, the state employs a set of procedures to determine whether students deemed proficient in one district would

receive a similar evaluation in another with a different local 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 actions.

FairTest's model is anchored in experience and evidence. New Hampshire is entering the third year of the Performance Assessment for Competency Education (PACE) pilot program. We describe PACE in some detail. Other important performance assessment examples include the New York Performance Standards Consortium, the Learning Record, the Work Sampling System, Big Picture Learning, and the International Baccalaureate program. The full report includes snapshot descriptions of these models.

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. While ESSA's requirements can create difficulties in implementing quality assessment for learning, the space for progress is large enough to make the 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, or performance tasks produced as part of ongoing academic activities. It can incorporate student work done out of school, such as internships, and can include group projects.

What differentiates this model from other proposals that emphasize performance tasks 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, in the content of their work, such as the specific science or history investigation; or in the mode of presentation, such as an oral report, written paper, video or computer game. Allowing student control has been shown to improve student learning.

Performance tasks take various forms, from short pieces of work to extended projects, and may include group tasks. The New York Performance Standards Consortium focuses on practitioner-designed, student-focused tasks. Other nations, such as Australia, use performance tasks as key components of their systems.

The value of portfolios is that they can clearly reflect curricular breadth (learning opportunities) and the quality of student work. With carefully designed scoring procedures, they can provide a more accurate and multifaceted indication of learning than standardized test scores. Examples

of well-structured assessments that include collections of student work include the Learning Record and the Work Sampling System.

Classroom-based assessments that emanate from student ongoing work in the curriculum differ from performance tests. The latter are tasks designed from outside the classroom (though often by teachers) and administered as summary tests or at points during the course of the year. To take advantage of student interests and help them learn to control their own ongoing learning, the former are the core of FairTest's model system. Rollinsford Grade School provides a strong example. However, performance testing can be a major improvement over current standardized exams and form a bridge to classroom-based assessing.

ESSA Innovative Assessment and Accountability Requirements

The most significant victory for improving assessment in ESSA is its "innovative assessment" demonstration project in which up to seven states can build new systems. Qualifying programs will have to meet ESSA's general mandates for state assessments as well as specific criteria for the pilot. New Hampshire already has launched a performance assessment pilot program under a waiver from NCLB granted by the U.S. Department of Education (DoE).

A full new system must include English language arts (ELA) and math assessments in at least grades 3-8 and once in high school, plus three grades of science. A state could, however, decide it will have a new system for only a portion of those (e.g., only science or only elementary grades). A pilot can start with a limited number of districts but must include a plan to become statewide in five years, though extensions are allowed.

NY Performance Standards Consortium school. Photo by Roy Reid.

The assessments can vary across districts — provided the results can be accurately compared. During the pilot period, the new assessments must also be comparable with current state tests. ESSA draft regulations list ways in which such comparability can be established. These include administering the state exam to all students in the pilot; or only to students in one grade each in elementary, middle and high school; or both the state test and the new assessments to a demographically representative sample of students in the pilot once in each grade span; or some other DoE-approved method a state creates.

Comparability within a New Assessment System

In order to establish comparability among students participating in the innovative assessment pilot 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 (commonly) samples of completed work are re-scored by someone other than the students' classroom teacher. This is done to ensure consistency of marking across educators, schools or districts. Moderation requires the use of common scoring guides, or "rubrics," and samples of student work that exemplify differences among student work at the various proficiency levels ("exemplars"). The Learning Record and NY Consortium use moderation. It is also a key part of the New Hampshire pilot. Other nations often use such procedures with performance assessments.

The main disadvantage of statewide scoring guides is the risk of lowest-common-denominator rubrics that push toward mediocrity. State scoring guides could enforce back-door standardization, as tests that require writing in response to a prompt often do. Lower-quality rubrics often focus on quantity (e.g., "provide two examples") rather than quality. On the other hand, strong rubrics can focus attention on the most important characteristics of much student learning.

Anchor tasks and tests. ESSA draft regulations recommend the use of "anchor tasks" to ensure comparability between new assessments and old tests and to establish comparability across districts within the new system. In this procedure, the same performance tasks are administered to students across participating districts.

While the new system is being built, all participating districts must administer the current state tests in at least some grades. Results are analyzed to ensure proficiency levels on anchor tasks are comparable to the state tests and participating districts are scoring them consistently. Anchor tasks or state test scores also can be compared with local assessment results, as New Hampshire does in its pilot program.

Anchor tasks are a reasonable means to establish comparability. Done well, they should fit cleanly into the curriculum in many schools. Writing and scoring them can provide important learning opportunities for teachers.

One significant disadvantage is that these tasks do not emerge from student interests within the curriculum. Thus, they may not engage all students, and may not connect well to what an individual is actually studying. These problems can lead to students performing less well. In addition, 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. They take substantial teacher time to write, time that could be used in other educationally valuable ways.

Validation studies. ESSA requires states to annually compare pilot results with their current tests. When the new system is complete, the current tests need not be used. A state could then conduct validation studies in which results across districts are compared in light of the state's standards-based definition of proficiency. This could happen once every few years rather than annually for districts that show strong comparability.

Addressing Potential Contradictions in Building a New System

There are potential obstacles to fitting high-quality local assessing into ESSA accountability mandates. However, these hurdles should not prevent states from moving ahead. The positive potential far outweighs the dangers. The greatest threat lies in the requirement to ensure comparability.

Good performance assessments measure a wide range of important learning and skills that are not covered by standardized tests. Thus, they should not be expected to be directly comparable, even if both are in some ways anchored in state standards.

Teachers may confront the problem of serving two masters: the old tests and the new performance assessments. They could face pressure to establish consistency between classroom evidence and the tests. This could distort how they design the new assessments and evaluate student results.

Performance assessments are intended to improve learning in ways that may not show up on standardized tests. Ideally, they can narrow gaps in achievement in areas that really matter for students' future success, such as designing an extended project and persevering to completion. The danger is that discrepancies with results from current tests could lead to dismissing other forms of learning gains that are more meaningful. This may be particularly harmful in schools that had most heavily focused on test scores, and thus for low-income children, children of color, English language learners and students with disabilities.

Comparability has value, but the great value of assessment is to enrich student learning. The dangers from comparability requirements could be lessened if districts are not forced to alter their local assessment scores to be comparable to state test results. However, as long as current standardized exams are falsely presented as the "gold standard," the problem will remain.

ESSA Opens a Door NCLB Had Closed

If the next U.S. secretary of Education understands the damage done by NCLB's focus on testing and wants to repair it, states could have the flexibility to move in the best possible direction. It will be up to assessment reform activists to persuade the new president to appoint a secretary who understands what is at stake. At the same time, parents, teachers, administrators, students, school boards, and other reform advocates will have to pressure their states and districts to take advantage of their new opportunities.

Kindergarten. Rollinsford Grade School photo.

In addition, teachers, schools and districts can move ahead on using performance assessments while cutting back on locally mandated standardized tests. As this report discusses in Part III, some schools have done so, with positive results for children.

New Hampshire: An Innovative Performance Assessment

New Hampshire received an NCLB waiver to begin constructing what is intended to become a new statewide system, the *Performance Assessment for Competency Education (PACE*). As such, it has become a national model.

PACE started with four districts in 2014-15, then eight the next year. It will include nine districts in 2016-17, with 10 more preparing to join. The state expects to become part of the ESSA innovative assessment program. PACE was designed to unite rich learning assessed locally with federal accountability requirements. It includes the state ELA and math tests administered once each in elementary, middle and high school; Common Tasks administered in the non-test grades, 3-11, plus science in three grades; local tasks; and an "Achievement Level Determination" (ALD).

There is one Common Task for each grade and subject, written by teachers and reviewed by independent experts. These and other tasks vetted for quality by experienced teachers and measurement experts are assembled into a "bank" for local use. In addition to helping design the assessments, teachers participate in moderation sessions to strengthen their ability to score accurately.

Local systems focus on multiple assessment tasks made by district teachers plus items from the bank. These are scored locally. Teachers across districts re-score samples for training purposes. At the end of the year, each teacher makes an ALD competency determination based on the body of work by each student over the course of the year, including task results.

The state developed multiple procedures to determine consistency. Common Task scoring and results from the state test (the Smarter Balanced Assessment Consortium, SBAC) were

compared across districts. Each was adequately consistent. Most important, the locally determined ALDs were consistent with Common Task and SBAC results at the district level. This process complies with the comparability evaluation required by ESSA and the state's waiver. Because ALDs are based on a full body of work, not just the tasks, the positive results add to the evidence that a state can design a system based on varied local assessments.

There are important benefits. The system offers students a range of ways to show knowledge and skills, many of which are not adequately covered by SBAC. The assessments tap higher order thinking and problem solving and strengthen teacher capabilities.

There are also concerns. Initial task quality is an issue, but evidence shows teachers get better at writing tasks and rubrics over time. Some rubrics are seen as too simplistic by some experts in performance assessment, focusing only on quantity not quality, and can foster narrow forms of instruction, such as writing "five-paragraph essays." They are also inserted into the curriculum (though based on it) as a form of test, rather than emanating from the ongoing student work during the year. The Common Tasks must fit into the traditional instructional program offered by most districts, which undermines their value for inquiry-based learning that allows significant student control. By prescribing only one way to assess, the tasks can narrow the possible range of student learning that is encouraged.

New Hampshire has one district that has not joined PACE, the *Rollinsford Grade School* (RGS), which has designed its own performance assessment system (see Part III). It illuminates some of PACE's limits.

Rather than rely on performance tasks that are externally developed, or even teacher-made tasks, RGS prioritizes teacher-guided, student-focused assessing that evolves out of its inquiry- and project-based curriculum. Students have substantial choice in identifying questions to explore. The resulting products, from books read and written about to science and social studies investigations, provide some of the evidence of student progress and challenges. Other evidence comes from ongoing observation of and conversations with students. These lead to "competency determinations" based on their school-developed competencies.

The key reasons Rollinsford has not joined PACE are the confines of the task-based system and the large staff time commitment for PACE work that would come out of school instructional time. Building its inquiry- and project-based instructional program has demanded a lot from RGS teachers; shifting that time to working on PACE tasks would, the school believes, undermine its own efforts. However, RGS participates in PACE discussions, which RGS staff have found valuable.

NH's current NCLB waiver is rooted in the local and common tasks system. A critical question is whether, under ESSA and a new US DoE, PACE could include schools, such as Rollinsford, which have different performance assessment systems. If so, RGS could be a model for the further evolution of PACE and other states.