



Assessment for the 21st Century: Using Performance Assessments to Measure Student Learning More Effectively

Linda Darling-Hammond

Charles E. Ducommun Professor of Education, Stanford University
Co-Director, School Redesign Network

George H. Wood

Executive Director, The Forum for Education and Democracy
Principal, Federal Hocking High School, Stewart, Ohio

October 20, 2008

For more information, contact:
Beth Glenn
202-372-7684

www.forumforeducation.org

Conveners:

Judith Browne-Dianis
Linda Darling-Hammond
Carl Glickman
John Goodlad
Gloria Ladson-Billings
Deborah Meier
Larry Myatt
Pedro Noguera
Wendy Puriefoy
Sharon Robinson
Ted and Nancy Sizer
Angela Valenzuela
George Wood, Executive Director

Executive Summary

*This paper is a modified version of a previously published paper, **Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills**. To access the original version, visit www.fairtest.org/refocusing-accountability.*

Assessments focused on higher order thinking and performance skills are essential for guiding teaching and learning that can really prepare students for 21st century careers and college. Extensive research and experience, both in the U.S. and abroad, have demonstrated that the use of locally-administered performance assessments helps focus education systems on their primary purpose — **assisting teachers in teaching and students in learning the problem-solving and performance skills they need in the real world**. In fact, the assessment systems of most of the highest-achieving nations in the world are a combination of centralized assessments that use mostly open-ended and essay questions with local assessments given by teachers that are factored into the final examination scores.

Unfortunately, while federal legislation in the U.S. calls for “multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding” (NCLB, Sec. 1111, b, 2, I, vi), most assessment tools used for reporting under No Child Left Behind focus on lower-level skills that are measured by standardized, mostly multiple-choice tests. High stakes attached to these tests have led many schools — especially those serving lower-achieving children — to neglect more challenging and engaging curricula and to limit school experiences to those that focus on a limited form of test preparation.

Performance assessments engage students in the demonstration of skills and knowledge through the completion of complex tasks that provide teachers with an understanding of how students learn as well as what they know. Large scale examples of the use of such performance-based assessments come from states such as Connecticut, Kentucky, Maine, New York, Vermont, and Wyoming. The evidence from research on these and other systems indicate that the use of these assessments focuses instruction on higher order skills, provides a more accurate measure of what students know and can do, engages students more deeply in learning, and can provide for more timely feedback to teachers, parents, and students that improves instruction. Furthermore, research has found that properly constructed performance assessment systems can be both reliably scored and more valid in predicting later life success.

While high quality assessment requires investments in teacher development and the development and scoring of performance tasks, this investment strengthens teachers’ understanding of standards and how to meet them, while setting more challenging expectations for what students should be able to do with what they learn. Thus these investments not only provide information about student achievement; they also improve teaching quality and student learning.

To meet our goals for education, federal and state policies should support and encourage the development and use of performance assessments at both the state and local levels. These can be combined in a systematic way to evaluate the full range of standards and to support more appropriate judgments about what students can do and what schools have contributed to their learning.

Refocusing Accountability: Using Performance Assessments to Enhance Teaching and Learning for Higher Order Skills

Over the past decade, educators, policymakers, and the public have begun to forge a consensus that our public schools must focus on better preparing all children for the demands of public participation and work in the 21st century. This has resulted in states developing standards-based educational systems and assessing the success of districts and schools in meeting these standards by requiring more systematic testing. Most of these tests are multiple choice, “on demand” standardized measures of achievement. This has limited what is tested to a narrow range of knowledge and skills. Studies have documented a number of unintended consequences of these assessments, especially when they are “high stakes” and strong consequences are attached to scores. These include a narrowing of the academic curriculum and experiences of students, especially in schools serving lower-achieving children; a focus on recognizing correct answers to lower-level questions rather than on developing higher-order thinking, reasoning, and performance skills; and a growing dissatisfaction among parents and educators with the school experience (for a review, see Darling-Hammond & Rustique-Forrester, 2005; von Zastrow and Janc, 2004).

The United States has also been falling further behind on international assessments of student learning. On the Program in International Student Assessment (PISA) tests in 2006, the U.S. ranked 25th of 30 OECD countries in mathematics and 21st in science, a decline in both raw scores and rankings from 3 years earlier. (Reading scores were not reported because of editing problems with the U.S. test.) In each disciplinary area tested, U.S. students scored lowest on the problem-solving items.

Because the PISA assessments focus explicitly on 21st century skills, they go beyond the question posed by most U.S. standardized tests — “Did students learn what we taught them?” — to ask, “What can students do with what they have learned?” (Stage, 2005). PISA defines literacy in mathematics, science, and reading as students’ abilities to *apply* what they know to new problems and situations. This is the kind of higher-order learning that is increasingly encouraged by reformers in the U.S. (see, e.g., Partnership for 21st Century Skills, 2007) and emphasized in other nations’ assessment systems, but often discouraged by the multiple-choice tests most U.S. states have adopted. In addition to problems with uneven teaching quality, analysts point to the fact that U.S. curriculum often requires coverage of more topics at a more superficial level of understanding and less emphasis on deep understanding through application and defense of ideas, extensive writing, critique, and reasoning (Schmidt, Wang, & McKnight, 2005).

These unfortunate consequences have occurred despite language in current federal legislation calling for “multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding” (NCLB, Sec. 1111, b, I, vi). Changing what counts as assessment evidence, coupled with other changes in the federal accountability structure measuring adequate yearly progress and sanctions, could help to overcome these problems and contribute toward school improvement.

Performance Assessment: A Definition

Almost every adult in the United States has experienced at least one performance assessment: the driving test that places new drivers into an automobile with a DMV official for a spin around the block. Few of us would be comfortable handing out licenses to people who have only passed the multiple-choice written test required by the state. The value of the performance assessment is that it illustrates — in ways the paper test cannot — essential information about the driver's skills, competency, and level of understanding.

Performance assessments in education are similar. They are tools that allow teachers to gather information about what students can actually do with what they are learning — science experiments that students design, carry out, analyze, and summarize; computer programs that students create, test and refine; persuasive essays that students write; research inquiries they pursue, seeking and assembling evidence about a question, and presenting it in written and oral form. Whether the skill or standard being measured is writing, speaking, mathematical literacy, or research, students actually perform tasks involving these skills while the teacher observes, gathers information, and then scores the performance based upon a set of pre-determined criteria.

As in our driving test example, these assessments typically consist of three parts; a task, a scoring guide or rubric, and a set of administration guidelines. The development, administration, and scoring of these tasks requires teacher development to insure quality and consistency. The research suggests that such assessments are better tools for showing the extent to which students have developed higher order thinking skills, such as the abilities to analyze, synthesize, and evaluate information. They lead to more student engagement in learning and stronger performance on the kinds of authentic tasks that better resemble what they will need to do in the world outside of school. They also provide richer feedback to teachers, leading to improved learning outcomes for students (Darling-Hammond & Rustique Forrester, 2005; Falk, 2000).

Local Assessments

Extensive research and experience, both in the United States and abroad, have demonstrated that the use of *performance assessments* which are *locally administered* and use *multiple* sources of evidence offer the opportunity for assessment systems to serve their primary purpose — **assisting students in learning and teachers in teaching for higher order intellectual skills** (Darling-Hammond & Rustique Forrester, 2005; Falk, 2000). In fact, the assessment systems of most of the highest-achieving nations in the world are a combination of centralized assessments that use mostly open-ended and essay questions and local assessments given by teachers which are factored into the final examination scores (Darling-Hammond & McCloskey, in press; Eckstein & Noah, 1993). These local assessments — which include research papers, applied science experiments, presentations of various kinds, and projects and products that students construct — are mapped to the syllabus and the standards for the subject and are selected because they represent critical skills, topics, and concepts. Central authorities often determine curricular areas and skills to assess, but the assessments are generally designed, administered, and scored locally.

The *local management* of such assessments refers both to their use and scoring. While not all performance assessments are locally developed, decisions about when to use them in the learning process and how to adapt them to particular content are made at the school or classroom level. This is a vital point — myriad studies confirm that, to be useful for learning, assessments must be responsive to emerging student needs and enable fast and specific teacher response, something that standardized examinations with long lapses between administration and results cannot do (Black & Wiliam, 1998).

In addition, as teachers use and evaluate these tasks, they become more knowledgeable about the standards and how to teach to them, and about what their students' learning needs are. The process improves their teaching. These rich assessment tasks can also be utilized as formative or benchmark assessments, which help teachers gauge ongoing progress, while avoiding the reduction of such assessments to multiple-choice formats alone.

Using *multiple sources of evidence* refers to the way in which performance assessments provide multiple ways to view student learning. For example, multiple samples of actual writing taken over time can best reveal to a teacher the progress a student is making in the development of his or her composition skills. This growth can then be reflected in a summative evaluation of the student's learning at important benchmark points. This provides ongoing feedback to learners as well, as they see how they are developing as writers and what they have yet to master. In addition, different kinds of writing tasks — persuasive essays, research papers, journalistic reports, responses to literature, etc. — encourage students to develop the full range of their writing and thinking skills in ways that writing a five-paragraph essay over and over again do not.

Of course, the most important benefit to utilizing performance assessments is that they assist in learning and teaching. They are *formative* in that they provide teachers and students with the feedback they need from authentic tasks to see if they have actually mastered content. They are also *summative* in that they can serve as a final assessment of student capabilities with respect to state and local standards, and as measures of their growth over time. Because of the fact that they are embedded in the curriculum and ask students to demonstrate what they have learned in an open-ended way that reveals student thinking, performance assessments are more sensitive to instruction and more useful for teaching than standardized examinations, while providing richer evidence of student learning that can be used by those outside the classroom or school.

These features of performance — local administration, multiple sources of evidence, and transparency of standards — are used in many assessment systems. Think back to the state driver's license exam, which involves both a written test and a performance assessment on the road. Everyone knows precisely what to expect in terms of the skills to be demonstrated — for example, whether or not the applicant can parallel park — as the examination is not a total secret. The fact that the assessment is open and transparent is not a problem, because the point is to see whether drivers have developed these real-world abilities. The performance is scored by the instructor, working from a rubric, and if the driver is sufficiently successful in all aspects of the examination (as determined by a state cut-off score), a license is conferred. The task is so well defined that instructional programs (driver's education), which include both hands-on and classroom instruction, clearly demonstrate their effectiveness in preparing students to perform. (This is reflected in the reduced insurance rates we grant to graduates of driver's education programs.) Imagine what life on our roads would

be like if we did not require prospective drivers to demonstrate what they know before taking the wheel.

Some states, districts, and schools have constructed a similarly rich set of assessments of competence that measure the higher-order thinking called for by new standards. In many cases they are explicitly intended to augment and complement more traditional tests. The New York Performance Standards Consortium is a network of 39 schools (located mainly in New York City) that rely primarily upon performance assessments to determine student readiness for graduation. Because of the quality of their work, these schools have a state waiver from four out of five required Regents exams. Students must complete a portfolio that includes a science investigation, a social science research paper, a literary critique, and a mathematical model or complex problem solution rooted in a real world challenge. All schools in the consortium use the same rubrics for scoring, and schools work together to set and maintain common standards. Research from this work indicates that these schools' students — largely low-income students of color entering high school with substandard skills — not only graduate high school at a much higher rate than the New York City average, but also attend college at higher rates, maintain strong GPAs, and persist in their college studies at rates higher than the national average (Foote, 2007; Fine, Stroudt & Futch, 2005).

Illinois' assessments provide a good example of how state tests and classroom performance assessments can complement each other, as well as how they measure different abilities. The state's 8th grade science learning standard 11B reads: "Technological design: Assess given test results on a prototype; analyze data and rebuild and retest prototype as necessary." The multiple choice example on the state test simply asks what "Josh" should do if his first prototype boat sinks, with the desired answer, "Change the design and retest his boat." By contrast, the classroom assessment says: "Given some clay, a drinking straw, and paper, design a sailboat that will sail across a small body of water. Students can test and retest their designs." In the course of this activity, students explore significant physics questions such as displacement in order to understand how a ball of clay can be made to float.

Such activities combine hands-on inquiry with reasoning skills; have visible, real-world applications; are more engaging; and enable deeper learning. They also enable the teacher to assess student learning along multiple dimensions, including the ability to frame a problem, develop hypotheses, reflect on outcomes and make reasoned and effective changes, demonstrate scientific understanding, use scientific terminology and facts, persist in problem solving, organize information, and develop sound concepts regarding the scientific principles in use.

Large Scale Examples of Performance Assessment

As we have noted, it is possible to create and implement assessment systems that include multiple sources of evidence which are performance-based and locally managed. In addition, many states — including Connecticut, New York, and Vermont — have developed and use performance assessments as part of their state testing systems. Indeed, the National Science Foundation provided millions of dollars for states to develop such hands-on science and math assessments as part of its Systemic Science Initiative in the 1990s, and prototypes exist all over the country. A number of states and countries provide models for how performance-based assessment systems can engage teachers, parents, and students in thinking carefully about what students have learned and how to measure that learning. Here are some examples:

- Connecticut uses rich science tasks as part of its statewide assessment system. For example, students design and conduct science experiments on specific topics, analyze the data, and report their results to prove their ability to engage in science reasoning. They also critique experiments and evaluate the soundness of findings.
- Maine, Vermont, New Hampshire, and Rhode Island have all developed systems that combine a jointly constructed reference exam with locally developed assessments that provide evidence of student work from performance tasks and portfolios.
- Kentucky developed portfolios in writing and mathematics to complement its state tests. The writing portfolio is still in use statewide; the mathematics portfolio was made optional for districts.
- Nebraska developed a system of assessments that are created and scored by local educators. Systems of local assessments mapped to state standards are peer-reviewed by assessment experts and include a check on validity through the use of a state-wide writing examination and the administration of a norm-referenced test.
- Wyoming uses a “body of evidence” approach that is locally developed in order to determine whether students have mastered standards required for graduation.
- In Silicon Valley, CA, many school districts use the Mathematics Assessment Resource System (MARS), an internationally developed program which requires students to learn complex knowledge and skills to do well on a set of performance-based tasks. The evidence shows that students do as well on traditional tests as peers who are not in the MARS program, while MARS students do far better at solving complex problems.

There are many large scale international examples of performance assessment, particularly from nations that lead in academic achievement. The following examples come from Australia and Sweden, illustrating the feasibility of large-scale nationwide use of performance assessment (from Darling-Hammond and McCloskey, in press).

Australia

Each Australian state has its own curriculum and assessment program. At the national level in Australia, the only assessment is a periodic matrix sample-based assessment, rather like the NAEP in the U.S. In most states, local school-based performance assessment is a well-developed part of the system. In some cases, states have also developed centralized assessment with performance components. The highest-achieving state, Queensland, has the most highly developed systems of local performance assessment.

In Queensland, there has been no assessment system external to schools for 40 years. Until the early 1970s, a traditional “post-colonial” examination system controlled the curriculum. When it was eliminated, all assessments became school-based. School-based assessments are developed, administered and scored by teachers in relation to the national curriculum guidelines and state syllabi (also developed by teachers), and are moderated by panels that include teachers from other schools as well as at least one professor from the tertiary education system.

The syllabi are not highly detailed; they spell out a small number of key concepts and/or skills to be learned in each course and provide examples of the kinds of projects or activities (including minimum assessment requirements) in which students should be engaged. Each school designs its program to fit the needs and experiences of its own students, choosing specific texts and topics with this in mind. At the end of the year, teachers collect a portfolio of each student’s work, which includes the specific assessment tasks, and grade it on a 5-point grading scale. To calibrate these grades, teachers put together a selection of portfolios from each grade level — one from each of the 5 score levels plus borderline cases — and send these to a regional panel for moderation. The panel of five teachers re-scores the portfolios and confers about whether the grade is warranted, making a judgment on the spread. A state panel also looks at specimens across schools as well. Based on these moderation processes, the school is given instructions to move grades up or down so that they are comparable to others.

A new initiative, Queensland’s “New Basics” and “Rich Tasks” approach to standards and assessment (which began as a pilot in 2003), also offers extended, multi-disciplinary tasks that are developed centrally and used locally when teachers determine the time is right and they can be integrated with locally-oriented curricula (Queensland Government, 2001). They are, say Queensland officials, “specific activities that students undertake that have real-world value and use, and through which students are able to display their grasp and use of important ideas and skills.”

Rich Tasks are defined as:

A culminating performance or demonstration or product that is purposeful and models a life role. It presents substantive, real problems to solve and engages learners in forms of pragmatic social action that have real value in the world. The problems require identification, analysis and resolution, and require students to analyze, theorize and engage intellectually with the world. As well as having this connectedness to the world beyond the classroom, the tasks are also rich in their application: they represent an educational outcome of demonstrable and substantial intellectual and educational value. And, to be truly rich, a task must be transdisciplinary. Transdisciplinary learnings draw upon practices and skills across disciplines while retaining the integrity of each individual discipline.

Science and Ethics Confer

Students must identify, explore and make judgments on a biotechnological process to which there are ethical dimensions. Students identify scientific techniques used as well as significant recent contributions to the field. They will also research frameworks of ethical principles for coming to terms with an identified ethical issue or question. Using this information they prepare pre-conference materials for an international conference that will feature selected speakers who are leading lights in their respective fields.

In order to do this students must choose and explore an area of biotechnology where there are ethical issues under consideration and undertake laboratory activities that help them understand some of the laboratory practices. This enables them to:

- A. Provide a written explanation of the fundamental technological differences in some of the techniques used, or of potential use, in this area (included in the pre-conference package for delegates who are not necessarily experts in this area).
- B. Consider the range of ethical issues raised in regard to this area's purposes and actions, and scientific techniques and principles and present a deep analysis of an ethical issue about which there is a debate in terms of an ethical framework.
- C. Select six real-life people who have made relevant contributions to this area and write a 150-200 word précis about each one indicating his/her contribution, as well as a letter of invitation to one of them.

This assessment measures research and analytic skills; laboratory practices; understanding biological and chemical structures and systems, nomenclature and notations; organizing, arranging, sifting through, and making sense of ideas; communicating using formal correspondence; précis writing with a purpose; understanding ethical issues and principles; time management, and much more.

A bank of these tasks now exists across grade levels, along with scoring rubrics, and moderation processes by which the quality of the tasks, the student work, and the scoring can be evaluated. Extensively researched, this system has had excellent success as a tool for school improvement. Studies found stronger student engagement in learning in schools using the Rich Tasks (Queensland Government, 2001). On traditional tests, New Basics students scored about the same as students in the traditional program, but they performed notably better on assessments designed to gauge higher order thinking. The Singapore government has employed the developers of the Queensland system to focus the new school improvement strategies upon performance assessments. High-scoring Hong Kong has also begun a process of expanding its already-ambitious school-based assessment system in collaboration with Queensland assessment developers.

Sweden

Over the past 40 years, Sweden's national assessment system has, like Finland's, shifted from a centralized system based on one test to a more localized system based on multiple forms of assessment. With this change, Sweden hoped to increase Upper Secondary school enrollment and provide more open access to higher education (Eckstein and Noah, 1993, p. 84). In 1977, Sweden abolished its *studentexamen*, a nationally administered exit exam that ranked Upper Secondary students and placed them in higher education programs. By abolishing its exit exam, Sweden aimed to reduce stress on pupils caused by the exam. The new policies also intended to produce more valid and reliable snapshots of whether students would succeed at the university level. Additionally, the country wanted to correct social and educational inequities caused by a one-size-fits-all assessment system (Eckstein and Noah, 1993, p. 230).

Sweden uses a national curriculum adjusted for the local context. In "Compulsory School," ages 7 to 16, Sweden has developed a common curriculum that includes nationally approved syllabi for the individual subjects. At the same time, every district adopts a local school plan outlining how to organize and develop schools in their region and each school has the flexibility to adapt the national curriculum and syllabi to local conditions (Swedish National Agency for Education, 2005). In "Upper Secondary School," ages 17 to 20, Sweden has 17 national programs consisting of 3-year programs providing a general education and eligibility to study at the post-secondary or university level. The programs include eight core subjects (English, the arts, physical education and health, mathematics, general science, social studies, Swedish or Swedish as a second language, and religion) and a set of subject-specific areas such as "the construction program" or "the business program" that combine general courses with specialized classes. The National Agency for Education determines the required courses for a national specialization, and most of the programs require at least 15 weeks of workplace training outside of school.

Sweden pairs nationally outlined and locally implemented curricula with multiple layers of assessment controlled by schools and teachers. Assessments in compulsory school consist of several components. First, during each school term, the teacher, student, and parent or guardian meet to discuss the student's learning and social development (Swedish National Agency for Education, 2005). Second, students receive grades from their teachers in each term of year 8 (age 15) and the end of the fall term of year 9 (age 16). Teachers base their grades on the goals in the syllabi. By meeting the goals, students receive grades of "Pass," "Pass with Distinction," or "Pass with Special Distinction" based on nationally approved assessment criteria (Swedish National Agency for Education, 2005).

Throughout the grades, schools use a number of diagnostic materials to assess students in Swedish, Swedish as a Second Language, and Mathematics prior to year 6, with English added in years 6 through 9. The diagnostic materials are not mandatory, but they help teachers assess students and support their learning. The diagnostic materials in years 6 through 9 assess where students stand in relation to the goals set by the syllabi (Swedish National Agency for Education, 2005).

Finally, students take nationally approved examinations in year 9 (age 16) and in Upper Secondary School. Teachers help design the tasks and questions, working with university faculty (Eckstein and Noah, 1993; O'Donnell, 2004). In year 9, the exams assess the subjects of Swedish, Swedish as a Second Language, English, and Mathematics. Teachers use these assessments as one factor in

determining students' grades, so that the grades reflect the national standards (Qualifications and Curriculum Authority, 2008). Regional education officials and schools provide time for teachers to calibrate their grading practices to minimize variation across the schools and region (Eckstein and Noah, 1993, p. 230). A similar process is used in Upper Secondary school to integrate national examinations in a wider array of subjects into teachers' grading.

The examinations administered during students' compulsory and Upper Secondary schooling use an open-ended, authentic approach to assessment. The exam questions grapple with real world contexts, asking students to use analytical and critical skills and draw on skill and content knowledge learned during their classes. The content of the assessments is closely matched with the national syllabi.

For example, Sweden's native language test at the upper elementary school level asks students about a broad theme. One year, the exam had the theme of "travel" and provided students with a contemporary poem, prose and poetry extracts from a variety of authors, a practical description of how to plan a trip, and data about travel presented in a set of texts, charts, and statistical tables (Eckstein and Noah, 1993, p. 119). Schools often give students materials a week in advance, so students have time to review the materials. Students have five hours to write an essay on the topic of their choice that will be evaluated on specific criteria emphasized in the syllabus from their course. The skills assessed include using appropriate language in certain circumstances, comprehending the different purposes of language, persuasive mechanisms, presenting information, word choice and grammar, as well as creative self expression.

Myriad examples of math assessment questions from Sweden show the connection to real world situations and the expectations for reasoning through problems. For example, an exam question from grade 5 asks students (aged 11-12) to grapple with a problem that they might have in their own lives:

Carl bikes home from school at four o'clock. It takes about a quarter of an hour. In the evening he's going back to school because the class is having a party. The party starts at 6 o'clock. Before the class party starts, Carl has to eat dinner. When he comes home, his grandmother calls, who is also his neighbor. She wants him to bring in her post before he bikes over to the class party. She also wants him to take her dog for a walk, then to come in and have a chat. What does Carl have time to do before the party begins? Write and describe below how you have reasoned (Pettersson, 2008).

The mathematics exam from the Upper Secondary level also frames the questions in real world, tangible topics and formats. Students have about 4 hours to answer 15 questions. The first 10 questions require short answers and the last 5 questions require longer answers for which students show their work (See figure 1, from Eckstein & Noah, 1993, pp. 270-272).

Figure 1: Swedish Mathematics Exam at the Third Year of the Upper Secondary Level

Short Format Questions
A coffee blender mixed x kg of coffee costing a kroner/kg with y kg of coffee costing b kroner/kg. Give a formula for the price per kg of the blend.
In 1976 Lena had a monthly salary of 6,000 kr. By 1984 her salary had risen to 9,000 kr. In current prices, her salary had risen by 50%. How large was the percent change in fixed prices? In 1976 the Consumer Price Index (CPI) was 382; in 1984 it was 818.
Long Format Question
A business paid into a pension fund at the beginning of every year a sum of 15,000 kr. The fund has a yearly growth rate of 10%. The first payment was made in 1987 and the last will be in 2010. The pension fund will continue to grow until 2015. How much more will the business have in the fund at the beginning of 2015, if it pays in the same amount as above, but the rate of growth is 15%?

Perhaps the most complex question surrounding these assessments when they are locally developed or scored is how to ensure comparability. Many of the systems described earlier, both in the U.S. and abroad, use common scoring guides. Queensland's system, like those in a number of countries, also employs "moderation," a process of bringing samples from different schools to be re-scored, with results sent back to the originating schools. This process leads to stronger comparability across schools and is part of building a strong performance assessment system.

Nebraska, through its peer review process, requires and verifies that scorers within each district participate in extensive scorer training on common rubrics and can demonstrate consistency in scoring. Although districts may be using different tools, consistency and comparability within classrooms, buildings, and districts is supported in this way. Valid comparison across districts is achieved through external validation checks such as the statewide writing assessment, the ACT and other commonly administered standardized tests. Each district's assessment system is evaluated and approved through a review process conducted by measurement experts.

Lessons Learned

The research and work that has been done on performance assessment has uncovered a number of benefits, challenges, and criteria for making such assessment systems successful. Among the benefits of performance assessment systems are that they:

- Elevate the focus of instruction to higher order thinking skills;
- Provide a more accurate and comprehensive assessment of what students know and can do;
- Lead to more student engagement in both the learning and assessment process;
- Invite more teacher engagement and encourage collaborative work;
- Support the improvement of teaching practices;
- Provide clearer information to parents as to student development, accomplishments, and needs; and
- Allow instruction to be altered in a timely fashion to meet student learning needs.

Key Questions

From the research and evidence on performance assessment, there are a number of key questions that should be considered when designing a system that substantially incorporates performance-based assessments:

1. What factors should be considered in developing performance assessments?

Careful attention must be paid to the quality, manageability, and scoring of performance tasks. They should be developed in response to criteria that establish the technical quality of assessments (including checking for bias and fairness), high proficiency standards, and ability to reveal useful information about student learning in response to core standards. Attention must be paid to how tasks will be administered fairly and consistently, and how students will have an opportunity to learn what is assessed. They should also be constructed to allow students with special needs and those who are learning English to have opportunities to demonstrate their knowledge appropriately.

2. Are performance assessments too expensive to be considered as a viable option for most schools?

Although some methods of managing performance assessments can cost more than machine scoring of multiple choice tests (i.e. when such assessments are treated as traditional external tests and shipped out to separately paid scorers), the cost calculus changes when assessment is understood as part of teachers' work and learning — and built into teaching and professional development time.

Much evidence suggests that developing and scoring these assessments is a high-yield investment in teacher learning and a good use of professional development resources (Darling-Hammond & Rustique-Forrester, 2005; Falk, 2000).

In most European and Asian systems, and in those used in several U.S. states, such as New York and Vermont, scoring of assessments is conducted by teachers and time is set aside for this aspect of teachers' work and learning. While teacher time to create and score assessments can be substantial, these activities ultimately lead to more skilled and engaged teachers. In contrast, external standardized tests provide teachers with little guidance on how to improve student learning when they simply receive numerical scores on secret tests months after the students have left school. Hence the professional development that seeks to help teachers improve achievement in this system is under-informed and less effective than it might be with more direct information about student thinking and performance.

These kinds of performance assessment systems use resources differently than accountability systems that rely upon standardized measures of achievement. For example, with its largely locally managed assessment system, Nebraska spends only \$.03 per child (or \$9,000) on outside testing contracts. By contrast, Ohio, which relies upon standardized measures, spends \$50.00 per child (or \$92,000,000). In systems that involve teachers in developing and scoring assessments, some of the funds otherwise used for outside contractors are instead spent on teachers' professional learning about assessment and on moderated scoring of tasks that raise the standard for learning and enable teachers to become more skillful in their teaching. Thus, this use of resources can be more cost effective in improving teaching and learning than external testing alone.

3. What types of professional development are needed to support the use of performance assessments?

Educators need opportunities to learn to build, use, and score assessments that will inform and guide their teaching. In some states and districts that have engaged in performance assessment work for some time, there is a substantial knowledge base in the field. In others, few teachers currently have that knowledge, but they can develop it with purposeful professional development that engages them in analyzing student work, developing and scoring performance tasks around standards, using the tasks in their classrooms, and debriefing with colleagues about how best to incorporate both the assessments and the feedback to students. These opportunities should be linked to other professional development occasions that help teachers work on teaching strategies within specific curricular contexts, so that assessment becomes part of instructional planning as well. Finally, there are opportunities for professional learning within the peer review, audit, or moderation systems that states or districts construct to check on consistency and provide feedback, as teachers discuss scoring together, compare and re-score tasks, and calibrate their judgments against the standards.

4. What should be the primary aim of performance assessments?

Productive use of performance assessments, like proper use of standardized tests, should be aimed at revealing areas where progress has been made and where improvement is needed. For students, there should be opportunities to continue to revise their work to meet standards and achieve mastery. In addition, both the scores that result from assessments and the finer-grained information

about how exactly students reason, think, and perform should be fed back to schools in ways that allow further analysis of both students needing support and areas of the curriculum needing development. Teachers should have opportunities to look at both aggregated information for different groups of students and individual samples of student work that can inform instructional planning and diagnostic teaching. All of these uses of assessment should lead to additional learning supports for students and teachers, rather than punishments that shut down inquiry and growth.

5. What are the roles of policymakers and administrators?

State and district leaders will also need to become skilled in developing and managing performance assessment systems, and bringing together resources for strong implementation of both state and local components. Since only classroom teachers can directly impact instruction and learning, a major task of legislators and personnel in departments of education is to provide assistance to on-the-ground educators who must make the system work. One key element of this support is common planning and learning time for teachers to practice looking at student work, to evaluate carefully the cognitive components of that work, and to change their instruction in response to diverse learners' performance and needs. Another key element is technical support for this critical work in the form of experts who can help guide these activities.

Policymakers should work to improve standardized assessments at the federal, state, and local levels so that they better represent students' abilities to reason, present, and defend their ideas, as well as to demonstrate their skills in authentic ways.

Perhaps most important, is that we must begin to think in terms of performance assessment *systems*, as do the nations and states described earlier. To support much improved educational outcomes, we will need systems in which standards, curriculum, instruction, and assessment are tightly intertwined, supporting high-quality learning in classrooms where students have many opportunities to demonstrate their knowledge and skills on standards-based tasks and teachers have many opportunities to learn how to teach them well.

References

- Black, Paul and Wiliam, Dylan (1998). "Inside the Black Box: Raising Standards Through Classroom Assessment." Phi Delta Kappan. Retrieved on October 10, 2008 from <http://www.pdkintl.org/kappan/kbla9810.htm>.
- Darling-Hammond, Linda and McCloskey, Laura (in press). "Assessment for Learning Around the World: What Would it Mean to Be "Internationally Competitive?" Phi Delta Kappan.
- Darling-Hammond, Linda and Rustique-Forrester, Elle (2005). "The Consequences of Student Testing for Teaching and Teacher Quality" In Joan Herman and Edward Haertel (eds.) *The Uses and Misuses of Data in Accountability Testing*. The 104th Yearbook of the National Society for the Study of Education, Part II, pp. 289-319. Malden, MA: Blackwell Publishing.
- Eckstein, M.A. and Noah, H.J.. (1993). *Secondary School Examinations: International Perspectives on Policies and Practice*. New Haven: Yale University Press.
- Falk, Beverly (2000). *The Heart of the Matter Using Standards and Assessment to Learn*. New York: Heinemann.
- Fine, M., Stoudt, B., and Futch, V. (June 2005). "The Internationals Network for Public Schools: A Quantitative and Qualitative Cohort Analysis of Graduation and Dropout Rates." New York: The Graduate Center, City University of New York.
- Foote, M. "Keeping Accountability Systems Accountable." Phi Delta Kappan, v. 88 (5), January 2007, pp. 359-363.
- O'Donnell, S. (December, 2004). International Review of Curriculum and Assessment Frameworks, Comparative tables and factual summaries. Slough, England: Qualifications and Curriculum Authority and National Foundation for Educational Research. Retrieved on May 8, 2008 from <http://www.inca.org.uk/pdf/comparative.pdf>.
- Partnership for 21st Century Skills (October, 2007). "21st Century Skills Assessment: A Partnership for 21st Century Skills e-paper." Retrieved on October 10, 2008 from http://www.21stcenturyskills.org/documents/21st_century_skills_assessment.pdf.
- Petterson, A. (2008). The National Tests and National Assessment in Sweden. Stockholm, Sweden. Stockholm Institute for Education. PRIM gruppen. Retrieved on May 31, 2008 from http://www.prim.su.se/artiklar/pdf/Sw_test_ICME.pdf.
- Qualifications and Curriculum Authority (2008). "Sweden: Assessment arrangements." Retrieved on May 8, 2008 from <http://www.inca.org.uk/690.html>.

Schmidt, W.H., Wang, H. C. and McKnight, C. (2005). "Curriculum Coherence: An Examination of US mathematics and Science Content Standards from an International Perspective." *Journal of Curriculum Studies*, 37 (5), 525-559.

Swedish Institute. (March 1984). "Primary and Secondary Education in Sweden." Fact Sheets on Sweden. Stockholm, Sweden.

Swedish National Agency for Education. (2005). "The Swedish School System: Compulsory School." Retrieved on May 31, 2008 from <http://www.skolverket.se/sb/d/354/a/959>.

Queensland Government (2001). "New Basics: The Why, What, How and When of Rich Tasks." Retrieved on September 12, 2008 from <http://education.qld.gov.au/corporate/newbasics/pdfs/rich-tasksbklet.pdf>.

Von Zastrow, Claus and Janc, Helen (March, 2004). "Academic Atrophy: The Condition of the Liberal Arts in America's Public Schools." Washington, DC: Council for Basic Education.