

Beyond Basic Skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning

April 2010

About this Brief

This brief outlines the findings of a SCOPE report summarizing a set of seven commissioned papers reviewing what has been learned over several decades about the use of performance assessments for measuring higher-order thinking and performance skills. Focusing on large-scale systems, these papers address technical advances, practices in the United States and abroad, feasibility and implementation issues, policy implications, uses with English language learners, and costs of performance assessments.

The project was funded by the Ford Foundation and the Nellie Mae Education Foundation and guided by an advisory board of leading education researchers, practitioners, and policy analysts. For more information and to read the papers, visit the SCOPE web site.



Stanford University
School of Education
520 Galvez Mall, CERAS Bldg
Stanford, CA 94305

<http://edpolicy.stanford.edu>
scope@stanford.edu
650.725.6600

By Linda Darling-Hammond & Frank Adamson, Stanford University

I am calling on our nation's governors and state education chiefs to develop standards and assessments that don't simply measure whether students can fill in a bubble on a test, but whether they possess 21st century skills like problem-solving and critical thinking, entrepreneurship and creativity.

— President Barack Obama, March 2009

Whether the context is the changing nature of work, international competitiveness, or, most recently, calls for common standards, it is clear today that the premium is not merely on students' acquisition of information, but on their ability to analyze, synthesize, and apply what they've learned and their capacity to solve problems, design solutions, and communicate effectively.

New Common Core Standards may point toward critically important knowledge and skills, but ensuring they are effectively implemented will depend on how schools design curriculum, organize teaching, and assess learning. High-achieving nations have pointed all of the elements of their systems toward challenging tasks that require students to use sophisticated knowledge to solve complex problems and explain their reasoning. These nations use primarily open-ended assessments that call for extensive writing, research, and applications of knowledge to novel situations.

However, the standardized tests that have been the linchpin of the federal No Child Left Behind Act (NCLB)—a law that has sought to promote school improvement by holding local educators accountable for their students' achievement—have largely failed to gauge students' mastery of the thinking skills that experts say they need to succeed in today's complex and fast-changing world.

While NCLB has appropriately cast in sharp relief the second-class educational status of students of color and those from disadvantaged backgrounds, the law's school accountability model and the standardized testing that undergirds it have relied heavily on multiple-choice questions measuring mostly low-level skills like the recall and recognition of information. Such questions can be administered and scored rapidly and inexpensively, but by their very nature are not well-suited to judge students' ability to express points of view, marshal evidence, and display other advanced skills.

As a result, studies have found that the tests have discouraged many teachers from teaching more ambitious intellectual skills, narrowing students'

opportunities to attain the higher standards that NCLB has sought for them. It’s not surprising, as a result, that scores have been rising on state tests used under NCLB, but American students’ performance has been declining steadily on tests that require students to apply knowledge, including the Programme for International Student Assessment, on which the United States now scores in the bottom tier of industrialized nations.

The Opportunity to Strengthen Assessment

A growing number of educators and policymakers have argued that new assessments are needed. For example, Achieve, a national organization of governors, business leaders, and education leaders, has called for a broader view of assessment:

States ... will need to move beyond large-scale assessments because, as critical as they are, they cannot measure everything that matters in a young person’s education. The ability to make effective oral arguments and conduct significant research projects are considered essential skills by both employers and post-secondary educators, but these skills are very difficult to assess on a paper-and-pencil test.

With the pending reauthorization of the federal Elementary and Secondary Education Act, of

which NCLB is the most recent incarnation, there is an opportunity to address this fundamental misalignment between the nation’s aspirations for its students and the assessments used to measure whether they are achieving those goals. The reauthorization opens the prospect for progress on measuring and encouraging the teaching of the advanced skills students need.

These new assessments would rely more heavily on what testing experts call “performance measures,” tasks requiring students to craft their own responses rather than merely select from among multiple-choice answers. They range from short-answer tasks, such as constructing and explaining a problem solution, to extended work like writing essays, engaging in research, and conducting laboratory investigations. Like the road test that virtually all adults have taken to gain a drivers license, these performance assessments ask students to demonstrate what they can actually *do* with their knowledge when it is applied in practice.

There are many examples of large-scale performance assessments in the United States and other countries, from the New York State Regents examinations to the hands-on science section of the National Assessment of Educational Progress, Connecticut’s and Vermont’s high school science assessments, writing assessments in many states, the

A Performance Assessment Prompt from the Collegiate Learning Assessment

You are the assistant to Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigational equipment. Sally Evans, a member of DynaTech’s sales force, recommended that DynaTech buy a small private plane (a SwiftAir 235) that she and other members of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. You are provided with the following documentation:

- 1: Newspaper articles about the accident
- 2: Federal Accident Report on in-flight breakups in single engine planes
- 3: Pat’s e-mail to you and Sally’s e-mail to Pat
- 4: Charts on SwiftAir’s performance characteristics
- 5: Amateur Pilot article comparing SwiftAir 235 to similar planes
- 6: Pictures and description of SwiftAir Models 180 and 235

Please prepare a memo that addresses several questions, including what data support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups, what other factors might have contributed to the accident and should be taken into account, and your overall recommendation about whether or not DynaTech should purchase the plane.



Collegiate Learning Assessment, England's General Certificate of Secondary Education exams featuring performance tasks in virtually all subject areas, and similar assessments in Hong Kong, Singapore, and Australia, among others.

Research shows that well-designed performance assessments yield a more complete picture of students' abilities and weaknesses, and can overcome some of the validity challenges of assessing English language learners and students with disabilities. The use of performance measures has been found to increase the intellectual challenge in classrooms and to support higher-quality teaching. Students who routinely engage in instruction where they are expected to demonstrate applications of their knowledge and explain and defend their answers have often been found to outscore peers on both traditional tests and more complex measures.

And by involving teachers in scoring essays and other performance measures, the way assessment systems in high-achieving nations and some states do today, teachers can become more knowledgeable about how to evaluate and teach to challenging standards. Teacher involvement in scoring has been found to offer a powerful professional development opportunity that translates into a stronger ability to design and implement standards-based curriculum. Such tests are thus tied more closely to the improvement of classroom instruction, and can support more expansive and productive student learning.

All of these factors are driving the increased use of performance assessments around the world. As the Hong Kong Education Examinations Authority explained while introducing new school-based performance assessments into its examination system:

The primary rationale for school-based assessments (SBA) is to enhance the validity of the assessment, by including the assessment of outcomes that cannot be readily assessed within the context of a one-off public examination, which may not always provide the most reliable indication of the actual abilities of candidates.... SBA typically involves stu-

dents in activities such as making oral presentations, developing a portfolio of work, undertaking fieldwork, carrying out an investigation, doing practical laboratory work or completing a design project, helps students to acquire important skills, knowledge, and work habits that cannot readily be assessed or promoted through paper-and-pencil testing. Not only are they outcomes that are essential to learning within the disciplines, they are also outcomes that are valued by tertiary institutions and by employers.

Challenges and Lessons

There are challenges to using performance measures on a large scale, including the need to ensure the tests' rigor and technical reliability and to manage their cost and time requirements. A number of states that implemented performance assessments in the early 1990s have since scaled them back as a result of technical concerns, implementation burdens, or costs, especially when NCLB increased testing requirements to reach every child every year. In addition, the federal Department of Education was often unwilling to approve innovative testing systems under NCLB.

But the experiences of a growing number of high-achieving nations using large-scale performance assessments effectively, the record of the International Baccalaureate and Advanced Placement (AP) testing programs, successful state experiences with performance assessments, and the growth of performance measures in the military and other sectors illustrate how such assessments can be reliably and cost-effectively incorporated into testing systems.

And studies have demonstrated that performance tasks can be designed in ways that allow them to measure student achievement accurately and permit the comparison of results across students and schools and from year to year—necessary features of tests used to hold schools accountable for their students' results.

Research shows that creating reliable, valid, feasible, and cost-effective performance assessments can be developed with attention to:

Careful task design based on a clear understanding of the specific knowledge and skills to be assessed and how they develop cognitively, what criteria define a competent performance, and rigorous field testing to ensure that the items or tasks are understandable and are measuring the intended concepts and abilities. When these principles are followed, studies have found that assessments can be made comparable and valid across time, tasks, and raters.

Reliable scoring systems based on standardization of tasks and well-designed scoring rubrics, training of scorers, moderation of the scoring process to ensure consistency in applying the standards, and auditing of the system to double check and upgrade comparability. Well-developed systems with these features have produced interrater reliability with levels of agreement of 90% or higher, comparable to the AP exams and other well-respected tests.

Methods for ensuring fairness based on the use of universal design principles, careful linguistic choices to avoid sources of confusion unrelated to the content being measured, cultural review of items, and piloting testing of tasks to see how they perform with different test-takers. Carefully designed performance assessments have often been found to produce more successful evaluations of knowledge than traditional tests for English language learners, special education students, and students with lower reading levels.

Effective use of technology to deliver and administer assessments; enable simulations, research tasks, and other sophisticated assessment opportunities; adapt assessments to better measure student abilities and growth; and support both human scoring and machine scoring of open-ended items, which is becoming more reliable and effective. As a measure of the potential for technology to streamline performance testing, the National Assessment of Educational Progress has found that human and computer scoring of a set of physics simulations matches 96% of the time.

Costs, especially for scoring, are another concern. Studies have found that performance-based tests tend to be about twice as expensive as tests that

rely exclusively on multiple choice questions. But a detailed cost modeling study grounded in real-world prices shows that it is possible to construct large-scale assessments that combine multiple-choice questions and performance measures for no more than today's much-less-informative tests—about \$20 per pupil for English language arts and math combined. Affordability would be accomplished by taking advantage of the economies of scale that will accompany states banding together in consortia, tapping the efficiencies of technology in administering tests and supporting scoring, and using teachers strategically in the scoring of performance items.

While looking to economize, it is also important to put the costs of high-quality assessment into perspective. Even if states spent \$50 per pupil on assessments (more than twice the study's estimate of the costs of a balanced system), this would still be less than 10% of the cost of interventions many are currently adopting to raise achievement, and far less than 1% of the costs of education overall.

While the use of performance tasks does require time and expertise, educators and policymakers in high-achieving nations believe that the value of rich performance assessments far outweighs their cost. Jurisdictions like Singapore, Hong Kong, Japan, England, and Australian states have expanded their use of performance tasks because these deeply engage teachers and students in learning, make rigorous and cognitively demanding instruction commonplace, and, leaders argue, increase students' achievement levels and readiness for college and careers.

At the end of the day, if standards are to influence learning in positive ways, they must be enacted in ways that enable students to learn to use their minds well and support teachers in developing strong instruction. For these reasons, consideration of performance assessment should be a critical aspect of the nation's analysis of how to achieve 21st century standards of learning