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

Linda Darling-Hammond & Frank Adamson

In collaboration with Jamal Abedi, Stuart Kahl, Suzanne Lane, William Montague, John Olson, Margaret Owens, Raymond Pecheone, Lawrence O. Picus, Ed Roeber, Brian Stecher, Thomas Toch, and Barry Topol

scope
Stanford Center for Opportunity Policy in Education

This study was conducted by the Stanford Center for Opportunity Policy in Education (SCOPE) with support from the Ford Foundation and the Nellie Mae Education Foundation.



The Stanford Center for Opportunity Policy in Education (SCOPE) supports cross-disciplinary research, policy analysis, and practice that address issues of educational opportunity, access, equity, and diversity in the United States and internationally.

Citation: Darling-Hammond, L. & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Stanford Center for Opportunity Policy in Education
Barnum Center, 505 Lasuen Mall
Stanford, California 94305
Phone: 650.725.8600
scope@stanford.edu
http://edpolicy.stanford.edu

scope
Stanford Center for
Opportunity Policy in Education

# Table of Contents

# Preface and Acknowledgements

This paper is the culminating report of a Stanford University project aimed at summarizing research and lessons learned regarding the development, implementation, consequences, and costs of performance assessments. The project was led by Linda Darling-Hammond, Charles E. Ducommun Professor of Education at Stanford University, with assistance from Frank Adamson and Susan Shultz at Stanford. It was funded by the Ford Foundation and the Nellie Mae Education Foundation and guided by an advisory board of education researchers, practitioners, and policy analysts, ably chaired by Richard Shavelson, one of the nation's leading experts on performance assessment. The board shaped the specifications for commissioned papers and reviewed these papers upon their completion. Members of the advisory board include:

Eva Baker, Professor, UCLA, and Director of the Center for Research on Evaluation, Standards, and Student Testing

Christopher Cross, Chairman, Cross & Joftus, LLC

Nicholas Donahue, President and CEO, Nellie Mae Education Foundation, and former State Superintendent, New Hampshire

Michael Feuer, Executive Director, Division of Behavioral and Social Sciences and Education in the National Research Council (NRC) of the National Academies

Edward Haertel, Jacks Family Professor of Education, Stanford University

Jack Jennings, President and CEO, Center on Education Policy

Peter McWalters, Strategic Initiative Director, Education Workforce, Council of Chief States School Officers (CCSSO) and former State Superintendent, Rhode Island

Richard Shavelson, Margaret Jacks Professor of Education and Psychology, Stanford University

Lorrie Shepard, Dean, School of Education, University of Colorado at Boulder

Guillermo Solano-Flores, Professor of Education, University of Colorado at Boulder

Brenda Welburn, Executive Director, National Association of State Boards of Education

Gene Wilhoit, Executive Director, Council of Chief States School Officers

A set of seven papers was commissioned to examine experiences with and lessons from large-scale performance assessment in the United States and abroad, including technical advances, feasibility issues, policy implications, uses with English language learners, and costs. These papers and their authors are listed below. This report draws extensively from these commissioned papers.

- Jamal Abedi, *Performance Assessments for English Language Learners.*
- Linda Darling-Hammond, with Laura Wentworth, *Benchmarking Learning Systems: Student Performance Assessment in International Context.*
- Suzanne Lane, *Performance Assessment: The State of the Art.*
- Raymond Pecheone and Stuart Kahl, *Developing Performance Assessments: Lessons from the United States.*
- Lawrence Picus, Frank Adamson, Will Montague, and Maggie Owens, *A New Conceptual Framework for Analyzing the Costs of Performance Assessment.*
- Brian Stecher, *Performance Assessment in an Era of Standards-Based Educational Accountability.*
- Barry Topol, John Olson, and Edward Roeber, *The Cost of New Higher Quality Assessments: A Comprehensive Analysis of the Potential Costs for Future State Assessments.*

All reports can be downloaded from http://edpolicy.stanford.edu.

We are grateful to the funders, the Advisory Board, and these authors for their careful analyses and wisdom. These papers were ably ushered into production by Barbara McKenna. Without their efforts, this project would not have come to come to fruition.

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

*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.*

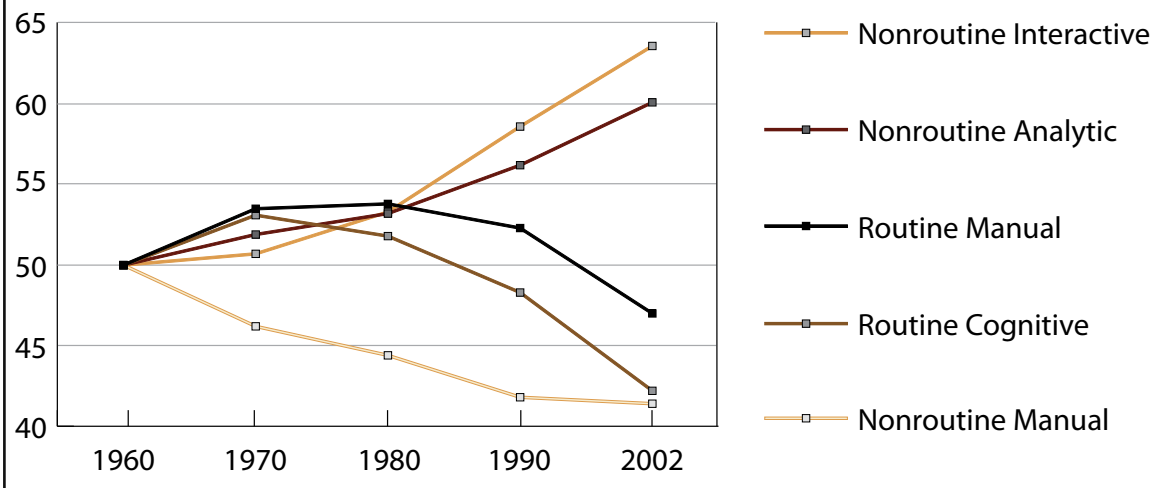
— President Barack Obama, March 2009

Reform of educational standards and assessments has been a constant theme in nations around the globe. As part of an effort to keep up with countries that appear to be galloping ever further ahead educationally, the nation's governors and chief state school officers recently issued a set of Common Core Standards that aim to outline internationally benchmarked concepts and skills needed for success in the modern world. The standards, which intend to create "fewer, higher, and deeper" curriculum goals, are meant to ensure that students are college and career-ready.

This goal has profound implications for teaching and testing. Genuine readiness for college and 21st century careers, as well as participation in today's democratic society, requires, as President Obama has noted, much more than "bubbling in" on a test. Students need to be able to find, evaluate, synthesize, and use knowledge in new contexts, frame and solve non-routine problems, and produce research findings and solutions. It also requires students to acquire well-developed thinking, problem solving, design, and communication skills.

These are the so-called "21st century skills" reformers around the world have been urging schools to pursue for decades—skills that are increasingly in demand in a complex, technologically connected, and fast-changing world. As research by economists Frank Levy and Richard Murnane shows, the routine skills used in factory jobs that once fueled an industrial economy have declined sharply in demand as they are computerized, outsourced, or made extinct by the changing nature of work. The skills in greatest demand are the non-routine interactive skills that allow for collaborative invention and problem solving. (See Figure 1, next page.)

Figure 1. How the demand for skills has changed
Economy-wide measures of routine and non-routine task input




In part, this is because knowledge is expanding at a breathtaking pace. Researchers at the University of California, Berkeley, estimate that, in the three years from 1999 to 2002, the amount of new information produced in the world approximately equaled the amount produced in the entire history of the world previously.[1] The amount of new technical information is doubling every two years.[2]

As a consequence, a successful education can no longer be organized by dividing a set of facts into the 12 years of schooling to be doled out bit by bit each year. Instead, schools must teach disciplinary knowledge in ways that also help students learn how to learn, so that they can use knowledge in new situations and manage the demands of changing information, technologies, jobs, and social conditions.

These concerns have driven educational reforms in nations around the globe. For example, as Singapore prepared to overhaul its assessment system, then Education Minister, Tharman Shanmugaratnam, noted:

> [We need] less dependence on rote learning, repetitive tests and a 'one size fits all' type of instruction, and more on engaged learning, discovery through experiences, differentiated teaching, the learning of life-long skills, and the building of character, so that students can… develop the attributes, mindsets, character and values for future success.[3]

Whether the context is the changing nature of work, international competitiveness, or, most recently, calls for common standards, the premium today is not merely on

students' acquiring information, but on recognizing what kind of information matters, why it matters, and how to combine it with other information.[4] Remembering pieces of knowledge is no longer the highest priority for learning; what students can *do* with knowledge is what counts.

## The Influence of Testing on Learning

The federal No Child Left Behind Act (NCLB), passed by Congress in 2001 to promote school improvement by holding schools accountable for students' achievement, cast in sharp relief the second-class educational status of students of color and those from disadvantaged background. At the same time, it elevated the importance of test-based accountability in the public education system.[5]

But the standardized tests that have been the linchpin of standards-based school reform, particularly the tests that states have introduced to comply with NCLB, have not focused primarily on the higher-order thinking and performance skills reformers called for. Tight NCLB testing timelines, the scope of every-child, every-year testing required under the federal law, and pressure from state elected officials to lower costs have led to tests that rely heavily on multiple-choice questions measuring mostly lower-level skills, such as the recall or recognition of information. These tests can be administered and scored rapidly and inexpensively, but by their very nature they are not well suited to judging students' ability to express points of view, marshal evidence, and display other advanced skills.

The General Accountability Office (GAO), the research branch of the U.S. Congress, reported in 2009 that the states' reliance on multiple-choice testing increased sharply in the NCLB era. Meanwhile, state education officials "reported facing trade-offs between efforts to assess highly complex content and to accommodate cost and time pressures."[6]

As RAND researcher Brian Stecher notes, multiple-choice tests do not reflect the nature of performance in the real world, which rarely presents people with structured choices.[7] With the possible exception of a few game shows, one demonstrates his or her ability in the real world by applying knowledge and skills in settings where there are no pre-determined options. A person balances her checkbook; buys ingredients and cooks a meal; reads an article in the newspaper and frames an opinion of the argument; assesses a customer's worthiness for a mortgage; interviews a patient, orders tests, and diagnoses the nature of his or her disease, and so on. Even in the context of school, the typical learning activity involves a mix of skills and culminates in a complex performance: a persuasive letter, a group project, a research paper, a first down, a band recital, a piece of art, etc. Rarely does a citizen or a student have to choose among four distinct alternatives.[8]

A key concern about the content and nature of tests is the growing recognition that assessment, especially when it is used for decision-making purposes, can exert powerful influences on curriculum and instruction. A long line of research has shown that—for

good or ill—tests can "drive" instruction in ways that mimic both the content and the format of tests.[9] Because schools tend to teach what is tested, the expansion of multiple-choice measures of simple skills into curriculum and extensive test preparation activities has especially narrowed the opportunities of lower-achieving students to attain the higher standards that NCLB sought for them. It has also placed a glass ceiling over more advanced students, who are unable to demonstrate the depth and breadth of their abilities on such exams. The tests have discouraged teachers from teaching more challenging skills by having students conduct experiments, make oral presentations, write extensively, and do other sorts of intellectually challenging activities that pique students' interest in learning at the same time.[10]

Assessment expert Lorrie Shepard and others have found that, when educators teach directly to the content and format of specific high-stakes tests, students are frequently unable to transfer their knowledge to items that test it in different ways.[11] Furthermore, students' ability to answer multiple-choice questions does not mean they have the ability to answer the same questions in open-ended form. Indeed, their scores often drop precipitously when answers are not provided for them, and they do not have the option to guess. Thus, a focus on multiple-choice testing gives false assurances about what students know and are able to do.[12]

This is why 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 postsecondary educators, but these skills are very difficult to assess on a paper-and pencil test.[13]

The NCLB school accountability model and the standardized testing that undergirds it have not catalyzed the law's pursuit of 21st century skills for all students. At best, they have established an academic floor for the nation's students, even though the law itself calls for schools to teach students to higher standards. And while many struggling students need large doses of reading and math to catch up, there's ample research revealing that sophisticated reading skills and the necessary vocabulary for comprehension are best learned in the context of history, science, and other subjects.[14] Yet, as the Center on Education Policy has documented, NCLB has narrowed the curriculum for many students, encouraging teachers to focus not only the content but also the format of the tests, at the expense of other essential kinds of learning.[15]

As one teacher noted in a national survey:

> Before [our state test] I was a better teacher. I was exposing my children to a wide range of science and social studies experiences. I taught using themes that really immersed the children into learning about a topic using their reading, writing, math, and technology skills. Now I'm basically afraid to NOT teach to the test. I know that the way I was teaching was building a better foundation for my kids as well as a love of learning.

Another, echoing the findings of researchers, observed:

> I have seen more students who can pass the [state test] but cannot apply those skills to anything if it's not in the test format. I have students who can do the test but can't look up words in a dictionary and understand the different meanings…. As for higher quality teaching, I'm not sure I would call it that. Because of the pressure for passing scores, more and more time is spent practicing the test and putting everything in [the test] format.[16]

A third raised the concern that many experts have pointed to—pressure to speed through the topics that might be tested in a curriculum that is a mile wide and an inch deep:

> I believe that the [state test] is pushing students and teachers to rush through curriculum much too quickly. Rather than focusing on getting students to understand a concept fully in math, we must rush through all the subjects so we are prepared to take the test in March. This creates a surface knowledge or many times very little knowledge in a lot of areas. I would rather spend a month on one concept and see my students studying in an in-depth manner.[17]

In contrast, international surveys have shown that higher-scoring countries in mathematics and science teach *fewer* concepts each year but teach them more deeply than in the United States, so that students have a stronger foundation to support higher order learning in the upper grades.[18] Ironically, states that test large numbers of topics in a grade level may encourage more superficial coverage leading to less solid learning.

It's thus not surprising that while student scores have been rising on the state tests used for accountability purposes under NCLB, scores have been declining on tests that gauge students' ability to *apply* knowledge to novel problems, such as the Programme for International Student Assessment, or PISA. In 2006, the United States ranked 21st of 30 OECD countries in mathematics and 21st of 30 in science, a decline in both raw scores and rankings from three years earlier. The 2003 scores were, in turn, a decline from the year 2000. Furthermore, U.S. students scored lowest on the problem-solving tasks.[19]

PISA differs from most tests in the United States, in that most items call on students to write their own answers to questions that require weighing and balancing evidence, evaluating ideas, finding and manipulating information to answer complex questions, and solving problems. These kinds of items resemble the tests commonly used in other countries, which routinely use extended essay questions and complex open-ended problems to evaluate knowledge. Students in many high-achieving nations also have to design and complete science investigations, technology solutions, and research projects as part of their examinations, ensuring their readiness for college-level work.

However, with the exception of a few states, we are still using basic-skills tests under NCLB that represent few of the higher-order skills and little of the in-depth knowledge needed for success in a rapidly changing and increasingly complex world.

The pending reauthorization of the federal Elementary and Secondary Education Act (ESEA), of which the No Child Left Behind Act is the most recent incarnation, offers an opportunity to address this fundamental misalignment between our aspirations for students and the assessments we use to measure whether they are achieving those goals. We have a chance to create a new generation of assessments that build on NCLB's commitment to accountability for the education of traditionally underserved groups of students, while measuring a wider range of skills and expanding the definition of accountability to include the teaching of such skills.

To match international standards, new assessments will need to rely more heavily on what testing experts call performance measures, tasks requiring students to craft their own responses rather than merely selecting multiple-choice answers. Researchers argue that, by tapping into students' advanced thinking skills and abilities to explain their thinking, performance assessments yield a more complete picture of students' strengths and weaknesses. And by giving teachers a role in scoring essays and other performance measures, the way the Advanced Placement and International Baccalaureate programs do today, performance-oriented assessments encourage teachers to teach the skills measured by the assessments and help teachers learn how to do so. Such measures would, in other words, focus attention more directly on the improvement of classroom instruction than NCLB has done.

There are challenges to using performance measures on a much wider scale, such as ensuring the measures' rigor and reliability, and managing them in ways that are affordable. At the same time, there are valuable lessons to be learned about how to address such challenges from a growing number of high-achieving nations that have successfully implemented performance assessments, some of them for many decades, as well as from state experiences with performance assessment, programs like the International Baccalaureate and Advanced Placement testing programs, and from the growth of performance measures in the military and other sectors.

These developments have been aided by substantial advances in testing technology over recent years as well. This large body of work suggests that performance assessments can pay

significant dividends to students, teachers, and policymakers in terms of improvements in teaching, learning, and the quality of information. Research also shows that the assessments can be built to produce confident comparisons of individual student performance over time and comparisons across schools, school systems, and states.

Our goal in this paper is to provide a thorough analysis of the prospects and challenges of introducing standardized performance assessments on a large scale when the Elementary and Secondary Education Act is reauthorized. This report describes the history and current uses of performance assessments in the United States and abroad. It summarizes the results of a set of commissioned papers from some of the nation's leading analysts, who synthesized decades of research on advances in and costs of performance assessments. Their work was overseen by an advisory board of leading testing and policy experts.

We hope that this work will inform the efforts of policymakers seeking a new, improved testing and accountability model under ESEA, one that measures the advanced skills that have become paramount and gives educators powerful incentives to pursue them.

## Performance Assessment: A Definition

For many people, performance assessment is most easily defined by what it is *not:* specifically, it is not multiple-choice testing. In a performance assessment, rather than choosing among pre-determined options, students must construct an answer, produce a product, or perform an activity.[20] From this perspective, performance assessment encompasses a very wide range of activities, from completing a sentence with a few words (short-answer), to writing a thorough analysis (essay), to conducting and analyzing a laboratory investigation (hands-on).

Because they allow students to construct or perform an original response rather than just recognizing a potentially right answer out of a list provided, performance assessments can measure students' cognitive thinking and reasoning skills and their ability to apply knowledge to solve realistic, meaningful problems.

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 and a demonstration of a set of driving maneuvers, including, in some parts of the country, the dreaded parallel parking technique. Few of us would be comfortable handing out licenses to people who have only passed the multiple-choice written test also required by the DMV. We understand the value of this performance assessment as a real-world test of whether a person can actually handle a car on the road. Not only does the test tell us some important things about potential drivers' skills, we also know that preparing for the test helps improve those skills as potential drivers practice to get better. (What parent doesn't remember the hair-raising

outings with a 16 year old wanting to practice taking the car out over and over again?) The test sets a standard toward which everyone must work. Without it, we'd have little assurance about what people can actually *do* with what they know about cars and road rules, and little leverage to improve actual driving abilities.

Performance assessments in education are very similar. They 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 write up; computer programs that students create and test out; research inquiries that they pursue, assembling evidence about a question that they present in written and oral form. Whether the skill or standard being measured is writing, speaking, scientific or mathematical literacy, or knowledge of history and social science research, students actually perform tasks involving these skills and the teacher or other rater scores the performance based upon a set of pre-determined criteria.

A good example of how differently skills are measured on performance assessments as compared to multiple-choice tests is provided by this example from Illinois. The state's eighth grade science learning standard for technological design 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 sinks. The desired answer is "Change the design and retest his boat." The classroom performance assessment, however, 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 and why a ball of clay can be made to float. If they are well conducted and carefully evaluated, such activities can combine hands-on inquiry with the demonstration of content knowledge and reasoning skills. They also enable the teacher to assess whether students can frame a problem, develop hypotheses, evaluate outcomes, demonstrate scientific understanding, use scientific facts and terminology, persist in problems solving, organize information, and develop sound concepts regarding the scientific principles in use.

Performance events can take several forms, including requests that can be answered by what are called "constructed-response" items—those that require students to create a response—within a relatively short time in a traditional "on-demand" test that students sit down to take. They can also include more extended tasks that require time in class. These performance tasks allow students to engage in more challenging activities

that demonstrate a broader array of skills, including problem framing and planning, inquiry, and production of more extended written or oral responses.

Examples of ***constructed response*** questions can be found in the "hands-on" science section of the National Assessment of Educational Progress (NAEP). In one item, twelfth-grade students are given a bag containing sand, salt, and three different metals. They are asked to separate the substances using a magnet, sieve, filter paper, funnel, spoon, and water and to document the steps they used to do so. This performance task requires students to conduct experiments using materials provided to them, and to record their observations and conclusions by responding to both multiple-choice and constructed-response questions. The example demonstrates a hybrid assessment model that tests student ability to physically conduct an experiment while also testing report writing and factual knowledge that are critical to scientific approaches to problems.

The New York Regents examinations include fairly ambitious constructed response elements in nearly all subject areas. The U.S. history test, for example, asks students to write essays on topics like the following: "Discuss the advantages and disadvantages of industrialization to American society between 1865 and 1920. In your essay, include a discussion of how industrialization affected different groups in American society."[21]

Another kind of common constructed response task occurs in writing tests that require students to formulate and develop ideas and arguments. For example, an English question on the International Baccalaureate exam asks students to choose essay questions within different literary genres and base their answer to questions requiring knowledge of literary techniques on a least two of three works studied in class. Questions like the following are common:

1. Using two or three of the works you have studied, discuss how and to what effect writers have used exaggeration as a literary device.

2. Acquiring material wealth or rejecting its attractions has often been the base upon which writers have developed interesting plots. Compare the ways the writers of two or three works you have studied have developed such motivations.

3. Discuss and compare the role of the speaker or persona in poems you have studied. You must refer closely to the work of two or three poets in your study and base your answer on a total of three or four poems.[22]

More ambitious ***performance tasks*** that occur in the classroom can test even more challenging intellectual skills that come even closer to the expectations for perfor-

mance found in colleges and careers. For example, a 9th and 10th grade Connecticut science assessment poses a problem for students to solve which requires that they develop hypotheses, design and conduct a brief experiment, record their observations, write up their findings, including displays of data, draw conclusions, and evaluate the validity of their results. (See Appendix A for an example of one task.) This classroom-embedded task, which all students complete, is scored by teachers and may factor into local grading. On the end-of-year statewide summative test, students receive a sample of a report from an experiment, which they have to analyze in terms of the appropriateness of its methods and the validity of its results, drawing on the experiences they have had in the classroom conducting experiments.

These tasks are similar to the expectations for science inquiry and analysis found in assessment systems in Australia, Canada, England, Finland, Hong Kong, Singapore, and many other high-performing countries. 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. These classroom-based assessments—which include research papers, applied science experiments, presentations, and products that students construct—are mapped to the core curriculum or syllabus and the standards for the subject. They are selected because they represent critical skills, topics, and concepts, and they are evaluated by teachers who are trained and calibrated to score comparably.

In most of these nations the expectations go even further than the Connecticut science task to require that students choose their own problem, design, and complete an extended investigation, and analyze the results in a paper that resembles a published scientific report. For example, science course examinations in Singapore (as in England and Australia) include an assessment of experimental skills and investigations that counts for at least 20 percent of the examination score. Teachers are trained to score these assessments using common criteria under conditions of both internal and external moderation for consistency. Following specifications from the Singapore Examinations and Assessments Board, students must:

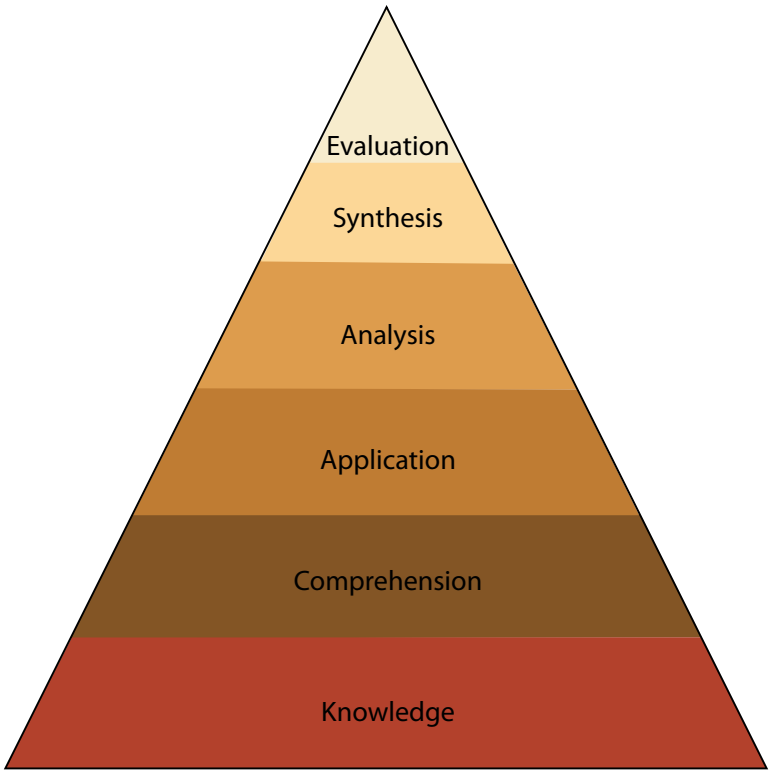
- Identify a problem, design and plan an investigation, evaluate their methods and techniques;
- Follow instructions and use techniques, apparatus, and materials safely and effectively;
- Make and record observations, measurements, methods, and techniques with precision and accuracy; and
- Interpret and evaluate observations and experimental data.[23]

## What it Means to Test Higher Order Skills

The key issue is not just whether a test expects students to provide an answer to an open-ended prompt or task, but what kind of knowledge and skill the student is expected to exhibit. Educators often refer to lower-level versus higher-order skills. The most well-known approach to describing these is Bloom's taxonomy of cognitive skills, shown in Figure 2.[24] At the bottom of the pyramid, defining lower-level skills, *knowledge* refers to memory and recollection of facts. *Comprehension* refers to demonstrating understanding of these ideas, while *application* refers to using this understanding to complete a task or solve a problem. The depth of understanding increases at each successive level.

The top half of the pyramid represents higher-order skills: *analysis* requires students to examine arguments, make inferences, and find evidence that supports explanations. In the *synthesis* phase, students compile information in different ways to produce a new pattern or alternative solution. *Evaluation* occurs when students weigh and balance evidence, evaluate ideas based on rigorous standards, present and defend ideas based on their judgments about information.

Figure 2. Bloom's Taxonomy of the Cognitive Domain




One of the differences in many U.S. tests and tests abroad is the extent to which they focus on higher-order skills. For example, a National Science Foundation study conduct-

ed during the 1990s found that on an extensive list of standardized mathematics tests, 95% of the items tested low-level thinking, 97% tested low-level conceptual knowledge and 87% tested low-level procedural knowledge. On science tests, 73% of items tested low-level thinking and 77% tested low-level conceptual knowledge.[25] These mathematics and science tests almost never assessed higher-order skills and thinking at the very top of Bloom's taxonomy.

Performance assessments that call for more analysis and manipulation of data and defense of ideas are often advocated because they offer a medium for students to display the higher-order skills of analysis, synthesis, and evaluation. These differences can show up on selected-response items as well as performance tasks. Compare, for example, a traditional item measuring basic recall (from a U.S. History test) with an analytic item developed by Alberta, Canada history teachers as part of Alberta's diploma examination—both evaluating knowledge of the same period of history—and notice how the second item requires deeper historical knowledge as well as the ability to compare and contrast situations across historical contexts.

***Who was president of the United States at the beginning of the Korean War?***

a) John F. Kennedy

b) Franklin D. Roosevelt

c) Dwight Eisenhower

d) Harry Truman

e) Don't know

***A feature common to the Korean War and the Vietnam War was that in both conflicts:***

a) Soviet soldiers and equipment were tested against American soldiers and equipment.

b) The United States became militarily involved because of a foreign policy of containment.

c) The final result was a stalemate; neither side gained or lost significant territory.

d) Communist forces successfully unified a divided nation.

Another example of how cognitive demands can differ for items that may look similar on the surface can be found in two different constructed response items on phys-

ics tests. An item from the New York State Regents Physics exam asks students to draw and label a circuit showing correct locations of resistors, an ammeter, and a voltmeter. Students are then asked to identify the equivalent resistance of the circuit and of a given resistor under specific conditions.[26]

This item does require application of knowledge by students; however, it does not go as far in testing higher-order skills as a similar item used on the high school physics examination in Hong Kong, one of the highest-scoring jurisdictions on PISA. First, the Hong Kong item asks students to identify the amount of current flowing through a resistor under different conditions and to explain their answers. Next, students are asked to sketch the time variation in the potential difference of the electrical pressure when a switch is opened. Finally, students are asked to show how they would modify the circuit to demonstrate particular outcomes under different conditions.[27] This type of question requires students to demonstrate a greater depth of knowledge, comprehension, application, analysis, and evaluation, as well as using their knowledge flexibly under changing situations, an important 21st century skill.

## Uses of Performance Assessments in the United States and Around the World

Performance assessments are common in high-achieving countries, which have long relied on open-ended items and tasks that require students to analyze, apply knowledge, and write extensively. Some, like top-scoring Finland, use only school-based performance assessments before 12th grade, developed by teachers in response to the national curriculum. These assessments emphasize students' ability to frame and conduct inquiries, develop products, represent their learning orally and in writing, and reflect on quality, with the goal of self-evaluation and ongoing improvement of their work.

At the 12th grade level, high school teachers and university faculty jointly develop the matriculation exam taken by students who want to go on to college. The open-ended items, which comprise the entire exam, ask students to apply and explain their knowledge in ways that demonstrate a deep understanding of the content under study. For example, mathematics problems require critical thinking and modeling, as well as straightforward problem solving. The basic mathematics exam poses this kind of problem:

> A solution of salt and water contains 25% salt. Diluted solutions are obtained by adding water. How much water must be added to one kilogram of the original solution in order to obtain a 10% solution? Work out a graphic representation which gives the amount of water to be added in order to get a solution with 2-25% of salt. The amount of water (in kilograms) to be added to one kilogram of the original solution must be on the horizontal axis; the salt content of the new solution as a percentage must be on the vertical axis.

And the advanced mathematics exam poses this one:

> In a society the growth of the standard of living is inversely proportional to the standard of living already gained, i.e. the higher the standard of living is, the less willingness there is to raise it further. Form a differential-equation-based model describing the standard of living and solve it. Does the standard of living rise forever? Is the rate of change increasing or decreasing? Does the standard of living approach some constant level?

Other nations, such as Singapore, Hong Kong, Australia, and England, use a combination of centralized assessments that feature mostly open-ended and essay questions and school-based tasks which are factored into the final examination scores. In England, for example, most students aim for the General Certificate of Secondary Education (GCSE), a two-year course of study evaluated by assessments both within and at the end of courses or units. The British system of examinations has informed systems in countries around the world, from Australia, Hong Kong, and Singapore, to the International Baccalaureate and the New York State Regents examinations. The exams involve open-ended items in on-demand tests and a set of structured, extended classroom-based tasks. England is currently introducing new tasks for the GCSE, called "controlled assessments," that emphasize applied knowledge and skills. These are either designed by the awarding body and marked by teachers or designed by teachers and marked by the awarding body, with teachers determining the timing of the assessments.

Table 1 shows the types of tasks that students complete to fulfill each of the English course units. Together, these comprise 60% of the examination score. They result in students engaging in significant extended writing, as well as speaking and listening, in multiple genres, from texts that are part of the syllabi developed from the national curriculum. Each of these tasks is further specified in terms of what students are asked to do and what criteria are used to evaluate their responses. An external examination body develops and monitors scoring protocols and processes to ensure consistency in evaluation.

**Table 1. Example of Tasks: GCSE English**

| Unit and Assessment | Tasks |
|---|---|
| Reading literacy texts | Responses to three texts from set choices of tasks and texts. Candidates must show an understanding of texts in their social, cultural and historical context. |
| Imaginative Writing | Two linked continuous writing responses from a choice of Text Development or Media. |
| Speaking and Listening | Three activities: a drama-focused activity; a group activity; an individual extended contribution. One activity must be a real-life context in and beyond the classroom. |
| Information and Ideas | Non-Fiction and Media: Responses to authentic passages. Writing information and Ideas: One continuous writing response—choice from 2 options. |

These tasks supplement written on-demand tests in which student answer essay questions about specific readings and situations. For example, referencing readings that students have completed during the course, one item asks students to compare and contrast pieces of literature, interpret authors' meanings, and analyze literary techniques:

> 1. Compare 'Blessing' with one other poem, explaining how the poets show their feelings and ideas about the different cultures in the poems. Write about:
>
> - what the poets' feelings are about the different cultures
> - what their ideas are about the different cultures
> - the methods they use to show their feelings and ideas.
>
> OR
>
> 2. Compare the ways in which the poets present people in 'Two Scavengers in a Truck, Two Beautiful People in a Mercedes' and one other poem that you have chosen from the Different Cultures section of the Anthology. Write about:
>
> - how the people in the poems are represented
> - how the different people in the poems are contrasted
> - what the poems say about the societies they describe
> - which of the poems you like more, and why.

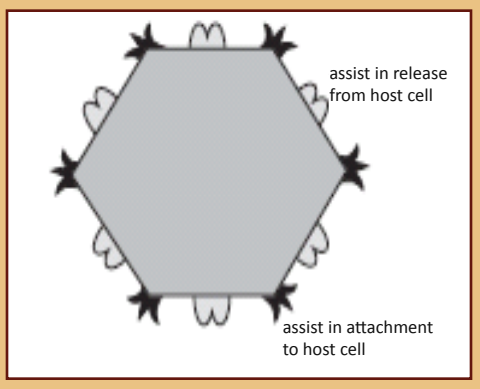
Similarly, in Victoria Australia, on-demand tests are supplemented with classroom-based tasks, given throughout the school year, that comprise at least 50% of the examination score. The performance tasks prepare students to succeed on the challenging end-of-course tests that demand high-level applications of knowledge. An example of an item from the high school biology test, for example (see page 16), describes a particular virus to students, asks them to design a drug to kill the virus and, in several pages, explain how the drug operates. It then asks them to design an experiment to test the drug.

## Victoria, Australia High School Biology Exam

When scientists design drugs against infectious agents, the term "designed drug" is often used.

A. Explain what is meant by this term:

Scientists aim to develop a drug against a particular virus that infects humans. The virus has a protein coat and different parts of the coat play different roles in the infective cycle. Some sites assist in the attachment of the virus to a host cell; others are important in the release from a host cell. The structure is represented in the following diagram:




The virus reproduces by attaching itself to the surface of a host cell and injecting its DNA into the host cell. The viral DNA then uses the components of host cell to reproduce its parts and hundreds of new viruses bud off from the host cell. Ultimately the host cell dies.

B. Design a drug that will be effective against this virus. In your answer outline the important aspects you would need to consider. Outline how your drug would prevent continuation of the cycle of reproduction of the virus particle. Use diagrams in your answer. Space for diagrams is provided on the next page.

C. Before a drug is used on humans, it is usually tested on animals. In this case, the virus under investigation also infects mice. Design an experiment, using mice, to test the effectiveness of the drug you have designed.

In preparation for this test, students taking Biology will have been assessed on six common pieces of work during the school year covering specific outcomes outlined in the syllabus. They will have conducted "practical tasks" like using a microscope to study plant and animal cells by preparing slides of cells, staining them, and comparing them in a variety of ways, resulting in a written product with visual elements. They also will have conducted practical tasks (labs) on enzymes and membranes, and on the maintenance of stable internal environments for animals and plants. Finally, they will have completed and presented a research report on characteristics of pathogenic organisms

and mechanisms by which organisms can defend against disease. These tasks, evaluated as part of the final examination score, link directly to the expectations that students will encounter on the external examination but go beyond what that examination can measure in terms of how students can apply their knowledge.

Like the behind-the-wheel test given for all new drivers, these performance assessments evaluate what students can *do* with what they know. The road test not only reveals some important things about drivers' skills; preparation for the test also helps *improve* those skills as novice drivers practice to get better. Teachers can get information and provide feedback to students as needed, something that traditional standardized tests cannot do. In addition, as teachers use and evaluate these tasks, they become more knowledgeable about how to teach to the standards and about their students' learning needs. Thus, the process improves the quality of teaching and learning.

As explained by the Hong Kong Education Examinations Authority, which is rapidly increasing the use of school-based assessments in 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 students 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, help 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.[28]

As we have noted, a number of states have also developed and use such hands-on assessments as part of their state testing systems. Indeed, the National Science Foundation provided hundreds of 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.

- Connecticut uses extended writing tasks and 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 many

constructed response items with locally developed assessments that provide evidence of student work from performance tasks or portfolios.

- Missouri and Kentucky each have developed systems of on-demand testing including substantial constructed response components, supplemented with state-designed, locally-administered performance tasks, scored in reliable ways.

- New York's Regents exams contain a variety of performance components. The English exam asks students to write three different kinds of essays. The history/social studies examinations use document-based questions to elicit essays that reveal students' ability to analyze texts and data, as well as to draw and defend conclusions. Science examinations contain a laboratory performance test.

- Also in New York, the New York Performance Assessment Consortium is a network of 47 schools that rely upon performance assessments to determine graduation. All students must complete and defend (dissertation-style) a literary analysis, science investigation, social science research paper, mathematical model, arts demonstration, and a technology demonstration that meet specific standards. Research from their work indicates that New York City students who graduate from these schools (which have a much higher graduation rate than the city as a whole, although they serve more low-income students, students of color, and recent immigrants) are more successful in college than most students nationally.

- In California, many school districts use the Mathematics Assessment Resource Services (MARS) tests, an assessment program developed with researchers from the Shell Center in England, which requires students to learn complex knowledge and skills to do well on a set of performance-based tasks. The evidence is 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.

- The Ohio Performance Assessment Project has developed curriculum-embedded, performance tasks at the high school level, aligned to college and workplace readiness standards. These assessments can serve as: 1) components of an end-of-course examination system; 2) an alternative means for students to demonstrate subject matter mastery; or 3) a way to satisfy the state's senior project requirement.[29]

The Ohio tasks are in many ways similar to those found in European and Asian systems. As components of course-based teaching and learning systems, they are designed to measure core concepts and skills in the disciplines that go beyond what can be assessed in a single period on a sit-down test. For example, in English language arts, students apply their understanding of a central theme in American literature to a task that requires selecting, analyzing, interpreting, and explaining texts.

### Ohio Performance Assessment Project
### English Language Arts Performance Task

Imagine that you are editing an on-line digital anthology for 11th-12th graders entitled, "Perspectives on the American Dream." Your job is to prepare the introduction to this anthology. In your introduction, please do the following things:

a) Decide which texts you want to include and in which order (you must include at least **six** texts). Texts can include books, poems, songs, short stories, essays, photographs, articles, films, television shows, or Internet media. The six texts must represent at least two different perspectives and must include at least two different types of text (e.g., print text, visual media, audio media, multi-media, digital media).

b) Identify and discuss different perspectives on the American dream represented in the six texts you selected.

c) Write a short paragraph about each text, in which you make clear why you have included it and how it relates to the other texts in your anthology.

d) Propose a set of questions to focus readers as they consider the perspectives represented in these texts.

In a mathematics task, students are asked to evaluate how heating costs may change as a simultaneous function of temperature, fuel costs, and savings due to insulation. The task requires students to apply their knowledge of ratio, proportion, and algebraic functions to a complex, real-world problem. They must engage in analysis and modeling of multiple variables. The response requires a display, explanation, and defense of their ideas.

### Ohio Performance Assessment Project "Heating Degrees" Task

The task starts with a scenario in which Ms. Johnson installs new insulation to save money on heating costs, but then learns that her bills have not declined by much from the previous year. Her contractor points out that heating costs have risen and weather has been colder. Ms. Johnson wants to find out how much she has actually saved due to the insulation she installed. On the basis of the situation painted above, details about Ms. Johnson's heating bills (rates, units of heat used), temperature changes, and some initial information to help them begin to research "heating degree days" on the internet, students are given two tasks:

(1) Assess the cost-effectiveness of Ms. Johnson's new insulation and window sealing. In their assessment, they must do the following:

- Compare Ms. Johnson's gas bills from January 2007 and January 2008
- Explain Ms. Johnson's savings after the insulation and sealing.
- Identify circumstances under which Ms. Johnson's January 2008 gas bill would have been at least 10% less than her January 2007 bill.
- Decide if the insulation and sealing work on Ms Johnson's house was cost-effective and provide evidence for this decision.

(2) Create a short pamphlet for gas company customers to guide them in making decisions about increasing the energy efficiency of their homes. The pamphlet must do the following:

- List the quantities that customers need to consider in assessing the cost-effectiveness of energy efficiency measures.
- Generalize the method of comparison used for Ms. Johnson's gas bills with a set of formulas, and provide an explanation of the formulas.
- Explain to gas customers how to weigh the cost of energy efficiency measures with savings on their gas bills.

The National Assessment of Educational Progress (NAEP) has also incorporated performance assessments in recent years, returning to the practices that defined the tests when they were first launched in the 1960s. For example, in a recent pilot embedded in the 2009 NAEP science assessment, students were required to design and conduct experiments, interpret results, and formulate conclusions. As part of the simulations, students needed to select values for independent variables and to make predictions as they designed their experiments. To interpret their results students needed to develop

tables, graphs and formulate conclusions. In addition to these scientific inquiry tasks, tasks were developed to assess students' search capabilities on a computer.

One eighth grade computer-based simulation task required students to investigate why scientists use helium gas balloons to explore outer space and the atmosphere. Below is an example of an item within this task that required students to search a simulated World Wide Web:

> Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from a spacecraft near other planets.
>
> Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.
>
> Base your answer on more than one web page or site. Be sure to write your answer in your own words.[30]

This task assesses students' on-line research skills. A related scientific inquiry task that required students to evaluate their work, form conclusions and provide rationales after designing and conducting a scientific investigation is provided below.[31]

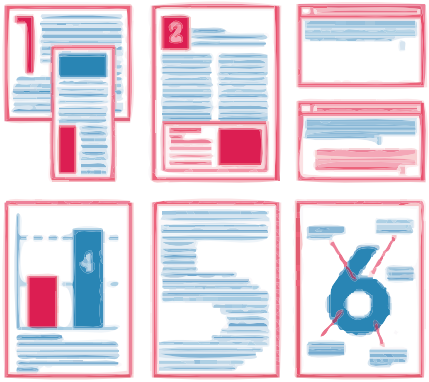
> How do different amounts of helium affect the altitude of a helium balloon? Support your answer with what you saw when you experimented.

These simulation tasks, which assess problem-solving, reasoning and evaluation skills valued within the scientific discipline, provide new possibilities for evaluating student cognition and learning.

Finally, a new Collegiate Learning Assessment (CLA), and a companion College and Work-Ready Assessment (CWRA) at the high school level, pose complex problems in scenarios that require data analysis, synthesis of many sources of information, evaluation of evidence, and explanation of a reasoned and well-grounded response. Administered on-line, students have 90 minutes to write an extended essay after evaluating these materials. Used with 60,000 college students annually, the CLA is largely scored by machine, with back reading by human scorers, whose ratings correlate with computer scores at a very high level. Many colleges use the CLA to evaluate value-added gains in learning for cohorts of students over the four years of college, and the CWRA will soon be available for similar analyses in high schools.[32]

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 & 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.

# The Challenges of Performance Assessments

Despite these advances and the ongoing work in the state and national assessments described earlier, legitimate questions and concerns remain about performance assessments. In the late 1980s and early 1990s, many states began to design and implement performance assessments; however, technical concerns, costs, and the demands of testing under NLCB led many to reduce or abandon performance components of their state accountability systems, even if they were retained locally. Some of the problems states encountered were due to difficulties with scoring reliability, implementation burdens, and costs, while others came from energized stakeholder groups who objected to aspects of the assessments or the manner in which they were implemented. In some states, people objected because the assessments were unfamiliar and stretched the boundaries of traditional testing. In others, the assessments were implemented in ways that did not take account of the needs for educator support, training, and time for participation. Most recently, under NCLB, many states had difficulty receiving approval from the federal Department of Education for performance elements of their systems.

But research in the United States and other countries on past and present performance assessments suggests that these challenges can be overcome, and that performance assessments can play an important role in ensuring that the nation's students learn the higher-order skills they need.

## Reliability and Validity

A central concern for any assessment is the credibility of results, which rests in large part on the reliability and validity of the measures—that is, whether they actually measure the skills and knowledge that are intended, and whether they do so consistently

and comparably across students, schools, tasks, and raters. Researchers agree that well-designed performance assessments offer more valid means to measure many kinds of learning, but many stakeholders have raised concerns about their reliability.

For example, in the early years of performance assessment in the United States, Vermont introduced a portfolio system in writing and mathematics that contained unique choices from each teacher's class as well as some common pieces. Because of this variation, researchers found that teachers could not score the portfolios consistently enough to accurately compare schools.[33] The key problem was the lack of standardization of the portfolios.

Since then, studies have reported much higher rates of scoring consistency for more standardized portfolios featuring common task expectations and analytic rubrics, like those that evolved later in Vermont and were ultimately developed in Kentucky. The Kentucky writing portfolio is actually a set of common performance tasks: There are three writing samples in different genres, with specific guidelines for each task and specific rubrics for scoring them. Over time, with teacher training and a statewide audit system, reliability has increased to the point that auditors who re-score randomly selected portfolios show rates of agreement with the original ratings of 99% for exact or adjacent scores.[34]

When performance assessments are used to judge schools and students, testing officials must develop strategies for standardizing the content that is measured, the administration of the assessments, and the scoring of student performances over time to ensure the quality and validity of the scores. This is not easy to do on a large scale with tests that require students to construct their own answers, because such tests often require human scorers. But improvements in test administration and scoring are improving the viability of large-scale performance assessments.

Researchers working in this field have found methods that help ensure the quality of performance tasks, producing more valid and stable results for a wide range of students.[35] In this section, we describe advances that have been made in these areas over the last two decades of work on performance assessments.

**Task Design.** A high-quality performance assessment is based on what we know about student learning and cognition in the specific domain, as well as a clear understanding of the specific knowledge and skills (or construct) to be assessed, the purpose of the assessment, and the interpretations to be drawn from the results. It is also closely aligned to the relevant curriculum.[36] It also is built to reduce what is known as "construct irrelevant variance"—that is, aspects of a task that might confuse the measurement of the central knowledge or skill being assessed. For example, the use of unnecessarily complicated vocabulary or syntax in a task may undermine the accurate measurement of mathematics skills for English learners. Simplifying the language while retaining the central features of the mathematics to be evaluated makes the task a more valid measure.

Assessments are stronger when test specifications are clear about what cognitive skills, subject matter content and concepts are to be assessed and what criteria define a competent performance.[37] Specifications of content, skills, and criteria can guide templates and scoring rubrics that are used with groups of tasks that measure the same sets of skills. Rubrics and templates help ensure that the content of the assessment is comparable across years to allow for measuring change in student performance over time.[38]

Suzanne Lane gives an example of a template for an "explanation task."[39] It asks students to read one or more texts that require some prior knowledge of the subject domain, including concepts, principles, and declarative knowledge, in order to understand them, and to evaluate and explain important issues introduced in the text. Consider an explanation task developed in Hawaii:[40]

> Imagine you are in a class that has been studying Hawaiian history. One of your friends, who is a new student in the class, has missed all the classes. Recently, your class began studying the Bayonet Constitution. Your friend is very interested in this topic and asks you to write an essay to explain everything that you have learned about it.
>
> Write an essay explaining the most important ideas you want your friend to understand. Include what you have already learned in class about Hawaiian history and what you have learned from the texts you have just read. While you write, think about what Thurston and Liliuokalani said about the Bayonet Constitution, and what is shown in the other materials.
>
> Your essay should be based on two major sources:
>
> 1. The general concepts and specific facts you know about Hawaiian history, and especially what you know about the period of Bayonet Constitution.
>
> 2. What you have learned from the readings yesterday.

Prior to receiving this task, students were required to read the primary source documents referred to in the prompt. This task requires students to not only make sense of the material from multiple sources, but to integrate material from these multiple sources in their explanations. This provides just one example of a task that can be generated from the explanation task template. An assessment system could use this type of task each year and replace the content while maintaining the central features of the task.

**Task Review and Field Testing.** Researchers have found that more valid and reliably scored tasks result from careful review and field testing of items and rubrics to ensure that they measure the knowledge and skills intended. This includes interviewing

students as they reflect on what they think the task is asking for and how they tried to solve it.[41] The individual piloting of tasks also provides an opportunity for the examiner to pose questions to students regarding their understanding of task wording and directions, and to evaluate their appropriateness for different subgroups of students, such as students whose first language is not English.

Large-scale field testing provides additional information regarding the quality of the tasks, including the psychometric characteristics of items. This includes analyzing student work to ensure that the tasks evoke the knowledge and skills intended and that the directions and wording are clear, and testing different versions of tasks to see which work best across different groups of learners. When these processes are followed, developers have been able to create tasks that are more clearly valid for their intended purposes and are more reliably scored.

**Scoring.** Perhaps the most frequently asked question surrounding these assessments is how to ensure comparability in scoring across different raters. Most of the systems described earlier, both in the United States and abroad, use common scoring guides, or rubrics, and engage graders in training, calibration, and moderation processes to ensure consistency.

Much has been learned about how to establish effective processes of training and moderation. We noted earlier the strong inter-rater reliability that has been achieved in the Kentucky writing portfolio, for example, which consists of a set of tasks within specified genres, with well-constructed scoring rubrics, teacher-moderated scoring processes, and a strong audit system that provides feedback to schools. Many developers of performance assessments have learned how to manage these processes in ways that achieve inter-rater reliabilities around 90%, matching the level achieved in the Advanced Placement system and on other long-standing tests.[42]

Human scoring of performance tasks has been found to be highly reliable when tasks are standardized and when scorers are effectively trained to share a common understanding of the scoring rubric so as to apply it consistently. Valid and reliable scoring is also enhanced by the design of quality scoring rubrics. Such rubrics:

- are designed for a family of tasks or a particular task template;
- include criteria aligned to the processes and skills that are to be measured—for example, in a mathematics task, students' computational fluency, strategic knowledge, and mathematical communication skills;
- develop criteria for judging the quality of the performance with the involvement of content and teaching experts who know the domain and understand how students of differing levels of proficiency would approach the task;

- identify score levels that reflect learning progressions as well as each of the important scoring criteria; and

- are validated via research with a range of students.[43]

A variety of systems for calibration and moderation of teacher scoring exist around the world. In New York State, teacher scoring of Regents examinations occurs at the school or regional level following training and is supplemented by a regular audit of scores from the state department of education, which can follow up with both rescoring and retraining of teachers. In Alberta, Canada, teachers are convened in centralized scoring sessions that involve training against benchmark papers and repeated calibration of scores until high levels of consistency are achieved. All scoring occurs in these sessions, with "table leaders" continually checking and re-checking the scoring for consistency, while it is going on. In England and Singapore, similar strategies are used, with benchmark papers and student "record files" used to train teachers and calibrate scoring. In addition, moderation processes are used within schools for teachers to calibrate their scores to benchmarks and to each other, while external moderators also examine schools' scored examinations and initiate additional training where it is needed. At the high school level, examination boards perform these functions of training and calibrating scorers.

In Queensland, Australia, samples of performance tasks from schools are rescored by panels of expert teachers, who guide feedback to schools and potential adjustments in scores. In Victoria, Australia, the quality and appropriateness of the tasks, student work, and grades is audited through an inspection system, and schools are given feedback on all of these elements. In both of these jurisdictions, statistical moderation is used to ensure that the same assessment standards are applied to students across schools. The schools' results on external exams are used as the basis for this moderation, which adjusts the level and spread of each school's performance assessments of its students to match the level and spread of the same students' collective scores on the common external test score.

In the International Baccalaureate program, which operates in 125 countries, teachers receive papers to score via computer delivery, and they calibrate their scoring to common benchmarks through an on-line training process that evaluates their ability to score accurately. The teachers upload their scored papers to be further evaluated or audited, as needed, and to have the scores recorded. Similarly, in Hong Kong, most delivery and scoring of open-ended assessments is becoming computer-based, as it is in 20 other provinces of China. There, as in many other places, double scoring is used to ensure reliability, with a third scorer called in if there are discrepancies.

More recently, automated scoring procedures have also been developed to score both short and long constructed-response items. Automated scoring has been used successfully in contexts ranging from state end-of courses exams to the Collegiate Learning Assessment[44] and the National Assessment of Educational Progress—in both the Math Online project that required students to provide explanations of their mathematical reasoning and the NAEP simulation study that required students to use search queries.[45] In the NAEP study that used physics simulations, the agreement between human raters and computer ratings in a cross-validation study was 96%. In the more complex, extended CLA task, correlations of human and computer ratings are nearly as high, at 86%.[46]

**Measuring Growth.** There is much work to be done on assessments, generally, to ensure that they can better measure gains in student learning. The problem with many tests currently used to measure gains is that they may measure items that reflect what states define as grade-level standards, but they do not measure student progress along a well-justified scale representing growing understanding of concepts or development of skills. These concerns are true regardless of the kinds of item types being used.

Some assessment experts, like Robert Mislevy, argue that performance assessments can allow for better measurement of growth and change in higher-order cognitive abilities and problem solving strategies, based, in part, on analyses of the numbers and kinds of strategies students use.[47]

Others have pointed out the potential for advances in measuring growth by designing performance assessments that reflect learning progressions. Learning progressions indicate what it means to acquire understanding within a content domain, and they identify where a student is on the continuum of the underlying construct. The *progress map* shown in Figure 3 (page 28) illustrates a learning progression from Australia's Developmental Assessment program. A student's progress in understanding number concepts can be charted on this continuum, which provides a picture of individual growth against a backdrop of normatively established expectations.[48]

These kinds of progressions can be used, as they have been in Australia and England, to design content standards and performance tasks that measure gains in students' learning as they develop understanding and competency in the content domain.[49] Further, they have the potential to lead to more meaningful scaling of assessments that span grade levels, and thus more valid score interpretations regarding student growth. Research and development that builds on the work already underway in other countries could allow significant progress on this front.

## Figure 3. Progress Map for Counting and Ordering

Below is the lower portion of a counting and ordering progress map. The map shows examples of knowledge, skills, and understandings in the sequence in which they are generally expected to develop from grades one through five. This type of map is useful for tracking the progress of an individual child over time. An evaluation using tasks designed to tap specific performances on the map can provide a "snapshot" showing where a student is located on the map, and a series of such evaluations is useful for assessing a student's progress over the course of several years.

| | |
|---|---|
| 1 | Counts collections of objects to answer the question "How many are there?"<br>Makes or draws collections of a given size (responds correctly to, "Give me 6 bears")<br>Makes sensible estimates of the size of small collections up to 10 (for 7 buttons, 2 or 15 would not be a sensible estimate, but 5 would be)<br>Skip counts in 2s or 3s using a number line, hundred chart, or mental counting (2, 4, 6…)<br>Uses numbers to decide which is bigger, smaller, same size (If he has 7 mice at home and I have 5, then he has more)<br>Uses the terms first, second, third (I finished my lunch second) |
| 2 | Counts forward and backward from any whole number, including skip counting in 2s, 3s, and 10s<br>Uses place value to distinguish and order whole numbers (writes four $10 notes and three $1 coins as $43)<br>Estimates the size of a collection (up to about 20)<br>Uses fractional language (half, third, quarter, fifth, tenth) appropriately in describing and comparing things<br>Shows and compares unit fractions (finds a third of a cup of sugar)<br>Describes and records simple fractional equivalents (the left over half pizza was as much as two quarters put together) |
| 3 | Counts in common fractional amounts (two and one-third, two and two-thirds, three, three and one-third)<br>Uses decimal notation to two places (uses 1.25 m for 1 m 25 cm; $3.05 for three $1 coins and one 5-cent coin; 1.75 kg for 1750 kg)<br>Regroups money to fewest possible notes and coins (11 x $5 + 17 x $2 + 8 x $1 regrouped as 1 x $50 + 2 x $20 + $5 + $2)<br>Uses materials and diagrams to represent fractional amounts (folds tape into five equal parts, shades 3 parts to show three-fifths)<br>Expresses generalizations about fractional numbers symbolically (1 quarter = 2 eighths and 1/4 = 2/8) |
| 4 | Counts in decimal fraction amounts (0.3, 0.6, 0.9, 1.2…)<br>Compares and orders decimal fractions (orders given weight data for babies to two decimal places)<br>Uses place value to explain the order of decimal fractions (which library book comes first-65.6 or 65.126? why?)<br>Reads scales calibrated in multiples of ten (reads 3.97 on a tape measure marked in hundredths, labeled in tenths)<br>Uses the symbols =, <, and > to order numbers and make comparisons (6.75 < 6.9; 5 x $6 > 5 x $5.95)<br>Compares and orders fractions (one-quarter is less than three- eighths) |
| 5 | Uses unitary ratios of the form 1 part to X parts (the ratio of cordial to water was 1 to 4)<br>Understands that common fractions are used to describe ratios of parts to whole<br>(2 in 5 students ride to school. In a school of 550, 220 ride bikes)<br>Uses percentages to make straightforward comparisons (26 balls from 50 tries is 52%; 24 from 40 tries is 60%, so that is better)<br>Uses common equivalences between decimals, fractions, and percentages (one-third off is better than 30% discount)<br>Uses whole number powers and square roots in describing things (finds length of side of square of area 225 sq cm as a square root of 225) |

*Source: Adapted from Masters and Forster (1996, p. 2). Knowing What Students Know.* Reprinted *with permission in Shepard (2005). Assessment.* In Preparing Teachers for a Changing World *(Jossey-Bass.)*

## How Assessments Affect Teaching and Learning

Fundamental to the validation of test use is the evaluation of the intended and unintended consequences of the use of an assessment—known as "consequential validity."[50] Because performance assessments are intended to improve teaching and student learning, it is particularly essential to obtain evidence of whether they have these effects, or any negative effects.[51]

It seems clear that, as with other kinds of testing, the use of more open-ended assessments affects teaching practice. In the 1990s, when performance assessments were launched in a number of states, studies found that teachers assigned more writing and mathematical problem solving of the kinds demanded on the new assessments in states ranging from California to Kentucky, Maine, Maryland, Vermont, and Washington.[52] Studies have found that well-designed performance assessments encourage instructional strategies that foster reasoning, problem solving and communication, as well as a focus on activities such as research and writing.[53]

Performance assessments that measure complex thinking skills have been shown to influence student learning, as well.[54] School level studies have found greater increases in performance on both traditional standardized tests and more complex measures for students in classrooms that offer a problem-oriented curriculum that regularly features performance assessment, as compared to other classrooms.[55] On a larger scale, Suzanne Lane and colleagues found that school achievement over a five-year period on Maryland's performance-based MSPAP test was strongly related to schools' incorporation of related teaching practices in reading, writing, mathematics and science.[56] Furthermore, a research team led by testing expert Robert Linn found that these gains carried over to the National Assessment of Educational Progress.[57]

One reason that performance assessments embedded in classroom instruction may help support stronger learning for students is that they ensure that students are undertaking intellectually challenging tasks. If teachers use these kinds of assignments consistently, with feedback and opportunities to revise to meet high standards, the level of rigor in the classroom increases. In addition, these assessments can provide information to teachers regarding how students think and try to solve problems.[58] This feedback allows teachers to diagnose students' strengths as well as gaps in understanding. Because performance assessment tasks often yield multiple scores in different domains of performance, reflecting students' areas of strength and weakness, they can also help teachers identify what kind of help students need, so they can tailor instruction accordingly.[59]

Furthermore, the clear criteria and rubrics that accompany good performance tasks can help students improve their work, especially if these carry over across multiple formative and summative assessments over time. For example, if writing is repeatedly evaluated for its use of evidence, accuracy of information, evaluation of competing viewpoints, development of a clear argument, and attention to conventions of writing, students begin to internalize the criteria and guide their own learning more productively. As an

example of how this process can operate, one study found that the introduction of such evaluation criteria produced significantly larger gains in individual learning scores as students spent more time discussing content, discussing the assignment, and evaluating their products.[60]

An analysis of hundreds of studies by British researchers Paul Black and Dylan Wiliam found that the regular use of these kinds of open-ended assessments with clear criteria to guide feedback, student revision, and teachers' instructional decisions—called "formative assessments"—produces larger learning gains than most instructional interventions that have been implemented and studied.[61]

For perhaps similar reasons, studies have found that teachers who were involved in scoring performance assessments with other colleagues were enabled to understand standards and the dimensions of high quality work more fully and to focus their teaching accordingly.[62]

These potentially positive consequences of performance assessments signal possibilities, however, not certainties. The quality of the assessments, how fairly they are constructed and how well they are implemented all influence the outcomes of assessment use and must be taken into account.

## Fairness

To make assessments fair and valid, it is important to eliminate features that can affect performance, but are not related to the specific knowledge and skills being measured. Problems can arise due to task wording and context, the mode of response required, and raters' attention to irrelevant features of responses or performances. As an example, in designing a performance assessment that measures students' mathematical problem solving, tasks should be set in contexts that are familiar to the population of students. If one or more subgroups of students are unfamiliar with a particular problem context, it will affect their performance, and hinder the validity and fairness of the score interpretations for those students. Similarly, if a mathematics performance assessment requires a high level of reading ability and students who have very similar mathematical proficiency perform differently due to differences in their reading ability, the assessment is measuring in part a construct that is not the target, namely, reading proficiency.

These issues are of particular concern for English language learners (ELLs). Although there are legitimate concerns about the linguistic demands of performance assessments, some studies have found that this is no more a problem with open-ended prompts than with traditional tests. For example, one recent study found that student responses to a writing prompt were less affected by student background variables, including English learner status, than were scores on a commercially developed language arts test, largely comprised of multiple-choice items.[63]

In fact, as testing expert Jamal Abedi explains, several components of well-designed performance assessments can make them more accessible to ELL students than multiple-choice assessments.[64] First, in many performance assessments, language is not the only medium of assessment. As shown above, many tasks incorporate graphical or hands-on elements as a different medium through which an ELL student can engage the content being tested and respond to the tasks. Drawing graphical representations of relationships as in some of the mathematics items shown earlier, or physically completing a science activity, such as sorting and categorizing substances, as in the NAEP science task described above, allows the student to demonstrate knowledge in multiple ways.

Second, multiple-choice tests often include responses that are plausibly correct, where the respondent is supposed to choose the "best" of several reasonable responses, or "distractor" choices that are intended to fool a respondent who is not reading carefully. These can often be particularly problematic for new English learners or students with disabilities who may know the material but not draw the fine linguistic distinctions that are required.

Finally, on performance assessments, raters can evaluate what respondents show about what they know, which allows them to credit students with the knowledge they can illustrate (for example, a solution on a mathematics problem), rather than getting only a count of right and wrong answers without information about the students' actual ability to read a passage or solve a problem. Particularly for special populations of students, scores on proxy items that are further from a direct performance can be deceiving, because they do not reveal whether the student understood all or part of the material but was confused by the instructions, format, or wording of the question or response choices, or may have made a minor error in the course of responding.

For these reasons, ELL students and students with disabilities sometimes perform better on performance tasks. This has proved the case in the New Jersey Special Review Assessments offered to students who fail the state high school exit exam. These open-ended performance tasks test the same standards and concepts as items on the multiple-choice test, but have proved more accessible to these populations of students. (See Figure 4, page 32, for an example of one task.)

In any kind of test, careful design can make a difference in validity for special populations. Jamal Abedi and his colleagues have identified a number of linguistic features of test items that slow readers down and increase the chances of misinterpretation. In their research, they have found that linguistic modifications that reduce the complexity of sentence structures and replace unfamiliar vocabulary with more familiar words increase the performance of English language learners, as well as other students in low- and average-level classes.[65] Linguistic modifications can be used in the design of performance assessments to help ensure a valid and fair assessment, not only for English language learners, but other students who may have difficulty with reading.

Figure 4, below, shows how a task from the New Jersey SRA can be made even more accessible with linguistic modifications, without altering the knowledge and skills being measured. These modifications reduce the length of the task instructions by more than 25%, eliminate conditional clauses and grammatical complexities (such as passive voice), and use more familiar words. While the modified task is easier to read, it still tests the same mathematics skills.

### Figure 4. New Jersey Department of Education, 2002-2003 SRA Mathematics Performance Assessment Task

ORIGINAL ITEM: Dorothy is running for president of the student body and wants to create campaign posters to hang throughout the school. She has determined that there are four main hallways that need six posters each. A single poster takes one person 30 minutes to create and costs a total of $1.50.

- What would be the total cost for Dorothy to create all the needed posters? Show your work.
- If two people working together can create a poster in 20 minutes, how much total time would Dorothy save by getting a friend to help her? Show your work.
- If Dorothy works alone for 3 hours, and is then joined by her friend, calculate exactly how much total time it will take to create all the necessary posters. Show your work.
- Omar, Dorothy's opponent, decided to create his posters on a Saturday and get his friends Janice and Beth to help. He knows that he can create 24 posters in 12 hours if he works alone. He also knows that Janice can create 24 posters in 10 hours and Beth can create 24 posters in 9 hours. How long will it take them, if all three of them work together to create the 24 posters? Round all decimals to the nearest hundredths. Show your work.
- When Omar went to purchase his posters, he discovered that the cost of creating a poster had increased by 20%. How many posters will he be able to create if he wants to spend the same amount of money on his posters as Dorothy? Justify your answer.

LINGUISTICALLY MODIFIED ITEM: You want to plant 6 roses in each of four large pots. Planting a single rose takes you 30 minutes and costs $1.50.

- What is the total cost to plant all roses? Show your work.
- With a friend's help, you can plant a rose in 20 minutes. How much total time do you save by getting a friend to help? Show your work.
- You work alone for 3 hours, and then a friend joins you. Now how much total time will it take to plant all the roses? Show your work.
- You can plant 24 roses in 12 hours. Your friend Al can plant 24 in 10 hours and your friend Kim can plant 24 roses in 9 hours. How long does it take the three of you to plant 24 roses together? Round all decimals to the nearest hundredths. Show your work.
- You just discovered that the cost of purchasing a rose increased by 20%. How many roses can you plant with the same amount of money that you spent when a rose cost $1.50? Justify your answer.

*Source: Abedi (2010)*

Finally, as Abedi points out, performance assessments provide stronger information for diagnostic purposes to help teachers decide how to continue instruction. They reveal more about students' processing skills and problem solving approaches, as well as their competence in particular areas, than do multiple-choice responses. They also simulate

learning activities, and as part of a system, may encourage teachers to use more complex assignments or formative assessments that resemble the tasks. These characteristics of performance assessments can be particularly beneficial for special needs student populations, including ELLs, because they provide more equitable learning opportunities and give teachers more information about how to support further learning.[66]

In general, fairness concerns can be addressed both by ensuring that all students gain access to rich assignments and learning opportunities—a goal supported by the use of classroom-based performance assessments—and by expert design of the tasks and rubrics and analyses of student thinking as they solve performance tasks. Use of universal design features, such as linguistic modifications, and pilot testing that leads to modifications of tasks based on features that contribute to subgroup differences also increases fairness.

## Feasibility

A host of feasibility issues have cropped up with performance assessment efforts in the United States, including the reliable production of high-quality tasks that are generalizable and scorable, managing the time and costs of scoring open-ended items, and ensuring that assessments can be well-implemented by teachers without overwhelming them. Feasible systems will also need to take advantage of efficiencies that have been discovered in assessment development, administration, and scoring, discussed here and further in the "Cost" section below.

**Creating State Capacity.** Experiences with performance assessments from U.S. states such as New York, Kentucky, Massachusetts, Vermont, and others provide a wealth of lessons about how to develop and administer assessments, involve teachers in professional development, and create systems that can support ongoing testing practices.

New York is an interesting case, given its 135-year history of assessments that include performance elements. Early in its history, all of New York's tests were open-ended essay examinations and problem solutions developed and scored by teachers, under state coordination. Today, the syllabus-based, end-of-course Regents exams in English, Global History & Geography, U.S. History & Government, mathematics, and science may be the closest U.S. equivalent to the British system. New York involves teachers in all aspects of the Regents testing process, from item and task development to review and training, as well as scoring. Teachers score on professional development days when they are released from teaching, and there are auditing systems that sample papers for re-scoring that may be followed by score adjustments and further training. A similar process in Kentucky, with substantial teacher training using benchmark performances and common scoring guides, has, with ongoing auditing, resulted in high levels of consistency in scoring, as well as more common understandings about high-quality work among teachers.

Systems that create consistency of local scoring across schools, and hence comparability of results, require substantial planning. States must commit to teacher training and, ideally at least at the beginning, to moderation sessions that bring teachers together to score, so that they can learn with one another. After training, states may decide to use only those certified scorers who demonstrate they can score reliably to common benchmarks. States must also provide a systematic approach to auditing scores, providing feedback, and adjustments needed to yield consistent scoring across a state. There is evidence that well-designed, consistent processes yield increasingly comparable scoring over time as the standards and processes are internalized by teachers and incorporated into instruction. The phasing in of performance assessment components of larger assessment systems should allow time for a state not only to refine and improve audit procedures, but also for local educators to internalize the state's general standards of performance.

Feasibility will also be enhanced by pursuing efficiencies in task design and scoring that have become available through recent research and development efforts and by using new technologies effectively.

**Efficiencies in Task Design.** There are a number of advances that can make performance assessments more efficient and effective as both measurement and teaching tools. For example, tasks can be designed to yield scores on different dimensions of performance in more than one content domain, which has practical as well as pedagogical appeal. Well-designed tasks that yield multiple scores reduce the time and costs of task development, test administration and scoring by raters.[67] Tasks that cut across content domains may also motivate a more integrated approach to teaching and learning.

For example, a text-based writing task in the Delaware state assessment is linked to a passage in the reading assessment, and student responses to the task are scored twice, once for reading and once for writing. The task below requires students to read an article prior to writing:

> The article you have just read describes some problems and possible solutions for dealing with grease. Do you think grease should be classified or labeled as a pollutant?
>
> Write a letter to the Environmental Protection Agency explaining whether or not grease should be classified as a pollutant. Use information from this article to support your position.[68]

This task is aligned to the reading and writing connection that occurs in instruction in Delaware classrooms. Students are first asked to read about a topic and then to use the information that they have read to support their position in their written product.

ETS researchers are currently developing methods that allow for accurate student level scores derived from mathematics and language arts performance assessments that are

administered on different occasions throughout the year.[69] This will not only allow for content representation across the performance assessments, but the assessments can be administered in closer proximity to the relevant instruction and information from any one administration can be used to inform future instructional efforts. This may allow assessments to provide both formative benefits and summative scores in a more integrated and efficient way that supports teachers.

**Technological Advances.** Advances in computer technology have made possible other efficiencies in measurement of performance. These advances have allowed for performance-based simulations that assess problem solving and reasoning skills in both formative assessments and in summative assessment programs. The most prominent large-scale assessments using computer-based simulations occur in licensure examinations in medicine, architecture, and accounting. In medicine, for example, the prospective doctor is presented with a description of the patient and then must manage the patient case by selecting history and physical examination options or by making entries into the patient's chart to request tests, treatments, and/or consultations. The condition of the patient changes in real time based on the patient's disease and the examinee's actions. The computer-based system generates a report that displays each action taken by the examinee and scores the appropriateness of the decisions made.

In addition to evaluating a student's responses, new technologies allow assessments to capture students' processes and strategies, along with their products. The computer can monitor and record the interactions a student has with tools used in solving a problem, assessing how students use these tools.[70] Teachers can use information on how a student arrived at an answer to guide instruction and monitor the progression of student learning.[71]

Computer technologies can also be used to create effective and efficient systems of on-line training, calibration, and scoring for performance items that will save time and money. It is now possible to have tasks uploaded by students and sent to teachers who will download and score them—often at home over a cup of coffee. These same teachers will have learned to score through on-line training and the computer will have certified their grading as reliable. The scored tasks they upload can be audited at any time to be sure that assessments are being scored consistently.

Finally, the use of automated scoring procedures for evaluating student performances in computer-based simulation tasks can provide an answer to the cost and time demands of human scoring. To ensure fairness in the brave new world of computer adaptive testing, it is important that examinees have had the opportunity to practice with the navigation system.[72] Advances in artificial intelligence can help reduce scoring burdens and enable faster turnaround, even though the requirements of complex programming do not yet produce much reduction in costs (see cost section below.)

At the same time, using teachers as scorers can reduce costs (see cost section below) while it helps improve instruction and communication. Teachers who are trained to score assess-

ments internalize the meaning of standards while they gain a better understanding of student thinking and misconceptions that can guide instruction. The rubrics used in scoring performance tasks support collaboration and learning among teachers by providing a unified language and common framework for teachers to recognize, debate, and assess student work.[73]

But teachers do not have to score 150 of the same items to gain these benefits. They might be asked to score a subsample of the tasks that are otherwise scored by computer, both as ongoing checks on the validity of computer scoring and as a learning opportunity for themselves. In the future, it will be possible to organize assessments that use a strategic blend of machine and human scoring that supports opportunities for teacher engagement while reducing burdens.

## Costs

Many policymakers have argued that the extensive use of performance items is too expensive in large-scale testing. A new study by the Assessment Solutions Group (ASG), a test development and consulting organization, demonstrates that it is possible to construct affordable, large-scale assessment systems that include a significant number of constructed-response items, reliably scored classroom-based performance tasks, and traditional multiple-choice questions for no more than the cost of the much-less-informative systems used today.

The ASG study shows that such systems can be designed for no more than the costs paid by an average state for today's tests (generally about $20 per pupil for English language arts and math tests), by making sound decisions that take advantage of the economies of scale that can be achieved when states join together in testing consortia, with new uses of technology in distributing and scoring standardized tests, and with thoughtful approaches to using teachers in the scoring of performance items.

Opportunity costs and benefits of assessment decisions are also important to consider. For example, studies have documented that high-stakes tests that are narrow in their focus and format reduce emphasis on strategies that support transferable learning, such as research, inquiry, and applications of knowledge to a variety of contexts; extended writing and defense of ideas; development of higher order thinking skills.[74] In addition, current testing systems provide very little textured information to help teachers improve learning: The tests deliver scores, rather than evidence of student work that can be closely examined and understood in terms of a learning continuum for specific skills. They reveal little of students' thinking, reasoning, and misconceptions, and almost nothing about their actual performance beyond the bounds of recognizing or guessing answers in items where they are already supplied.

Because the time used for testing and test preparation often does little to help students acquire transferable knowledge and skills, teachers often feel it is "lost" to instruction, rather than that it reflects, supports, and reinforces instruction. Data in the form of scores is supplied months after students have left school for the summer. Thus, the opportunity costs of current tests are fairly high and they produce relatively few benefits in terms of expanded knowledge about important student learning for students and teachers. The flip side of these opportunity costs illustrates some of the potential benefits accrued when using a performance assessment system that is information-rich in the ways that we have described.

At the same time, it is important to acknowledge that there are greater expenditures associated with the development and human scoring of open-ended items and tasks, especially when they need to be scored in ways that assure high levels of consistency.

Previous studies and cost estimates from current programs provide relatively similar estimates of these costs from multiple sources. For example, adjusted to current dollars,[75] cost estimates from several studies for development, administration, and scoring of assessment batteries that include significant performance components have ranged from about $30 to $50 per pupil, depending on how extensive the performance components are.[76] These estimates are mostly based on the practices used in the United States during the late 1980s and early 1990s. By comparison, as noted above, a largely multiple-choice test would cost about $20 to $25 per pupil to develop, administer, and score. The ratio of about 2 to 1 in terms of costs between performance-based and selected-response tests is also fairly constant across studies.

A 1993 GAO study highlighted potential cost savings based on the large spread in the cost of performance assessment, from $16 to $64 (with an average of $33). The upper end estimates are mostly from studies of small-scale experiments using specialized materials and equipment (e.g. science kits) that had to be delivered to schools.[77] This spread suggests the potential for economies of scale and experience in developing and implementing performance assessments. When including more students in test administrations, the study found that costs fell, with fixed costs distributed over a larger number of students.

Estimates for scoring individual performance events ranged from about $0.79 per student to over $8 per student, adjusted to current dollars (see Table 2 on page 38).[78] Performance assessments in European and Asian countries tend to cost considerably less, because of the more highly-developed routines and systems and the engagement of teachers in scoring.

Table 2: Scoring Estimates for Performance Assessments

| Assessment | Scoring | Cost (converted to 2009 dollars) | Study |
|---|---|---|---|
| *Connecticut Assessment of Educational Progress*: 25-minute essay | Twice holistically (Does not include staff costs for recruiting raters, procuring scoring sites, training table leaders, and selecting range finder papers and other categories). | $1.65 per student. | Baron, 1984 |
| *Research study for SAT*: 45-minute essay | Scored once holistically. | $0.79 to $2.14 per student. | Breland, Camp, Jones, Morris & Rock, 1987 |
| *California Assessment Program*: 45-minute essay | Scored twice. | $7.29 per student. | Hymes, 1991 |
| *College Board English Composition*: 20-minute essay | Scored twice. | $8.58 per student. | US Congress Office of Technology Assessment, 1992. |
| *Geometry Proofs* | Not reported. | $4.38 per student. | Stevenson, 1990 |
| *Kentucky Assessment*: on-demand tasks in a variety of subject areas | Total scoring time per student is 12 minutes. | $4.38 per student. | Hill & Reidy, 1993 |

The new ASG study finds similar cost comparisons—before including important efficiencies in its model that lower the cost of tests with significant performance elements.[79] ASG developed cost models providing an "apples-to-apples" comparison for two types of tests: a "typical" summative multiple-choice test with a few constructed response items and "high-quality assessment" that includes more constructed response items and new item types, such as performance events (relatively short curriculum-embedded tasks) and more ambitious performance tasks. In this study, ASG used empirically-based cost data and their cost model to determine the costs of each type of assessment.

Table 3 shows the number of multiple-choice and extended response items for each grade in a "typical" state system, followed by a reduced number of multiple-choice items and the addition of performance tasks and events in the new "high-quality" assessment (HQA). The models are based on an NCLB-type assessment system (English language arts and mathematics tests in grades 3-8 and grade 10).

## Table 3: Summative Assessment Design

| Summative Assessment | Item Counts | | | | |
|---|---|---|---|---|---|
| Mathematics | Multiple Choice | Short Constructed Response | Extended Constructed Response | Performance Event | Performance Task |
| Current Typical Assessment | 50 | 0 | 2 | 0 | 0 |
| High Quality Assessment | 25 | 2 (1 in grade 3) | 2 (0 in gr. 3, 1 in gr. 4) | 2 | 2 (0 in gr. 3, 1 in gr. 4) |
| **Summative Assessment** | **Item Counts** | | | | |
| English Language Arts | Multiple Choice | Short Constructed Response | Extended Constructed Response | Performance Event | Performance Task |
| Current Typical Assessment (Reading) | 50 | 0 | 2 | 0 | 0 |
| Current Typical Assessment (Writing)* | 10 | 0 | 1 | 0 | 0 |
| High Quality Assessment (Reading) | 25 | 2 (1 in gr. 3&4) | 2 (1 in gr. 3&4) | 2 | 1 |
| High Quality Assessment (Writing)* | 10 | 2 (1 in gr. 3&4) | 2 (1 in gr. 3&4) | 2 | 0 |

*Administered in grades 4, 7 and 10

The "high-quality assessment" is assumed to include two "performance events" and one or two "performance tasks" in addition to more constructed response items. The distinction we draw is between performance events, in which an individual student writes a response in a summative testing situation, such as an extended writing prompt completed within one or two class periods, and performance tasks that involve more ambitious work, such as a research project in which students prepare a paper and make a presentation. In the latter case, the assessment costs include developing the curriculum and materials to scaffold student progress throughout the task, as well as allowing for more scoring time.

This model estimates that a current multiple-choice testing battery in a typical state—including both reading and mathematics tests, plus benchmark testing—costs around $20 per pupil. In the same "typical" state, the high-quality assessment (HQA) including the same subjects and benchmark assessments would cost around $55 per pupil before cost-reduction strategies are applied. When such strategies are applied, the cost of performance assessments drops significantly. The strategies include:

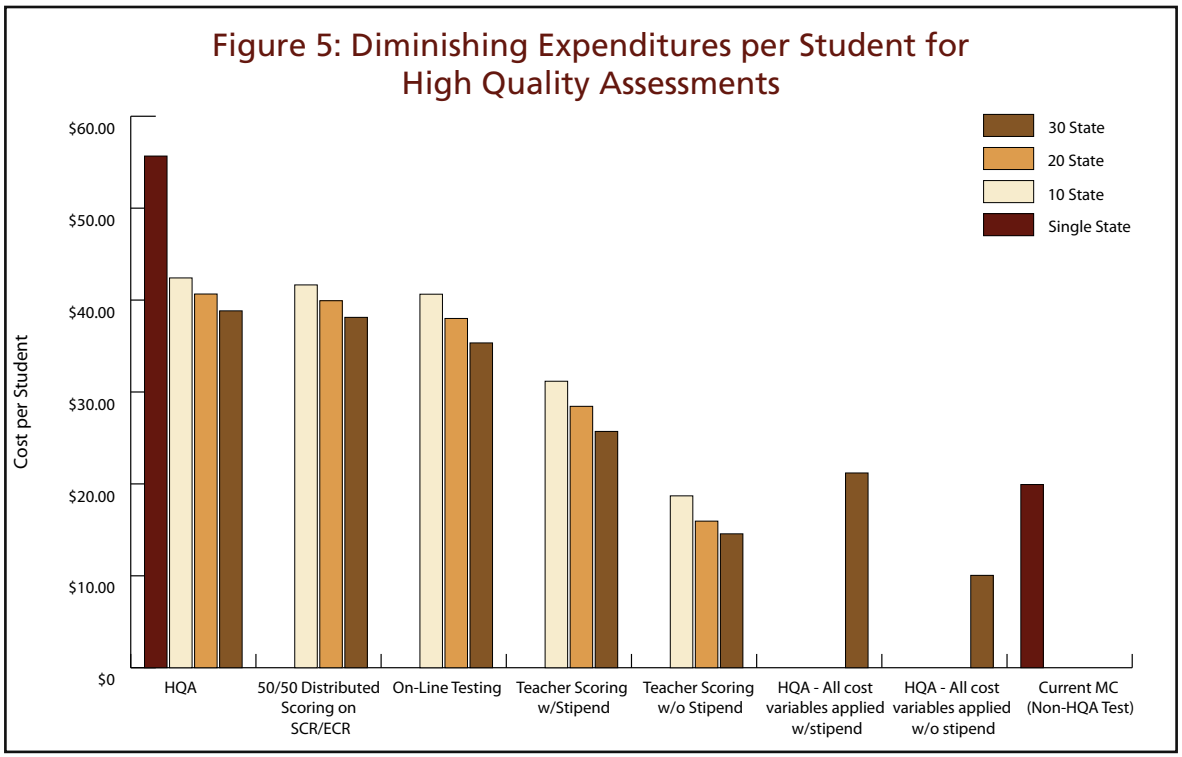
- *Participation in a consortium.* The model includes state consortium sizes of 10, 20, and 30 states. The use of a state consortium reduces costs by an average of $15 per pupil. The consortium approach represents a significant decrease in assessment cost.

- *Uses of technology.* Computers are used for online test delivery, to distribute human scoring of some of the open-ended items, and for automated scoring for certain constructed response items. Together these innovations account for cost savings of about $3 to $4 per pupil, and are likely to provide more reductions as efficiencies are developed in programming and using technology for these purposes.

- *Two approaches to the use of teacher-moderated scoring.* The final cost-reduction strategy, teacher-moderated scoring, can net both substantial cost reductions as well as the potential professional development benefits we have discussed earlier. ASG estimates two different models for teacher-moderated scoring, one a professional development model with no additional teacher compensation beyond that supported by the state or district for normal professional development days and the other assuming a $125 per day stipend to teachers. These strategies for using teachers as scorers reduce costs by an additional $10 to $20 per pupil (depending on whether teachers are paid or engaged as part of professional development).

Combining all possible cost-saving strategies results in a per-pupil cost for the high-quality assessment of just under $10 per pupil, about half the estimated cost of the "typical" summative state test (see Figure 5, page 41). Paying teachers a stipend to score increases this cost to about $21 per pupil, about the same cost that is paid for the typical multiple-choice testing battery a student takes today. The estimated time for scoring—based on contemporary evidence about teacher scoring time on similar tasks—decreases as scorers become more highly skilled.

Ultimately, a state could score assessments with these features by underwriting about two professional development days per teacher in the relevant subject areas. Whether incorporated into professional development costs, or paid directly through stipends, teacher involvement in scoring is not only a critical aspect of making performance assessments affordable, it is also critical to improve teaching and learning and to support instruction of higher-order thinking and performance skills.




While looking to economize, it is also important to put the costs of high-quality assessment into perspective. Whereas different choices about the size of consortia, the use of technology, and the approaches to scoring could put the costs of such a system in the range of $20 per student, the costs of other instructional interventions are much greater. For example, a study of three comprehensive school reform models with modest impact on achievement found that spending on professional development averaged almost $600 per pupil.[80] In the context of a tightly integrated teaching and learning system, the use of performance assessment offers an important and much more cost-effective method for influencing instructional practice.

## Benefits

Costs associated with a comprehensive assessment system are reflected in money and time. Of particular concern in this regard is the cost of using performance tasks that meet the requirements of valid and reliable assessment. While the development, use and scoring of performance tasks does require time and expertise, educators and policymakers in virtually all 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 have argued, increase students' achievement levels and readiness for college and careers.

The costs of performance assessment are also accompanied by the benefits of giving teachers incentives to engage in more ambitious kinds of teaching and more rigorous assignments in the classroom; more feedback about student thinking and performance; and models for crafting their own assessments, which can lead stronger instruction and learning.

In these respects, performance assessment is a Trojan Horse for instructional reform. As policy analyst Milbrey McLaughlin has noted, "It is exceedingly difficult for policy to change practice…. Change ultimately is a problem of the smallest unit."[81] Approaches to assessment that include teacher-scored, curriculum-embedded tasks reach into classrooms to stimulate change for teachers and students. By engaging teachers in the development, use, and scoring of these assessments, teachers can develop a shared conception of high quality instruction over time and through practice. They can internalize what counts as evidence of high quality student work. Teachers and administrators can develop knowledge of high quality assessment design principles and of how assessment should inform curriculum and instruction. They can also see first-hand which instructional patterns lead to particular characteristics of a performance.

The investment of resources for assessment-based scoring and professional development might be viewed as an opportunity to use professional development resources more wisely. The one-shot, "flavor-of-the-month" workshops that still constitute the bulk of American professional development leverage less knowledge and change in practice than engagement in involvement in developing and scoring assessments has been found to yield.[82] A coherent assessment system could re-direct a portion of professional development dollars toward more meaningful use, focused tightly on student learning, and create a paradigm shift about how to organize teachers' learning to support more effective instruction.

Finally, the engagement of educators in assessment development can also enable assessment designers to create more valid, reliable and fair assessments, because they gain fine-grained information about the contexts in which the assessments are used.

## Conclusion

Performance assessments have been an integral part of educational systems in most high-achieving countries and some states in the United States. Evidence suggests that the nature and format of the assessments affects the depth of knowledge and types of skills developed by students, and that performance assessments are better suited to assessing high level, complex thinking skills. Such assessments are also more likely to encourage the acquisition of such skills.[83] Furthermore, there is evidence that engaging teachers in these assessments can strengthen curriculum and instruction, and support more diagnostic teaching practices.

Our review suggests that large-scale testing in the United States could be enhanced by the thoughtful incorporation of adequately standardized performance assessments in tests that also include more analytically-oriented multiple-choice and constructed-response items. A more balanced and comprehensive assessment system could better represent academic content standards, particularly those describing higher-order, cognitively demanding performance; provide clearer signals to teachers about the kinds of student performances that are valued; and reduce pressures to mimic multiple-choice teaching in classroom instruction.

The appropriate role for performance assessments should be determined, in part, by an analysis of content standards. Such an analysis should reveal which standards are served well by which types of assessments. To the extent that the standards call for demonstration of higher-order, strategic skills, they may favor the use of performance assessment. Research reminds us that subject domains are different, and mastery of each domain is manifest in unique ways. Rich, thoughtful writing about literary texts can be observed under different circumstances and in different ways than rich, thoughtful scientific inquiry or rich, thoughtful mathematical modeling.

Much has been learned about how to develop, administer, and score performance tasks so that they provide valid, reliable, and fair evidence of student knowledge and skills, and so that they can engage teachers productively without creating overwhelming burdens. A new testing system will benefit from incorporating these practices and the psychometric standards underlying them. Next generation approaches will also benefit from new uses of technology that reduce costs and increase efficiencies by distributing and administering assessments, enabling new kinds of simulations and tasks, and strategically supporting both machine- and human-scoring.

There are costs and benefits associated with testing for accountability in whatever form that testing takes. We are used to the current high-stakes, multiple-choice model, but that does not mean it is cost-free or benefit-rich. Adopting performance assessments for some or all accountability testing will have trade-offs, and we are more likely to make wise decisions if we understand these trade-offs better.

In general, the addition of performance tasks will increase the overall cost of assessment, and the more complex the tasks the greater the additional costs. However, costs can be reduced significantly—to levels comparable to current spending on tests—if states join together in consortia, use technology wisely, create and score tasks in efficient ways, and involve teachers in scoring. If teachers' participation is conceptualized as part of their ongoing professional development, costs can be reduced even further and benefits for instruction can be increased. Furthermore, the cost of even the most performance-rich assessment system imaginable would be dwarfed by other spending on education—spending often directed in response to impoverished learning outcomes when high-quality tests are not in use.

Even if states spent $50 per pupil on assessments (more than twice our estimate of the costs of a balanced system), this would still be less than 10% of the cost of the kinds of interventions many are currently adopting to try to raise achievement, and far less than 1% of the costs of education overall. Given the power of assessment to change practice and guide learning, such an investment is miniscule relative to the benefits of improved assessment and the opportunity costs of the current approach.

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. Performance assessment should be a critical piece of the effort to achieve 21st century standards of learning.

# Appendix A

## Connecticut 9th / 10th Grade Science Assessment

### Acid Rain Task

Acid rain is a major environmental issue throughout Connecticut and much of the United States. Acid rain occurs when pollutants, such as sulfur dioxide from coal burning power plants and nitrogen oxides from car exhaust, combine with the moisture in the atmosphere to create sulfuric and nitric acids. Precipitation with a pH of 5.5 or lower is considered acid rain. Acid rain not only affects wildlife in rivers and lakes but also does tremendous damage to buildings and monuments made of stone. Millions of dollars are spent annually on cleaning and renovating these structures because of acid rain.

#### Your Task

Your town council is commissioning a new statue to be displayed downtown. You and your lab partner will conduct an experiment to investigate the effect of acid rain on various building materials in order to make a recommendation to the town council as to the best material to use for the statue. In your experiment, vinegar will simulate acid rain.

You have been provided with the following materials and equipment. It may not be necessary to use all of the equipment that has been provided.

| Suggested materials | Proposed building materials |
|---|---|
| containers with lids<br>graduated cylinder<br>vinegar (simulates acid rain)<br>pH paper/meter<br>safety goggles<br>access to a balance | limestone chips<br>marble chips<br>red sandstone chips<br>pea stone |

#### Designing and Conducting your Experiment

**1. In your words, state the problem you are going to investigate. Write a hypothesis using an "If... then... because ...." statement that describes what you expect to find and why.** Include a clear identification of the independent and dependent variables that will be studied.

**2. Design an experiment to solve the problem.** Your experimental design should match the statement of the problem and should be clearly described so that someone else could easily replicate your experiment. Include a control if appropriate and state which variables need to be held constant.

**3. Review your design with your teacher before you begin your experiment.**

**4. Conduct your experiment.** While conducting your experiment, take notes and organize your data into tables.

### Communicating your Findings

Working on your own, summarize your investigation in a laboratory report that includes the following:

- **A statement of the problem you investigated. A hypothesis ("If... then... because …" statement) that described what you expected to find and why.** Include a clear identification of the independent and dependent variables.

- **A description of the experiment you carried out.** Your description should be clear and complete enough so that someone could easily replicate your experiment.

- **Data from your experiment.** Your data should be organized into tables, charts and/or graphs as appropriate.

- **Your conclusions from the experiment.** Your conclusions should be fully supported by your data and address your hypothesis.

- **Discuss the reliability of your data and any factors that contribute to a lack of validity of your conclusions**. Also, include ways that your experiment could be improved if you were to do it again.

# Endnotes


1. Lyman, P. & Varian, H. R. (2003). *How much information*. School of Information Management and Systems, University of California, Berkeley. Retrieved February 11, 2010, from http://www.sims.berkeley.edu/how-much-info-2003/

2. McCain, T. & Jukes, I. (2001). *Windows on the future: Education in the age of technology.* Thousand Oaks, CA: Corwin Press.

3. Ng, P. T. (2008). Educational reform in Singapore: From quantity to quality. *Education Research on Policy and Practice, 7*, 5-15.

4. Silva, E. (2008). *Measuring the skills of the 21st century*. Washington, DC: Education Sector, p. 5.

5. The No Child Left Behind Act of 2001, (sec 1111 B 2 c (1)). Retrieved February 11, 2010, from http://www2ed.gov/policy/elsec/leg/esear02/pg2.html

6. U.S. General Accountability Office. (2009). *No Child Left Behind Act: Enhancements in the Department of Education's review process could improve state academic assessments*. Report GAO-09-911. Washington, DC: U.S. Government Accountability Office, p. 20.

7. Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

8. Stecher, B. (2010).

9. Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4-12 (SPA8954759)*. Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

10. Jones, B. D., & Egley R. J. (2004). Voices from the frontlines: teachers' perceptions of high-stakes testing. *Education Policy Analysis Archives*, *12*(39). Retrieved August 10, 2004, from http://epaa.asu.edu/epaa/v12n39/; Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: findings from a national survey of teachers.* Boston, MA: Boston College, National Board on Testing and Public Policy; Woody, E., Buttles, M., Kafka, J., Park, S., & Russell, J. (2004). *Voices from the field: Educators respond to accountability*. Berkeley, CA: Policy Analysis for California Education, University of California, Berkeley.

11. Shepard, L.A. (1996). *Measuring achievement: What does it mean to test for robust understandings?* William H. Angoff Memorial Lecture Series. Princeton, NJ: Educational Testing Services.

12. Shepard, L.A. (2008). Commentary on the National Mathematics Advisory Panel recommendations on assessment, *Educational Researcher*, *37*(9), 602-609.

13. Achieve Inc. (2004). *Do graduation tests measure up? A closer look at state high school exit exams. Executive summary.* Washington, DC: Author.

14. See, for example, Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading.* Santa Monica, CA: Science and Technology Policy Institute, RAND Corporation.

15. McMurrer, J. (2007). *Choices, changes, and challenges: Curriculum and instruction in the NCLB era*. Washington, DC: Center for Education Policy.

16. Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved February 11, 2010, from http://epaa.asu.edu/epaa/v8n41/, part 6, p. 10.

17. Southeast Center for Teacher Quality. (2004). High-stakes accountability in California: A view from the teacher's desk. *Teaching Quality Research Matters, 12,* 1-2. Retrieved September 2, 2004, from: http://www.teachingquality.org/ResearchMatters/issues/2004/issue12-Aug2004.pdf, p. 15.

18. Schmidt, W. H., Wang, H. C. & McKnight, C. (2005). Curriculum coherence: An examination of U.S. mathematics and science content standards from an international perspective. *Journal of Curriculum Studies*, *37*(5), 525-59.

19. OECD. (2007). *PISA 2006: Science competencies for tomorrow's world.* Volume 1: Analysis. Paris: Author. Retrieved February 11, 2010, from http://www.pisa.oecd.org/dataoecd/30/17/39703267.pdf.

20. Stecher, B. (2010), citing Madaus, G. F. & O'Dwyer, L. M. (1999). A short history of performance assessment. *Phi Delta Kappan* (May). 688-695.

21. Pecheone, R. L. & Kahl, S. (2010). *Developing performance assessments: Lessons learned.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

22. International Baccalaureate Organization. (2005, November). *IB Diploma Programme: English A1 – Higher Level – Paper 2*. Retrieved May 31, 2008, from http://www.ibo.org/diploma/curriculum/examples/samplepapers/documents/gp1_englisha1hl2.

23. Singapore Examinations and Assessment Board (n.d.), *Science Investigations*. Author.

24. Bloom, B. (1956). *Taxonomy of educational objectives, Handbook 1: Cognitive domain.* White Plains, NY: Longman.

25. Madaus, G. F. et al. (1992); Lomax, R. G., West, M. M., Harmon, M. C., Viator, K. A., & Madaus, G. F. (1995). *The impact of mandated standardized testing on minority students*. *The Journal of Negro Education, 64*(2), 171-185.

26. Pechone, R. L. & Kahl. S. (2010).

27. Hong Kong Eductional Assessment Authority (n.d.). *High School Physics Examination.* Author.

28. Hong Kong Educational Assessment Authority (2009). *School-based assessment: Changing the assessment culture***.** Retrieved October 1, 2009, from http://www.hkeaa.edu.hk/en/hkdse/School_based_Assessment/SBA/.

29. From information on the Ohio Department of Education OPAPP website: http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=9&ContentID=61383&Content=78805.

30. Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project (NCES 2007-466)*. Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved May 2, 2009, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466, p. 41.

31. Bennett, R. et al. (2007). p. 46.

32. Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate Learning Assessment: Facts and fantasies. *Evaluation Review*. *31*(5), 415-439; Klein, S., Freedman, D., Shavelson, R., & Bolus, R. (2008). Assessing school effectiveness. *Evaluation Review, 32,* 511-525.

33. Koretz, D., Mitchell, K. J., Barron, S. I., & Keith, S. (1996). *Final report: Perceived effects of the Maryland school performance assessment program.* CSE Technical Report. Los Angeles, CA: UCLA National Center for Research on Evaluation, Standards, and Student Testing.

34. Measured Progress. (2009). *Commonwealth Accountability and Testing System: 2007-08 Technical report*. Version 1.2. Retrieved February 20, 2010 from http://www.education.ky.gov/KDE/Administrative+Resources/Testing+and+Reporting+/Kentucky+School+Testing+System/Accountability+System/Technical+Manual+2008.htm.

35. This section relies heavily on Suzanne Lane's paper developed for this project: Lane, S. (2010). *Performance assessment: The state of the art*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

36. Lane, S. & Stone, C. A. (2006). Performance assessments. In B. Brennan (Ed.), *Educational Measurement*. Westport, CT: American Council on Education & Praeger.

37. Baker, E. L. (2007). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. *Educational Assessment*, *12*(3&4), 179-194.

38. Lane, S. & Stone, C. A. (2006).

39. Lane, S. (2010).

40. Niemi, D., Wang, J., Steinberg, D. H., Baker, E. L., & Wang, H. (2007). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment, 12*(3 & 4), 215-238, p. 199.

41. Chi, M. T. H., Glaser, R., & Farr, M. J. (Eds.). (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum; Ericsson, K. A. & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: The MIT Press.

42. Measured Progress. (2009). *New England Common Assessment Program 2008-2009 technical report,* Dover, NH. Retrieved September 30, 2009, from http://www.ride.ri.gov/assessment/DOCS/NECAP/Tech_Manual/2008-09_TechReport/2008-09_NECAP_TechReport.pdf; Measured Progress. (2009). *New England Common Assessment Program 2008-2009 technical report,* Dover, NH. Retrieved September 30, 2009, from http://www.ride.ri.gov/assessment/DOCS/NECAP/Tech_Manual/2008-09_TechReport/2008-09_NECAP_TechReport.pdf; New York State Education Department. (1996, October 10). *Report of the Technical Advisory Group for the New York State Assessment Project.* Author.

43. Lane, S. (2010).

44. Collegiate Learning Assessment. (2010). *CLA: Returning to learning.* Retrieved March 15, 2010, from http://www.collegiatelearningassessment.org/

45. Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project (NCES 2007-466)*. Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved May 2, 2009, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466; Deane, P. (2006). Strategies for evidence identification through linguistic assessment of textual responses. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 313-362). Mahwah, NJ: Lawrence Erlbaum Associates.

46. Klein et al. (2007).

47. Mislevy, R. (1993). Foundations of a new test theory. In N. Frederiksen, R. Mislevy, & I. Bejar, (Eds.), *Test theory for a new generation of tests* (pp. 19-40). Hillsdale, NJ: Erlbaum..

48. L. Shepard (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do.* San Francisco, CA: Jossey-Bass.

49. National Research Council (2006). *Systems for state science assessment*. M.R. Wilson & M.W. Bertenthal (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press; For a description of England's *Assessing Pupils' Progress* program, see Darling-Hammond, L. & Wentworth, V. (2010). *Benchmarking learning systems: Student performance assessment in international context*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

50. Kane, M. T. (2006). Validation. In B. Brennan (Ed.), *Educational measurement*. Westport, CT: American Council on Education & Praeger; Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-104). New York: American Council on Education and Macmillan.

51. Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

52. Chapman, C. (1991, June). *What have we learned from writing assessment that can be applied to performance assessment?* Presentation at ECS/CDE Alternative Assessment Conference, Breckenbridge, CO.; Herman, J. L., Klein, D. C. D., Heath, T. M., & Wakai, S. T. (1995). *A first look: Are claims for alternative assessment holding up?* CSE Technical Report. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing; Koretz, D., Mitchell, K. J., Barron, S. I., & Keith, S. (1996). *Final Report: Perceived effects of the Maryland school performance assessment program.* CSE Technical Report. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing; Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey of Kentucky teachers of mathematics and writing.* CSE Technical Report. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing; Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998, Summer). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, *20*: 95-113; Lane, S., Stone, C. A., Parke, C. S., Hansen, M. A., & Cerrillo, T. L. (2000, April). *Consequential Evidence for MSPAP from the Teacher, Principal and Student Perspective*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA; Koretz, D., Stetcher, B., & Deibert, E. (1992). *The Vermont portfolio program: Interim report on implementation and impact, 1991-92 school year.* Santa Monica, CA: The RAND Corporation; Stecher, B., Baron, S., Chun, T., & Ross, K. (2000). *The effects of the Washington state education reform on schools and classroom.* CSE Technical Report. Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing; Darling-Hammond, L. & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. *Yearbook of the National Society for the Study of Education. 104*(2), 289-319.

53. Frederiksen, J. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher,18*(9): 27-32; National Council on Education Standards and Testing. (1992). *Raising standards for American education. A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people.* Washington, DC: U.S. Government Printing Office, Superintendent of Documents, Mail Stop SSOP; Resnick, L. B. & Resnick, D. P. (1982). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M.C. O'Conner (Eds.). *Changing assessment: Alternative views of aptitude, achievement and instruction* (pp. 37-55), Boston: Kluwer Academics.

54. Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment, 8*(4), 279-315; Stecher, B., Barron, S., Chun, T., & Ross, K. (2000, August). *The effects of the Washington state education reform in schools and classrooms* (CSE Tech. Rep. NO. 525). Los Angeles, CA: UCLA National Center for Research on Evaluation, Standards and Student Testing; Stein, M. K. & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, *2*(1), 50-80; Stone, C. A. & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16*(1), 1-26.

55. Newmann, F. M., Marks, H. M., & Gamoran, A. (1996). Authentic pedagogy and student performance. *American Journal of Education, 104*(8), 280-312.

56. Lane, S. et al. (2002); Parke, C. S., Lane, S., & Stone, C. A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation, 12*(3), 239-269; Stone, C.A. & Lane, S. (2003).

57. Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3-16.

58. Lane, S. (2010).

59. Lane, S. (2010).

60. Cohen, E. G., Lotan, R. A., Abram, P. L., Scarloss, B. A., & Schultz, S. E. (2002). Can groups learn? *Teachers College Record, 104*(6), 1045-1068; see also Barron, B., Schwartz, D.L., Vye, N.J., Moore, A., Petrosino, T., Zech, L. & Bransford, J.D. (1998). Doing with understanding: Lessons from research on problem and project-based-learning. *Journal of Learning Sciences*, 7(3&4), 271-311.

61. Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80,* 139-148.

62. Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action*. New York: Teachers College Press; Falk, B., & Ort, S. (1997, April). *Sitting down to score: Teacher learning through assessment*. Presentation at the annual meeting of the American Educational Research Association, Chicago; Goldberg, G. L. & Rosewell, B.S. (2000). From perception to practice: The impact of teachers' scoring experience on the performance based instruction and classroom practice. *Educational Assessment, 6*, 257-290; Murnane, R. & Levy, F. (1996). *Teaching the New Basic Skills*. NY: The Free Press.

63. Goldschmidt, P., Martinez, J. F., Niemi, D., & Baker, E. L. (2007). Relationships among measures as empirical evidence of validity: Incorporating multiple indicators of achievement and school context. *Educational Assessment, 12*(3 & 4), 239-266.

64. Abedi, J. (2010). *Performance assessments for English language learners*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

65. Abedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234; Abedi, J., Lord, C. Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16-26.

66. Abedi, J. & Herman, J. L. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. *Teacher's College Record 112*(3), 723-746.

67. Goldberg, G. L., & Roswell, B. S. (2001). Are multiple measures meaningful? Lessons from a statewide performance assessment. *Applied Measurement in Education, 14,* 125-150; Lane, S. (2010).

68. Delaware Department of Education. (2005). *Text-based writing item sampler.* Retrieved July 5, 2009, from http://www.doe.k12.de.us/AAB/files/Grade%208%20TBW%20-%20Greaseaters.pdf, p. 5.

69. Bennett, R. et al. (2007).

70. Vendlinski, T. P., Baker, E. L., & Niemi, D. (2008). *Templates and objects in authoring problem-solving assessments.* (CRESST Tech. Rep. No. 735). Los Angeles: University of California, National Center Research on Evaluation, Standards, and Student Testing (CRESST).

71. Bennett, R. et al. (2007).

72. Bennett, R. E. (2006). Moving the field forward: Some thoughts on validity and automated scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Behar (Eds.), *Automated scoring of complex tasks in computer-based testing,* (pp. 403-412). Hillside, NJ: Erlbaum; Bennett, R. E. & Gitomer, D. H. (in press). Transforming K-12 Assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century.* New York: Springer.

73. Pechone, R. & Kahl, S. (2010).

74. Feuer, M. J. (2008). Future directions for educational accountability: Notes for a political economy of measurement. In *The future of test-based educational accountability* . K. E. Ryan & L. A. Shepard (Eds.), New York: Routledge; Picus, L., Adamson, F., Montague, W., Owens, M. (2010). *A new conceptual framework for analyzing the costs of performance assessment*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education; Darling-Hammond, L. & Rustique-Forrester, E. (2005).

75. All calculations of the cost in 2009 dollars were made using the CPI calculator available from the U.S. Bureau of Labor Statistics at http://data.bls.gov/cgi-bin/cpicalc.pl.

76 For a review, see Picus, L. et al. (2010). See also U.S. General Accounting Office. (1993). *Student extent and expenditures, with cost estimates for a national examination*. Report GAO/PEMD-93-8. Washington, DC: Author.

77. U.S. GAO. (1993); Stecher, B. (1995). *The cost of performance assessment in science: The RAND perspective.* Presentation at the annual conference of the National Council on Measurement in Education, San Francisco.

78. Table adapted from Hardy, R. A. (1995). Examining the costs of performance assessment. *Applied Measurement in Education*, *8*(2), 121-134.

79. Topol, B., Olson, J., & Roeber, E. (2010). *The cost of new higher quality assessments: A comprehensive analysis of the potential costs for future state assessments.* Stanford, CA: Stanford Center for Opportunity Policy in Education.

80. Odden, A., Goertz, M., Goetz, M., Archibald, S., Gross, B., Weiss, M., & Mangan, M. T. (2008). The cost of instructional improvement: resource allocation in schools using comprehensive strategies to change classroom practice. *Journal of Education Finance*, *33*(4), 381-405, p. 399.

81. McLaughlin, M. (2005). Listening and learning from the field: Tales of policy implementation and situated practice. In Lieberman, A. (Ed.), *The roots of educational change*, (58-72), p. 60.

82. For a recent review of the status and effectiveness of professional development programs in the United States, see Wei, R. C., Darling-Hammond, L., Andree, A., Richardson, N., Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. Dallas, TX. National Staff Development Council.

83. This finding and other key points in this section are drawn from Brian Stecher's paper commissioned for this project: Stecher (2010).

Linda Darling-Hammond, Co-Director
*Stanford University Charles E. Ducommun Professor of Education*

Prudence Carter, Co-Director
*Stanford University Associate Professor of Education and (by courtesy) Sociology*

Carol Campbell, Executive Director


**Stanford Center for Opportunity Policy in Education**
**Barnum Center, 505 Lasuen Mall**
**Stanford, California 94305**
**Phone: 650.725.8600**
**scope@stanford.edu**

**http://edpolicy.stanford.edu**