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

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.

PROUD PARENT
OF AN HONOR STUDENT WHO REFUSES TO TAKE ANOTHER STANDARDIZED TEST

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.

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.

NY Performance Standards Consortium school. Photo by Roy Reid.

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.

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.

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.

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

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).

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.

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

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>

Part II

New Hampshire PACE

New Hampshire received a waiver in 2015 from No Child Left Behind (NCLB) to begin constructing a new statewide system, the Performance Assessment for Competency Education (PACE). Implementation started with four participating districts in the 2014-15, school year. It grew to eight in 2015-16 and includes nine in 2016-17, with 10 more preparing to join. The waiver has been extended through 2016-17. The state expects to be part of the Every Student Succeeds Act (ESSA) "Innovative Assessment" pilot.

New Hampshire describes its new system in these terms:

> "One of the motivating reasons why NH is piloting a new kind of accountability system is because the state wants to support meaningful learning and continuous improvement models, as well as promote shared accountability between districts and the state" (Evans, Lyons & Marion, 2016).
>
> "The PACE system is based on a rich system of local and common (across multiple districts) performance-based assessments that are necessary for supporting deeper learning as well as allowing students to demonstrate their competency through multiple performance assessment measures in a variety of contexts" (Marion and Leather, 2015).

For federal approval under ESSA, PACE must demonstrate that the new assessments are valid and reliable, and are comparable among themselves and with the state's current tests. PACE can be understood as a structure to unite two purposes: deeper learning and comparability for accountability. The first begins with locally designed and controlled performance tasks scored by teachers. These are intended to improve teaching and learning and provide key evidence for an "achievement level determination" (ALD). The determinations, placing every student's level of proficiency on a four-point scale, are made by classroom teachers using information gathered across the school year. New Hampshire has designed a structure to ensure these "competency determinations" are comparable across the districts and accord with state test results. The ALDs are also the basis for determining each school's level, an ESSA requirement. Thus, the ALDs are the core of the PACE system.

Description of PACE

PACE includes three kinds of assessments:

- State tests, which are the Smarter Balanced Assessment Consortium (SBAC) exams in English Language Arts (ELA) and math, given in one grade each in elementary and middle school, and the SAT college admissions test in grade 11.

- Teacher-made, PACE-wide common performance tasks in ELA and math in grades 3-8 that do not have state tests, as well as science tasks at grades 4, 8 and in high school; and
- District teacher-made tasks in all three subjects in grades 3-11 that do not use the state tests, with many districts also developing tasks for earlier grades and state-tested grades.

Common Tasks

Teachers from participating districts design 17 Common Tasks (CT) with task-specific scoring guides, one per required subject/grade. "There are three main purposes for the common tasks across districts: 1) to help measure the degree of cross-district comparability of scoring, 2) to serve as models of high quality tasks and build local capacity, and 3) to contribute to the long-term goal of building a large task bank from which districts can draw for local assessment purposes" (NH DoE, 2016a, p. 4).

As models, "The tasks are designed and reviewed specifically to allow for independent student inquiry, multi-step problem solving and argument building, and typically allow for multiple possible solutions" (NH DoE, 2016a, p. 5).

Tasks are designed to be accessible to students with disabilities and English language learners. The state uses Universal Design for Learning in writing the Common Tasks, to maximize accessibility, and also allows accommodations in line with those provided for SBAC. It has separate assessments for students with the most severe disabilities. It administers the WIDA test to English language learners, who do not take the CTs, as allowed under NCLB/ESSA.

Once CTs have been administered, the state posts them to a website for use by districts as local tasks and by teachers in their classrooms. The assessment bank is complemented with tasks from other teachers and states that the state has reviewed and approved (Marion-Leather, 2015, p 13).[i]

Local Tasks and the "Achievement Level Determination"

While Common Tasks are important, the heart of the new system is the use of locally made tasks included in local systems combined with each teacher's "achievement level determination" (ALD) for individual students.

In PACE, each participating district designs a system of performance-based assessments tied to the state's subject area "competencies" and standards (NH DoE, NDb). Districts submit their systems to the New Hampshire Department of Education (NH DoE) for peer review and approval (NHDoE, 2016a). Local assessments sometimes include gathering information on students' "work-study practices," but these are not part of any statewide data or subject competency determinations.

The number of local tasks varies by district, grade and competency. These tasks are "curriculum-embedded and administered in local districts." Deputy Commissioner Paul Leather (2016) explained this means teachers base the tasks on their curricula and administer them at an appropriate time.

Among directives to districts regarding the local assessments are (NH DoE, 2016a, p. 10):

> "3. Students must be allowed multiple opportunities to demonstrate evidence of achieving a competency over the course of a year.
>
> "4. Districts must use a mixture of locally-designed performance assessments and assessments drawn from validated state/multi-state task banks."

Photo from Rollinsford Grade School.

"Leaders and teachers in each district determine how to score their local competency-based assessments" (NHDoE, 2016b). Each district task creates its own rubric, though in some cases a general rubric can be used across tasks (e.g., different writing samples). Local scoring, the state recommends, begins with selecting 10-20 previously completed tasks from across the range of student achievement (NHDoE, 2016, p 36 ff). Teachers sort these into four performance levels. They then use the anchors with the local task rubrics to score new student work. The results of the local tasks are included by the teachers as they render their ALDs.

Teachers decide each of their students' proficiency levels – the ALD – on a 1-4 scale in the subjects and grades mandated for NCLB/ESSA accountability. These determinations take the form of a summary judgment by each PACE teacher about each student. It includes all the academic information gathered from the student's work over the year. The judgments are based on SBAC's "Achievement Level Descriptors," modified to fit PACE. These Descriptors also serve as a scoring guide that enables the analysis of comparability (NHDoE, 2016, p 10 and pp. 41-46).

Establishing Comparability

PACE has designed a complex system for establishing comparability. It begins with the requirements districts face to join PACE. It includes tools for re-scoring ("moderating") local and common assessments, and protocols to determine the degree of comparability across SBAC or SAT, PACE common tasks, district tasks, and the ALD. Disaggregated group scores are also examined. New Hampshire points out that comparability does not require psychometric exactness, such as the comparison of student scores on a single standardized test. PACE's methods also serve to demonstrate validity and reliability (required by ESSA), as well as assist teacher professional development.

First, the design, validation and approval of district systems. Staffs in participating districts go through extensive training and two years of peer review. Local systems must meet a set of criteria for quality, have well-prepared core staff, demonstrate validity, reliability and comparability, and adequately address such issues as bias/fairness.

Second, determining consistency in scoring common tasks across districts. The Common Tasks serve as reference points for comparing results across districts. Students take one CT per subject in each grade that has no state test. New Hampshire officials did not want a larger set of such tasks as they do not want a new form of state exam.

Each PACE CT has a scoring guide teachers use to rate student work. To help teachers do this well, the state sets up sessions for teachers to collectively re-score a sample of completed tasks and discuss the results. Teachers bring this knowledge back to their districts to apply to local tasks.

NH Department of Education staff worked with measurement experts from the National Center for the Improvement of Educational Assessment (NCIEA) to design a procedure to ensure accurate scoring. The process builds on models from Australia and Britain (Evans, Lyons & Marion, 2016). In the first year of the pilot, local teachers' common task scoring was adequately comparable across the four districts (NHDoE, 2016, p 18). Time and experience should strengthen consistency, though the influx of new districts and teachers will make this an ongoing project.

Third, and most central, ensuring accuracy and consistency of the local competency determinations. As the state says, "Comparability in scoring performance assessment tasks is important but the ultimate goal is that 'annual determinations' are comparable across school districts" (NH DoE, 2016b). The state devised procedures for linking district teacher-determined ALDs to common task scores and SBAC results.[ii] Teams of teachers compared the 2015 local ALDs with samples of completed Common Tasks from each district in each subject across the four achievement levels. Teachers did not score work from their own districts. They found no systematic variation. That means overall scoring was adequately consistent across the grades and districts (Evans, Lyons & Marion, 2016).

PACE also compared the proportion of students across the state scoring at levels 3 and 4 on SBAC in the tested grades with ALDs from pilot districts in the other grades. The study found the scores across districts “were quite similar, indicating a high degree of comparability between PACE and non-PACE districts” (NH DoE, 2016b, p 11; NH DoE, 2016a, p. 11). In addition, “the differences in performance among major subgroups and the all students group were similar for both PACE and Smarter Balanced annual determinations” (NH DoE, 2016a, p. 22).

Consequences

The NH DoE 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, the analysis of the results from the first year concluded that no district systematically scored more stringently or more leniently, indicating there would have been no need to modify any district’s scores.

New Hampshire began by including districts the state education department viewed as well-prepared for the new program. It recognizes that capacity will need to be built to expand PACE to districts that are less ready for performance assessment. Under ESSA, NH will have five years to expand PACE statewide.

Benefits and Concerns

There are significant benefits from the work being done by PACE:

- First, the performance assessments offer students a range of ways to show their knowledge and skills, many of which are not adequately covered by SBAC or SAT. Since PACE local assessments are tied to the curriculum, students are being taught content they may not have covered in the past due to pressure to raise standardized test scores. That includes higher-order thinking, applications of knowledge, problem solving, communicating, and connecting learning across subject areas.
- Second, designing, administering and scoring the assessments provides a vehicle for professional learning. Teachers deepen their knowledge about assessment, curriculum and instruction, and strengthen their ability to work and learn together.
- Third, PACE is creating a valuable model for the nation. Several states were developing performance assessment systems in the 1990s, but NCLB halted most of that work. While a knowledge base exists, for the most part states and educators have to start nearly from scratch. They also have to address accountability and comparability issues that were generally not concerns in the 1990s.

- Fourth, the moderation and comparability systems developed by PACE and its primary technical partner, the NCIEA, can be useful to other states.

Photo from Rollinsford Grade School.

PACE is still very much a learning process. For example, beginning teachers typically have limited capacity to craft good tasks. However, evidence shows educators quickly learn how to make tasks better by trying them out in classrooms, sharing with other teachers, reflecting and discussing.

New Hampshire Principal Jonathan Vander Els (2015) observed, “Each performance assessment that I see being constructed is of higher and higher quality. This is due not only to our teachers’ overall increased understanding of assessment in general, but also to their increased understanding of the nuances within each assessment. Considerations such as specific wording of a question, students’ background experiences, ability to provide appropriate accommodations, and the level of the depth of knowledge are intuitively included.”

Despite progress, there are concerns. Some observers fear a system like PACE will end up allowing low-performers to skate by, allowing some students to slip through the cracks. These critics want to retain a system rooted in standardized tests, despite evidence of damage to educational quality and student learning. While most have acceded to giving “innovative assessments” a try, they may fight for constraints that end up undermining assessment quality and thus its benefits for teaching and learning.

Other concerns emerge from those who have been engaged in performance assessments. These include:

- Some question the quality of tasks, though they recognize they will improve. Tasks may be artificial, not exemplifying real-world problems. Other issues may reflect the history, in New Hampshire and elsewhere, of using narrow rubrics to judge limited forms of writing. This includes the “five-paragraph essay,” which lacks real world applicability.[iii]
- The performance tasks are administered as tests. They do not evolve out of ongoing student work in the curriculum, as they do at Rollinsford NH Grade School, the New York Performance Standards Consortium, Big Picture Learning and elsewhere (see Part III). In those instances, students have strong say over their tasks and projects – a key principle for high-quality assessment. A related concern is that tasks may not be particularly

engaging for many students. The famous Coleman report (1966) found that next to parent's socio-economic status, the most significant predictor of academic success was a sense of control over learning, which can be strengthened when students choose their work.

- The common tasks must fit into a traditional curriculum. The PACE tasks can demonstrate knowledge, problem-solving skills and communications, but they are not part of an evolving inquiry. At Rollinsford, the NY Consortium, etc., the projects and tasks are themselves learning experiences that can be assessed during the process and at completion.
- Due to how they are constructed, says Rollinsford Principal Kate Lucas, "The current tasks offer a *prescribed way to assess students.* For example, the ELA grade 6 PA was to write a persuasive essay (rain forests). We question if an essay is the best way for all students to demonstrate their ability to synthesize and analyze information and then persuade others. Is it possible *to offer multiple modalities* for demonstration that require the same skills? This opens up the door to success for *all* students. It also requires deeper thinking and decision making" (Lucas, 2016, emphases in original). Of course, writing is a highly valued skill that schools do need to teach and assess – but the "essay" form is not the only possible mode for demonstrating content knowledge.

- The process demands a large commitment of teacher time from participating schools. Certainly it is a learning experience for many, but Rollinsford fears that the time commitment would detract from its own labor-intensive efforts to improve. More generally, the question of teacher time is a significant issue that designers of new systems will have to address.

Can There Be a "System of Systems?"

The broader issue raised by Rollinsford concerns the vision of education, for example, whether it should be inquiry-driven (project-based) or more traditional. It also concerns whether a state developing a new system can allow a Rollinsford (or a Big Picture, a NY Consortium, or a school using the Learning Record) to join as a partner despite its different approach to assessments. NH Deputy Commissioner Paul Leather thinks it could not under the current NCLB waiver.

This could change. ESSA allows differing local assessments (as in New Hampshire) provided there is a vehicle to establish comparability. PACE does this by linking local competency determinations to Common Tasks and the SBAC tests. The determinations are made by teachers based on the evidence gathered over the year. Rollinsford does exactly that. If Rollinsford joined PACE under ESSA requirements rather than the current NCLB waiver, it could continue to use its own assessment processes rather than design local tasks. They would be something of an outlier among PACE partners but could be a powerful example to districts interested in moving toward inquiry-based schooling.

What Rollinsford, the NY Consortium and other examples lead to is a “system of systems” in which local assessment systems may vary but all must provide evidence of comparability. A mix of anchor tasks and moderation can do that.

PACE has opened the door toward creating a state system of performance assessing that is significantly decentralized, places teachers at the heart of the process, ensures significant professional development, and directs students toward deeper learning. As such, it is a strong model for the nation to consider. FairTest calls on states to take account of the even richer possibilities of allowing a range of local systems, including those that are inquiry driven and build on student work as it evolves out of the curriculum.

References

Carla M. Evans, Susan Lyons, & Scott Marion. 2016. “Comparability in Balanced Assessment Systems for State Accountability.” Paper Presented at the National Council for Measurement in Education (NCME) Coordinated Session “Advances in Balanced Assessment Systems,” April, Washington, DC

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.

Leather, P. 2016. Personal email communications, July 26.

Lucas, K. 2016. Personal email communications, July 26.

Marion, S. 2015, Feb. Two sides of the same coin: Competency based education and Student Learning Objectives. Published by Competency Works. http://www.competencyworks.org/resources/two-sides-of-the-same-coin-competency-based-education-and-student-learning-objectives/

Marion, S., & Leather, P. 2015. Assessment and accountability to support meaningful learning. Education Policy Analysis Archives, 23(9). http://dx.doi.org/10.14507/epaa.v23.1984

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

NH DoE. 2016a. New Hampshire Performance Assessment of Competency Education: Progress Report to the United States Department of Education, March 1.

NH DoE. 2016b. Moving from Good to Great in New Hampshire: Performance Assessment of Competency Education (PACE), revised, January. http://education.nh.gov/assessment-systems/documents/overview.pdf

NH DoE. NDa. New Hampshire Accountability Pilot Overview Performance Assessment of Competency Education (PACE). http://education.nh.gov/assessment-systems/documents/pilot-overview.pdf

NH DoE. NDb. State Model Competencies. http://education.nh.gov/innovations/hs_redesign/competencies.htm

Vander Els, J. 2015. "Setting the PACE: Teacher Assessment Practices in a Competency-Based Education System." Competency Works. http://www.competencyworks.org/insights-into-implementation/setting-the-pace-teacher-assessment-practices-in-a-competency-based-education-system/

[i] There is a list of 2015 tasks at http://tinyurl.com/alltaskslist, but it contains only task titles. For an example, see Water Tower, http://www.ewa.org/blog-educated-reporter/building-better-student-assessments, which this report summarizes in part I. See also Algebra task at Appendix I of NH DoE Report to USDoE, March 2016. The state plans to make tasks available online in 2016-17.

[ii] See chart, NHDoE, 2016b, p 10, and related discussion; NHDoE, 2016a, p, 6 ff, also Appendix J, on details developing ALDs, and appendix F, on producing the ALDs.

[iii] For some brief "rules" about rubrics, see "Gail's Axioms" in Neill, *et al., N.D., Implementing Performance Assessment*, p. 30. FairTest, Cambridge, MA. For examples of high-quality rubrics, see New York Performance Standards Consortium, *Education for the 21st Century*, http://performanceassessment.org/articles/DataReport_NY_PSC.pdf.

Part III

Performance Assessment Examples

The U.S. has fine schools that make great use of performance assessment. They do so despite the damage wrought by No Child Left Behind with its insistence on dominating education with standardized tests.

Less common are systems or networks that exemplify high-quality assessment and demonstrate it can be done on a large scale. In some cases, these networks have established comparability of the meaning of achievement levels across schools.

Part III presents examples:

- Rollinsford Grade School exemplifies high-quality teaching, learning and assessing.
- The New York Performance Standards Consortium is a subsystem of 38 public high schools that uses performance-based assessments.
- The Learning Record is a portfolio-like tool that was used in a growing number of schools prior to NCLB.
- The Big Picture Company is a network of mostly public schools which prizes performance assessment, though its schools vary in their assessment practices.
- The Work Sampling System combines portfolios, checklists and short summaries of elementary students' progress across academic and other domains.
- The International Baccalaureate is a worldwide system of schools that uses a variety of forms of performance assessments.

We could have included others. At the school level, for example, there is Boston's wonderful Mission Hill School, which has regularly opened its doors to us. We decided the excellent videos about MHS would be a superior alternative to a short write-up, as is the case for other schools. Systems that extensively use performance assessments also flourish in other countries. Linda Darling-Hammond's research provides examples, as does FairTest's fact sheet on multiple measures. (For more details, see Resources at end of this Part.)

Rollinsford Grade School

> “The kids blow my mind every day. The thinking they share with me or others, the kindness they show each other, and how they know to ask for and get help.”
>
> – Principal Kate Lucas

This small-town New Hampshire public school has used performance assessments as part of a remarkable transition to an inquiry-based instructional approach, in which the learning process is as important if not more important than the product.

RGS has organized its work around four pillars, which guide students toward becoming:

- Collaborative and compassionate members of our global society;
- Lifelong learners;
- Architects of their personal wellness; and
- Critical thinkers and problem solvers.

Photograph by Rollinsford Grade School

Rather than rely on externally developed performance tasks, or even teacher-made tasks administered to the students as tests, it prioritizes teacher-guided, student-focused assessments that evolve out of the curriculum. That is, within the curriculum, students have substantial choice in identifying questions to explore. The resulting products, from books read and written about to science and social studies research, provide some evidence of student progress, difficulties and attainments. Other evidence comes from ongoing observation of and conversations with students (also known as “kid watching”).

FairTest staff visited Rollinsford in May 2016, visiting classrooms and talking to staff and students. That day, grades 1-2 and 5-6 were sharing samples of student work with one another and grades 3-4. The following week, all grades shared in an open house for the town. Students chose the work to present, sometimes completed and sometimes in progress. Many presented the books they had read, focusing on one they had written about and sharing a list of others. While there was overlap, most students had read some books that few or no other students had.

One group of grade 5-6 students investigated Southeast Asian river dolphins. They researched biology and ecology, wrote a report, prepared visual presentations, and sold tie-dyed T-shirts to raise money to save these endangered animals. They happily talked about their work and findings. On the wall outside a grade 4-5 classroom were student questions for a social studies investigation that was just beginning into civil rights and civil liberties. One student had asked, how can we stop the KKK and protect civil liberties? These are the sorts of "questions, problems and project-based learning (QPP)" around which the school's teaching and learning is organized.

Another girl's showcase was about the importance of "first impressions" and "type A" personalities. She reflected on her first days as a fifth grader in a grade 5/6 multi-age classroom and how her "bold" contributions from the start left a bad taste in the mouths of her sixth-grade female classmates. She embarked on a close reflection on her personality and how she might change the impression she makes on others. Ms. Lucas (2016) said, "Her greatest realization was that because she had learned to allow others to contribute first and disagree by saying things like 'I see that perspective and I'm curious how it might change if we think about it this way,' she has made and kept friends."

Rollinsford's approach to instruction and assessment affects both academic content and students' self-awareness. The purpose of the school is to help children grow into good adults, not to implement an assessment system or measure "competencies." The qualities described in the epigram above – thoughtfulness, kindness and knowing how to ask for and get help – are what counts.

When students graduate from Rollinsford, they cross the Maine state line to attend the Marshfield middle and high schools. RGS students do quite well in this highly regarded but traditional school – though their good grades are, to RGS staff, secondary to their broader goals or "pillars." Ms. Lucas noted that Marshfield teachers said RGS students "ask a lot of questions," another sign of an actively engaged student.

NH has adopted a "competency" system in which students advance and graduate by demonstrating sufficient knowledge and skills in subject areas. Rollinsford staff designed their own competencies four years ago. These share some alignment with state competencies and other subject-area national standards. RGS reports student progress in light of the competencies on a trimester basis.

However, Rollinsford Grade School (RGS) has so far chosen to limit its participation in the NH PACE performance assessment pilot (see Part II). New Hampshire districts/schools that develop and implement the state performance tasks are Tier 1 schools. Rollinsford participates in a second group, Tier 2, which engages in discussion and planning, often in preparation to become a Tier 1 school.

Principal Kate Lucas was pleased they participated in Tier 2: "We found the conversations helped to solidify our philosophies and approach. They forced us to be certain in our position and have substantial support/evidence of student success. We also very much liked learning with the other Districts."

RGS has not joined Tier I for several reasons. First, teachers would have to take a large amount of time from school days to design the common tasks used across districts. Second, they conclude that administering performance tests is not the same as assessing student work that arises from the curriculum. They fear that focusing on such test tasks would undermine their own approaches to teaching and learning. Ms. Lucas said, "It would have been extremely difficult to embed the assessment within our school culture given the PACE performance task 'constraints' (rules, procedures, expectations, etc.)."

The downside is that they continue to administer the SBAC exam in grades 3-6, which is not compatible with RGS educational practice but can be administered without pulling teachers from their classrooms for task development. They largely ignore SBAC. RGS will continue to participate in Tier II and the conversations with other schools, and consider whether conditions will facilitate their participation in Tier I and the full PACE project. (See further discussion of this issue in Part II.)

Resources

Lukas, K. 2016. Personal communication, email, August 1.

Interested readers can explore Rollinsford's website for discussion of their understanding of inquiry-based learning, what they mean by levels of competency in the subject areas, samples of student projects, and more. https://sites.google.com/a/rollinsford.k12.nh.us/rollinsford-grade-school/

FairTest's Monty Neill interviewed Principal Kate Lucas, literacy specialist Shawna Coppola, and consultant Lynne Stewart (from the Center for Collaborative Education), and talked with staff and students who were engaged in a showcase of their work, on May 12, 2016.

New York Performance Standards Consortium

Performance-based assessment works well for all students, but its success with the most vulnerable students is what makes the outcomes of the New York Performance Standards Consortium so impressive. The Consortium now includes 38 public, non-charter high schools, 36 in New York City.

The Consortium's assessments are created by teachers and rooted in in-depth, project-based curricula and teaching. Its 2015 report, *Education for the 21st Century*, demonstrates that Consortium schools significantly outperform those in other New York City public schools while serving a similar population. In particular, more students from all demographic groups graduate, go to college and stay in college.

NY Performance Standards Consortium biology students at work. Photo by Roy Reid.

The Assessments

To "demonstrate college and career readiness and to qualify for graduation," all Consortium schools require students to complete four Performance-Based Assessment Tasks **(**PBATs): an analytic literature essay, a social studies research paper, a student-designed science experiment, and higher-level mathematics problems that have real-world applications. They include both written and oral components.

The Consortium has permission from the state Department of Education to administer only one of the state graduation tests, English Language Arts. The PBATs, generally completed in 11th and 12th grades, replace the Regents exams in other subjects and for school accountability.

Education for the 21st Century explains that the PBATs "emerge from class readings and discussion. In some classes, the tasks are crafted by the teacher and in other instances by the student." For example, in social studies, each student must write and then orally defend a research-based analytic paper on questions that have grown out of a history, government, or economics class. The Consortium's data report includes samples of the wide range of social studies interests addressed by the students, as well as similar samples for the other required tasks. In the oral defense for each PBAT, the student responds to questions from a panel of teachers and outside experts.

As Urban Academy history teacher Avram Barlowe (2016) explains, the PBATs require students to learn perseverance, how to assess and apply evidence, and explain their thinking in these assessments in written and oral forms. They “demand that students learn, through practice, how to read, write, calculate, observe and research in a critical manner.” A DVD series, Teacher to Teacher, shows how teachers and students build their courses to attain these ends.

All the PBATs and oral defenses completed for the common graduation requirement are evaluated using Consortium-wide scoring guides (“rubrics”). The report includes rubrics for the four subjects. These well-developed assessment standards, written and revised as needed by Consortium teachers, allow accurate evaluations of student work across schools. Samples of the work are blindly re-scored to evaluate both reliability of scoring and the challenge level of teacher assignments. Samples of student work (“exemplars”) that have gone through a series of moderation studies help both scorers and students to think about high-quality work.

Each school maintains collections of work that chronicle a student's growth. The college persistence data show that the extensive reading, writing and long-term planning required for the performance assessments prepare students well for higher education.

The Results

Consortium schools follow the same admissions process as other non-exam New York City high schools. The student population of the Consortium’s New York City schools mirrors the city’s student body (only two schools are outside NYC). These schools have nearly identical shares of blacks, Latinos, English language learners and students with disabilities. Students enter Consortium high schools with lower ELA and math average scores than citywide averages.

The Consortium dropout rate is half that of NYC public schools. Graduation rates for all categories of students are higher than for the rest of NYC and nearly double the city’s rates for ELLs and students with disabilities. The Consortium also tracks student persistence in college. The report demonstrates that rates of enrollment in third semester exceed the national average. (The Consortium is updating this information, for release in fall 2016).

Resources and References

The Consortium website is at http://performanceassessment.org

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

Teacher to Teacher, a series of videos and books on the Consortium. http://www.teacherscollegepress.com/teachertoteacher.html

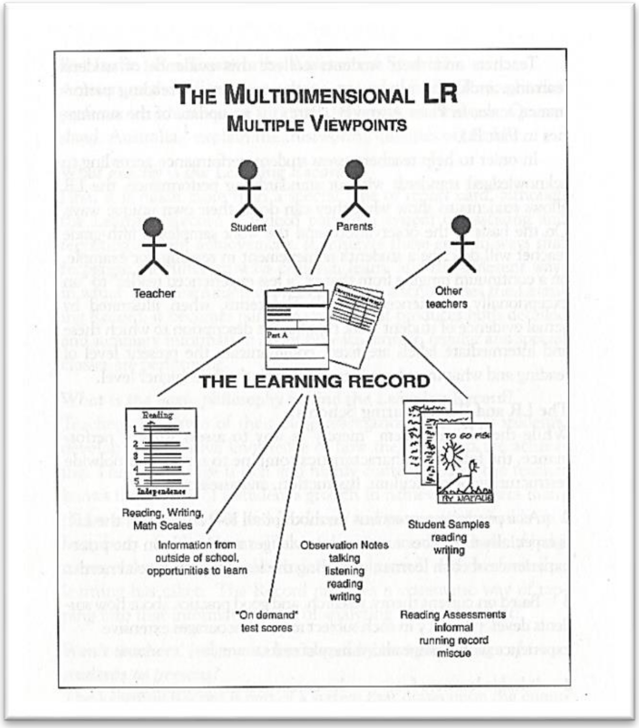
Barlowe, A. 2016. "The New York Performance Standards Consortium," workshop presentation at Save Our Schools Conference, July 9, Washington, DC.

See also the Webinar on Performance Assessment, sponsored by the Forum on Educational Accountability (FEA). It features Ann Cook of the Consortium, Sally Thomas of the Learning Record, and Monty Neill from FairTest. http://fairtest.org/view-new-webinar-authentic-performance-assessment . The webinar is a good place to start; you can obtain just the slides as well, no voice (though listening and watching provides more depth).

The Learning Record

The *Learning Record* (LR) is a system of literacy and mathematics assessment, K-12, maintained and monitored by classroom teachers to document student progress toward agreed upon goals and standards in reading/language arts and math. Parents and students also contribute evidence to the Record. It is a superb means for gathering information to assist individual students and teachers. Its re-scoring ("moderation") procedures demonstrate high inter-rater reliability between classroom teachers and independent re-scorers. Thus, it could be used for public reporting.

The LR was developed in England as the *Primary Language Record* for use with multi-lingual, multi-cultural young children in reading, writing, speaking and listening. Its structure provides a consistent framework for teachers to gather and evaluate evidence of student learning. It was adapted in the U.S. and expanded to include higher grades and mathematics. It grew to 25 U.S. schools in 2001, including many Bureau of Indian Affairs schools, before being largely swept aside by NCLB testing requirements.



In the LR, each teacher documents and evaluates each student's progress over the course of the year. It begins with the teacher summarizing two conversations, with the student and the parent(s), each focusing on the child as a learner. Such information has been documented as improving teachers' understanding, reducing referrals to special education, and strengthening parent relationships to the school. (Darling-Hammond, Ancess & Falk, 1995, Ch. 6).

Each Record includes "observation notes" by the teacher of the student as a speaker and listener, reader, writer, and mathematician. It includes writing samples and reports on at least three books the student read and her understanding of them. Math covers four core domains. For each piece of documentation, the teacher describes the context of the observation or sample (e.g., individual or group work; whether the reading was literature or social studies, whether the book was new or familiar).

Teachers write extensive summary descriptions of the student's status and progress three-quarters of the way through the year. These are reviewed and discussed with parents. Toward the end of the year, the record is prepared for the coming year's teacher, adding any significant new information.

In reading and writing, each student's progress is placed numerically on a developmental scale. Separate reading scales cover grades K-3, 4-8 and 9-12. The scales are similar to many others that describe the process of learning to read and write. In K-3, teachers observe and document such things as students' use of letter-sound correspondence. They consider how children apply their prior life experience and how they build on their familiarity with the conventions of print as ways to understand what they read. Teachers use tools such as Running Records and Miscue Inventories as well as informal observations to determine which skills and strategies students are using and which ones they need to learn. In Grades 4-8, teachers look for signs that students are learning to read widely, for recreation as well as literature and informational text in all subject areas. In Grades 9-12, students provide evidence they can read critically across the curriculum.

Use of the LR scales have been validated (Hallam, 2001; Thomas, 2001). Moderation processes established good inter-rater agreement (above .75) between the classroom teachers and the independent re-scoring. This shows that with a good structure, diverse sources of information can be organized and evaluated to provide accurate, comparable evidence of student progress. Results can be aggregated and used to describe group attainment and progress. That is, if each originating teacher's judgment is solid, as supported by a review of 3 to 5 randomly sampled Records, then the aggregate information about classrooms and schools can be considered sound.

Hallem's (2001) validity study also compared LR scale scores with standardized test results. While the LR and norm-referenced tests such as the CTBS and Stanford Achievement Test are fundamentally different, they measure some of the same skills. Thus, the expectation is for positive but not very high correlations, which is what Hallam found in her investigation of 3009 individual LRs. Correlations ranged from .48 to .65. Hallam also found strong support among teachers for the LR's positive effects on professional development.

Practice and research in reading and writing demonstrate the LR is a reliable, valid, comparable and educationally sound method of evaluating individual progress and status using multiple sources of evidence, and aggregating that information to provide public information about schools. Math was a later development for the LR and thus the scales developed for the subject were not evaluated in depth. As with the LR in reading and writing, teachers provide varied specific evidence of learning that could be scored on LR with common rubrics, to provide reliable and valid data.

Resources and References

For more detailed information about the LR, see http://www.fairtest.org/learning-record. It includes links to articles, reports, sample recording forms, reading and writing scales, professional development, etc. For example, a sample recording form for young children is at http://www.fairtest.org/sites/default/files/LR-%20reporting%20form%20-%20Elementary_Eng%20(1).pdf. This page also links to further discussions of the validity and reliability of the LR.

FairTest thanks Mary Barr, Myra Barrs and Sally Thomas for their contributions to our knowledge, shared over many years.

Articles and Books

Barr, M., Craig, D., Syverson, M, and Fisette, D. 1999. Assessing Literacy with the Learning Record: A Handbook for Teachers, Grades K-6. Portsmouth, NH: Heinemann.

Barr, M., and Syverson, M. 1999. *Assessing Literacy with the Learning Record: A Handbook for Teachers, Grades 6-12.* Portsmouth, NH: Heinemann.

Darling-Hammond, L., Ancess, J., and Falk, B. 1995. *Authentic Assessment in Action.* New York: Teachers College Press. See esp. Ch. 5, "The Primary Language Record at P.S. 261.

Hallam, P.J. 2001. "Findings on Literacy Learning, Environment and Learning Record™ Validity, 2001 at Combined Learning Record™ Sites." Center for Language in Learning. http://www.fairtest.org/sites/default/files/LR%20validity%2099-01all%20scores.pdf.

Thomas, S. 2001. "Learning Record Shows Promise For Accountability Uses." *FairTest Examiner*, Fall. http://www.fairtest.org/learning-record-shows-promise-accountability-uses

Work Sampling System

The Work Sampling System (WSS) was developed by Samuel Meisels, one of the nation's foremost authorities on the assessment of young children, and his colleagues. Teachers observe children, collect and evaluate samples of student work in structured portfolios, use checklists, and prepare summary reports (shared with parents) three times per year. Originally for pre-school through grade 3, it now goes extends through grade 6. WSS has been demonstrated to have strong validity and reliability as well as parental acceptance.

The system is based on seven domains or categories, each with performance indicators:

- Personal and Social Development focuses on self-identity, the self as a learner, and social development.
- Language and Literacy is based on the theory that students learn to read and write the way they learn to speak, naturally and slowly.
- Mathematical Thinking focuses on children's approaches to mathematical thinking and problem solving.
- Scientific Thinking emphasizes the processes of scientific investigation, because process skills are embedded in and fundamental to all science instruction and content.
- Social Studies includes understanding from personal experience and learning about the experiences of others.
- Arts focus on how using and appreciating the arts enables children to demonstrate what they know and to expand their thinking.
- Physical Development includes developing fine and gross motor skills and competence to understand and manage personal health and safety.

Pearson now owns the WSS. Some users charge that the "grade-specific guidelines" for student achievement now based on the Common Core State Standards are developmentally inappropriate. However, they add, the tools for observing, describing and summarizing remain useful. Because checklists produce numbers, some observers fear that the numbers would be too seductive. That could undermine WSS' observational and narrative uses as well as parent interpretations. Investigations, including by FairTest staff, found that experienced teachers wrote highly individualized summaries of each child. Each seemed to be a clear snapshot, not a generic summary of attainments. WSS thus seems to effectively combine personalization and systematization, though the question of parental interpretation remains open.

In studies conducted when WSS was owned by Dr. Meisels, researchers found strong evidence of validity, reliability and parental approval (FairTest *Examiner*). One compared WSS reports by 17 teachers with their students' results on the individually administered Woodcock-Johnson (WJ) battery for literacy and math. They found correlations between WSS and the WJ comprehensive scores in reading, writing and math to range mostly from .50 and .75. These are high enough to support a claim of teacher assessment accuracy and not so high as to suggest that the WSS only measures what the WJ measures. The WSS also was found to be an accurate means of identifying children in need of special services, based on correlations with the WJ.

A second study found that parents in Pittsburgh viewed WSS positively. They believed its use benefited their children whether they were high or low achievers. Most of the families were low-income African Americans. The more parents knew about WSS, the greater their satisfaction. The researchers concluded, “[T]his study demonstrates that when schools using a systematic, curriculum-embedded performance assessment make an effort to keep parents informed about the assessment, and when consistent informal communications between parents and teachers takes place, parental reactions to performance assessment can be very positive.”

Another study compared children who had been enrolled in WSS with a demographically matched sample of children who had not experienced WSS and with all other children in the district. The WSS students showed substantially higher gains over time on conventional, group-administered achievement tests.

If the WSS is accurate and parents respect it, could its results, such as teacher-generated summary reports, be aggregated for use in reporting student learning at the school level? The WSS was not designed for this purpose. Prof. Meisels expressed concern about the potentially corrupting effects on teacher judgments if they know the results could be used to evaluate schools or make high-stakes decisions about students.

The accuracy of this curriculum-embedded assessment adds to the evidence that it is possible to construct public reporting systems based substantially on teacher judgment. If “accountability” consequences are genuinely helpful rather than punitive, the legitimate concerns about corrupting effects could be resolved.

Resources and References

The WSS is now owned by Pearson; information on it is available at http://www.pearsonclinical.com/childhood/products/100000755/the-work-sampling-system-5th-edition.html.

FairTest *Examiner* articles on WSS are available at http://www.fairtest.org/work-sampling-system and http://www.fairtest.org/trusting-teacher-judgment.

Big Picture Learning

Big Picture Learning (BPL) is a network of 65 schools across the U.S., with as many in a handful other countries. While a few of BPL's U.S. members are charters, most are regular public schools, and most are high schools. What distinguishes them are factors such as:

- grouping students in "advisories" of no more than 15, which provides a sort of home group for students throughout their school years;
- engaging in out-of-school internships. High school students are often at internships two days a week, learning with a mentor from the local community;
- project-based learning that bridges the real world of internships with academic learning that mixes teacher- and student-led studies, which enable pupils to build on their interests; and
- performance assessments, including student presentations and discussions of their work ("mini-portfolios") several times per year.

Big Picture Learning students with advisor. Photo from Big Picture Learning.

The performance assessments are not standardized across BPL, nor do they use a common scoring guide. Their students are subject to federal and state testing requirements, which have in some cases undermined portfolio and performance practices.

One school reporting harmful consequences is the founding BPL school, Met High School in Providence, R.I. Met leaders told FairTest that preparing for state exams, particularly the math test, took time from further developing common assessment practices in the school. Fortunately, Rhode Island has dropped its pending high-stakes high school exit exam, which should alleviate pressure. The Met seeks to return to developing its own assessments.

Met students produce extended biographies, often 75-100 pages, and senior thesis projects, as well as complete internships and coursework. Many have earned college credits, as is true across BPL schools. Met plans a 2016-17 pilot in which students, in collaboration with teachers, post their portfolios to a website, enabling more shared study of student products. The school has its own rubrics, crafted by the teachers, for evaluating projects and internships. Met curriculum director Joe Battaglia said they are used first for planning and guidance and second for assessing student work. Met and other BPL schools are also grappling with how to document, assess and give credit for the large amounts of learning beyond what is initially planned when internships and courses are designed.

Met forms its own district, drawing students from across the state, with about 70% qualifying for free- and reduced-price lunch. Thirty percent of their students enter with lower than grade 4 reading levels. A detailed investigation of RI high schools found the Met at the top in 80 out of 81 categories, reported Dennis Littky, founding co-director of the Met and BPL.

A study of BPL graduates from across the network reported strong results. A 2015 report concludes, "Findings show that the Big Picture Learning model is highly effective at fostering positive relationships, helping students discover and pursue their interests, and raising high school graduation and college entrance rates" (Arnold, *et al*., 2015). The schools averaged a 92% graduation rate, of whom over 90% started at college or other post-secondary institutions. Of those, from the classes of 2006 and 2007, more than two thirds had earned a Bachelor's (35% and 24%), Associate's or other credential, or were still enrolled in 2012. This compares well with the national average, particularly of students from lower-income strata. (For example, the four-year college graduation rate of the lowest income quintile is 10.6%.)

Most of BPL students "come from communities with high levels of academic under-achievement, geographic transition, and high school dropouts." Much of BPL's success is due to the deep relationships built in the schools, especially between advisors and students, and the resulting strong sense of school community that sustains the youth. Most remain in some contact with advisors after graduating. Parental engagement in BPL schools also is far higher than in most U.S. high schools. However, the report concluded that many students' preparation for college work in math and science was weak, an issue BPL is grappling with.

Resources and References

Arnold, K.D., Soto, E.B., Wartman, K.L., Methven, L., and Brown, P.G. 2015. *Post-secondary Outcomes of Innovative High Schools: The Big Picture Longitudinal Study*. Boston College, March 3. Available at https://d3jc3ahdjad7x7.cloudfront.net/9hloszW4FyNM5EdJWri39BVKbVpArurU9gAFe3FmKmcuICyK.pdf

Big Picture Learning website, http://www.bigpicture.org/

FairTest visited Big Picture Learning on May 10, 2015, talking with co-founder Dennis Littky, Joe Battaglia and other staff. FairTest also talked with BPL co-founder Elliot Washor.

International Baccalaureate

> "The International Baccalaureate® aims to develop inquiring, knowledgeable and caring young people who help to create a better and more peaceful world through intercultural understanding and respect."

Founded in 1968, International Baccalaureate (IB) is a worldwide system of thousands of schools known for strong academics and in-depth learning to prepare students for college. IB contains four programs: Primary Years, Middle Years (ages 11-16), Diploma (final two years of high school), and Career-related, which includes some Diploma courses and additional work aimed at career preparation. IB periodically reviews its member schools to re-approve them. Slightly more than half of IB Diploma candidates are from the U.S. (76,000).

IB assessments vary significantly by program level. In the Primary program, there is formative but no external, summative assessment. All students complete a personal project at the end of the program.

The Middle program assessments are primarily local. Students complete a project of personal interest to them in each of the final two Middle years. Typically, these are presented in a "showcase." Until 2015-16, IB provided voluntary moderation (re-scoring) for samples of student work submitted by participating schools. That year, IB initiated an optional exam on computer. Schools now choose between the portfolio and the test. Only a few have signed up for the test at this point, including some who see it as necessary to meet government requirements.

The Diploma program requires students to pass courses in six areas: literature, foreign language, humanities (which includes history), science, math and a student option, such as art. These are offered at "Standard" and "Higher" levels, with students required to complete higher level work in at least three areas.

Students take a combination of internal and external exams in each subject. For art, IB uses portfolios. The external tests, taken on paper and marked on a 1-7 scale, carry more weight, typically around 80%. They are centrally scored by international teams of teachers and retired teachers. This requires IB to address language and cultural differences among students and graders. Matthew Glanville, head of assessment, says it is also difficult to score consistently across subject areas, but they are generally satisfied with their accuracy.

In 2015, IB registered 580,000 students to take exams and granted about 56,000 diplomas. The exams are primarily responses to essay prompts, but include short and longer structured problems. Multiple-choice items are rarely used because IB expects students to demonstrate deeper knowledge and higher order thinking.

Each Diploma student also must pass a Theory of Knowledge course which includes an extended, research-based essay the student chooses with teacher support. The essays are graded by IB rather than the school. Glanville says that students often complain about this project, then report it was their most valuable preparation for college.

IB staff emphasize the importance of teacher assessing. It provides guides for teachers to use in developing and marking tests, projects and other classwork.

Boston's public Josiah Quincy Upper School offers the Middle and Diploma IB programs, covering grades 6-10 and 11-12. The student body is 92% low income, 50% Asian, 30% Black, 15% Hispanic and 5% White; its location on the edge of Chinatown explains in part the disproportionate Asian participation. Twenty-two percent have disabilities and most of them are fully included in the IB program. Not all students who graduate complete the IB program, but two-thirds of the Upper students are on track to do so. Quincy joined the Middle program exam, but only for grade 10, the final Middle year. The school also participates in voluntary moderation activities in grades 6-10.

Josiah Quincy staff pointed out that the Diploma program focuses heavily on preparing for the exams, while the Middle program is more holistic. Even with the exam focus, an internal survey found students prioritize learning for its own sake over grades, and they prize self-direction and taking ownership of their learning. Staff also said that moving toward a portfolio approach might be a good idea.

In conclusion, IB assessment, particularly in the Diploma program, provides limited student choice, though students say they greatly value those options. IB shows that a wide range of exams can be fairly scored. Their exams can do a far better job of evaluating in-depth knowledge than current tests common in the U.S. In a "system of systems" that states could establish under ESSA, IB or similar approaches are a valid option. The question states will face is how to establish comparability between IB results, those of current state tests, and other locally determined assessments.

Resources and References

International Baccalaureate website, http://www.ibo.org/.

Interview with Matthew Glanville, Head of Assessment Principles and Practice, June 14, 2016.

Visit to Josiah Quincy Upper School, Boston, MA, June 23, 2016. Discussions with Richard Chang, co-principal; Robin Coyne, Guidance Counselor; and Sarah Chang and Kristina Danahy, IB co-Coordinators.

Additional Resources on Performance Assessments

These are in addition to references included in each section.

FairTest fact sheets:

A Better Way to Evaluate Schools. 2010. http://fairtest.org/fact-sheet-better-way-evaluate-schools-pdf

Multiple Measures: A Definition and Examples from the U.S. and Other Nations. http://www.fairtest.org/fact-sheet-multiple-measures-definition-and-exampl.

Books and Reports:

Annotated Bibliography: Performance Assessment. N.D. FairTest. http://www.fairtest.org/annotated-bibliography-performance-assessment. (These books and reports are primarily from the 1990s; many remain very useful.)

Darling-Hammond, L. 2010. *The Flat World and Education*. New York: Teachers College Press.

Meier, D. 2002. *The Power of Their Ideas.* Boston: Beacon Press.

Wood, G., Darling-Hammond, L., Neill, M., Roschewski, P. 2007, May. Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills. http://www.fairtest.org/refocusing-accountability-using-local-performance-

Videos:

Beyond Measure. http://beyondmeasurefilm.com/. By Vicki Abeles, with a companion book.

FairTest: You Can't Judge Learning with a Standardized Test. Four-minute video available at https://www.youtube.com/watch?v=WkJlst6vDyY.

Good Morning Mission Hill. http://goodmorningmissionhill.com. Tom and Amy Valens. See also a series of shorts based on the film’s footage, *A Year at Mission Hill.* http://www.ayearatmissionhill.com. Video #9, “Seeing the Learning,” focuses on assessment.

Most Likely to Succeed. http://www.mltsfilm.org/. Produced by Ted Dintersmith, Greg Whitely and Adam Leibowitz.

Schools That Change Communities. http://www.docmakeronline.com/schoolsthatchangecommunities.html. Bob Gliner.

Part IV

Principles and Uses of Assessment: “To Assess” Means “To Sit Beside”

The word “assess” derives from the Latin term meaning “to sit beside.” Assessing implies a direct and active relationship between or among people. Assessing could involve an observation by a teacher, a conversation between teacher and student, or looking at a student’s academic work. It typically includes interaction to provide feedback or find out more about the student’s thinking or depth of knowledge.

In this document, we often use the term “assessing” to mean the process rather than the instrument or system (“assessment”). In schools, assessment is a relationship that revolves around teaching and learning. Thus, it must be rooted in the content and skills students should learn.

The primary purpose of assessing should be to improve the depth and breadth of student learning, including their ability to learn. Other applications can be built on that foundation. These could include ascertaining whether students have met a threshold of knowledge and skills needed for high school graduation, or evaluating a school or district. These secondary decisions can be derived from the primary assessment evidence gathered by educators. Other forms of evidence, such as external or large-scale exams and measures of school or district quality, might be considered.

“Test” and “exam” are largely interchangeable, but they do not mean the same thing as “assessment.” Tests are tools that are used in a direct relationship, as when a teacher administers a test to her students. Or districts or states could place tests between teachers and student. In the latter case, the *Standards for Educational and Psychological Testing* say that such tests “standardize the process by which test takers’ responses to test materials are evaluated and scored” (AERA, APA, NCME, 2014, p. 2).

Tests are commonly used by themselves or in conjunction with other kinds of evidence to judge students, educators or schools. Using them on a stand-alone basis – as a mandatory hurdle – to make high-stakes judgments, generally violates the *Standards* (e.g., Standard 12.10, p. 198, for students; or 13.9, p. 213, for programs or schools). Evaluation of students, educators and schools can and should include more than academic achievement – as ESSA now requires, at least minimally, of schools.

This chapter first considers three primary assessment purposes, then proposes a set of principles to guide assessing.

Assessing for, of and as learning

Assessing is a relationship and a process that employs various tools to support and evaluate student learning. Experts increasingly describe assessment as *for learning*, *of learning*, and *as learning*.

Assessing for learning, or "formative assessment," is used to provide feedback to students or be used by the teacher to modify instruction to improve student learning (Third International Conference, 2009). Formative assessment can range from a teacher observation or conversation with a student to a multiple-choice quiz to probing questions when a student is engaged in an extended project. It must provide meaningful, useful feedback. In order to fit the curriculum and provide actionable information, it should be primarily controlled by the teacher. Students also can assess one another and self-assess. Practices such as the "descriptive review of the child" also serve formative purposes as teachers closely appraise a child as a learner and a whole person in order to better understand and serve her (Himley & Carini, 2000).

Teachers often administer mini-tests created outside the classroom. Commonly, these are mandated by local administrators. Many are now taken on a computer. Their purpose is to determine how well a student has advanced in a pre-set curriculum or toward higher scores on standardized tests. Use of such instruments does not represent good formative assessment. Those are interim tests (also called benchmark, periodic or predictive).

Assessment of learning, or summative assessment, evaluates a student's learning or attainments. Grades and most standardized tests serve this purpose. Standardized tests have often been misused to determine graduation, promotion, or program placement. Culminating projects or evaluations based on portfolios serve this purpose as well. Such assessments can also play a formative role in improving teaching year to year, strengthening school curriculum, and improving how districts use resources. Summative assessments should be rooted in ongoing student work. External exams should serve only as a check on the system. They should not be a primary measure of student progress or a requirement for placement, advancement, graduation, or college admissions.

Assessment as learning emerged as a concept in the 1990s as educators sought to craft projects, tasks or exams in which the effort required to do well provided instructional benefits. Well-designed projects are a much more powerful practice than exams because they provide depth of learning and student engagement. For example, a student could develop a project, with teachers providing feedback along the way. The final product is evaluated, and the students reflect on the process and their learning. Such evidence of learning can also become part of evaluating schools.

Big Picture Learning student presenting. Photo from Big Picture Learning.

Clearly there can be many forms of assessment "for, of and as" learning. Indeed, one criterion for a good assessment system is the use of "multiple measures." That means students have the opportunity to demonstrate their learning in varied formats over time. This, in turn, creates different ways in which teachers can assess to help students learn. At times, the term has been reduced to "several standardized tests" or "mostly tests with a bit of something else." Those misuses undermine educational quality and damage students (Neill, Guisbond & Schaeffer, 2001).

The related term "classroom-based" refers to assessing that is integrated with the particular curriculum and work students do. It includes "for, of and as." Classroom-based assessing is under the control of teachers and primarily constructed by them – "practitioner developed," in the words of the New York Performance Standards Consortium (NY Consortium, N.D.).

As with all such definitions, lines blur. Teachers may share assessments or borrow from a library of tasks assembled from many classrooms. The key point is that teachers have control. Students deserve a strong voice in the work they do and should also learn to self-assess. The NY Consortium (N.D.) often refers to their Performance-Based Assessment Tasks as "school-based," since the shaping of these tasks and the scoring guides used are set by schools as collectives of teachers/practitioners.

Principles for Assessment

States and districts are rethinking assessment in light of the opportunities provided in ESSA, such as its innovative assessment program. As they do so, they should be guided by these core principles. In drafting them, FairTest has relied on previous work, especially the *Principles and Indicators for Student Assessment Systems* (National Forum on Assessment, 1995). The primary differences are that we now more strongly emphasize student and teacher agency, and we have compiled even more evidence of the damage centralized control, especially via standardized tests, can wreak on school quality.

We are mindful of the strong advice in the report of the Forum on Educational Accountability (FEA, 2007). It states that the foundational requirements for high-quality assessment are that "all students have equitable access to the resources, tools, and information they need to succeed and by building capacity to improve teaching and learning" (p 2). With that in mind, FairTest offers these Principles:

1. ***The primary purpose of assessing is to enhance learning***. Formative assessing, including assessment as learning, and classroom-based practices should have top priority. It involves documenting student progress, allowing students multiple methods to demonstrate their learning over time, and providing usable feedback to students. Using assessment evidence for summative decision-making or reporting about students or systems is secondary and must not undermine the primary purpose.
2. ***Assessing is rooted in significant learning outcomes and is therefore integrated with standards and curricula.*** As explained by the National Forum on Assessment (1995), "Learning goals or content standards describe broad, important intellectual competencies – knowledge, skills, understandings and habits of mind – that students should acquire and be able to demonstrate. These include important learning in and across subject areas, with a focus on thoughtful application and meaningful use of knowledge" (p. 5). Assessing focuses heavily on critical thinking, problem solving, research, writing, public speaking, presentations, initiative, self-development and group collaboration. In doing so, it pays attention to building blocks of declarative and procedural knowledge.
3. ***Assessments are primarily practitioner developed and controlled.*** Classroom educators must have the primary authority and responsibility to ensure high-quality assessing as they determine curriculum and engage directly with students. Professionals must know and use strong assessment practices.
4. ***Assessing is student focused so that they exercise significant control where appropriate,*** such as choice of books to read, research topics and projects. James Coleman (1966) found that while family background was the primary predictor of student outcomes, the second most powerful predictor was students' sense of control over their learning. In college admissions, this means students should decide whether to include SAT, ACT or similar test results in their applications.
5. ***Assessing should be fair, unbiased and culturally responsive for use in an increasingly diverse society.*** This means ensuring teachers are culturally competent and can recognize and counter biases they may have individually or as a group. Instruments such as standardized tests or scoring guides used for projects and portfolios are carefully vetted to eliminate cultural bias. In addition, assessing by teachers and instruments they use must be well constructed to best instruct students with disabilities and English language learners (ELL) and allow them to demonstrate their progress. Universal design is one key to that process. (Oregon, 2015, principle 2; FEA, 2007, Principle III.)
6. ***An assessment system uses tasks, projects, portfolios and learning records as core tools.*** Teachers may find a test is useful as a supplementary tool. On a large scale, state tests could be used as a check on the system but should not themselves carry decision-making weight.

7. ***An assessment system is as decentralized as possible.*** States build toward "systems of systems" comprised of local assessments. School and system evaluations should primarily use classroom/school-based evidence. State- or district-mandated exams are used sparingly, if at all. As required under the federal Every Student Succeeds Act, a state must establish adequate comparability across local systems. States are also responsible for ensuring local systems are valid, reliable and used fairly.
8. ***Professional collaboration and development are necessary to ensure high-quality assessing.*** A good deal of research finds that teachers often have too narrow a repertoire of assessment practices (Stiggins, 2014). This problem became much worse under NCLB's focus on standardized tests with the ensuing proliferation of interim exams and falsely labeled formative instruments. However, much evidence demonstrates teachers prefer to use a rich array of practices, and they can collaboratively learn and improve (Gallagher, 2007).
9. ***Communities participate in developing assessments.*** As schools, districts and states overhaul their practices, they must engage communities. This starts with a shared process of defining the purposes of education, its desired goals and outcomes, then considering how to evaluate the system. It includes participating in system reviews. As assessments should support teachers and students in reaching the goals, communities should understand and contribute to how assessment can do so. For example, parents can share their knowledge about how their child learns, her interests and needs. Students and parents can share their views on the benefits and drawbacks of particular assessment approaches. (See NFA, Principle 5.)
10. ***Systems incorporate feedback loops to ensure continuing improvement.*** This includes reviewing professional learning programs, examining goals as well as tools such as scoring guides, and periodically revisiting how assessment can best strengthen teaching and learning.

Taken together, these principles direct states to build systems of local assessments that are practitioner-controlled and student-focused. The new systems rely primarily on performance assessments that are classroom- and school-based. They must be unbiased and culturally responsive. They also must gather diverse forms of evidence of student learning over time. Such systems minimize statewide exams and the use of multiple-choice and short answer questions, including computer-based tests. They require professional learning and collaboration and well as meaningful community participation. And they are rooted in expectations of rich student learning in a strong, deep, culturally responsive curriculum in well-supported schools.

Resources and References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: 2014.

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.

Forum on Educational Accountability. 2007, August. *Assessment and Accountability for Improving Schools and Learning*. http://www.edaccountability.org/AssessmentFullReportJUNE07.pdf

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

Himley, M., and Carini, P. 2000. *From Another Angle*. New York: Teachers College Press.

National Forum on Assessment. 1995. *Principles and Indicators for Student Assessment Systems.* http://www.fairtest.org/principles-and-indicators-student-assessment-syste

Neill, M., Guisbond, L., and Schaeffer, B. 2001. *Failing Our Children*. FairTest. http://www.fairtest.org/node/1778

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

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

Stiggins, R. 2014. *Revolutionize Assessment.* Thousand Oaks, CA: Corwin Press.

Third International Conference on Assessment for Learning, Dunedin, New Zealand. 2009. *Position Paper on Assessment for Learning.* http://www.fairtest.org/position-paper-assessment-learning