chapter
1

Beyond the Bubble Test

How Performance Assessment Can Support Deeper 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, US governors and chief state school officers recently issued a set of Common Core State Standards that aim to outline internationally benchmarked concepts and skills needed for success in today's and tomorrow's world.[1] The standards, which intend to create "fewer, higher, and deeper" curriculum goals, are meant to ensure that students are college and career ready.

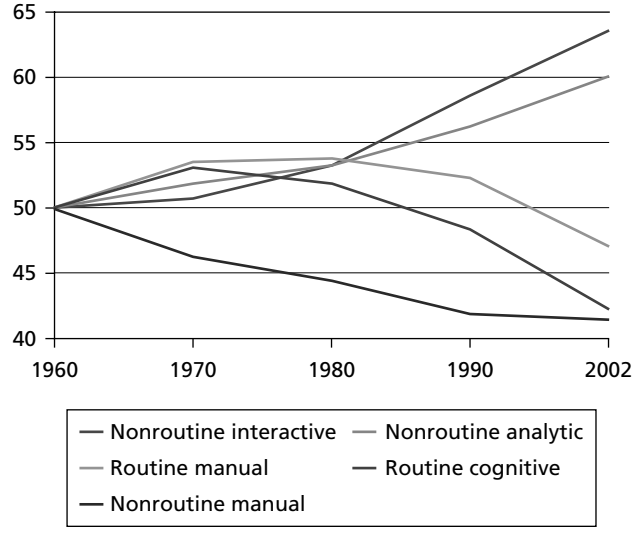
Changes in teaching and testing are profoundly implicated by this goal. Genuine readiness for college and twenty-first-century careers, as well as participation in today's democratic society, requires, as President Obama has noted, much more than "bubbling in" on a test. Students need to be able to find, evaluate, synthesize, and use knowledge in new contexts, frame and solve nonroutine 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 twenty-first-century skills that 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 now are the nonroutine interactive skills that are important for collaborative invention and problem solving. (See figure 1.1.)

In part, this is because knowledge is expanding at a breathtaking pace. Researchers at the University of California, Berkeley estimated that from

Figure 1.1
How the Demand for Skills Has Changed: Economy-Wide Measures of Routine and Nonroutine Task Input



Source: Murnane, R., & Levy, F. (1996). *Teaching the new basic skills: Principles for educating children to thrive in a changing economy*. New York, NY: Free Press.

1999 to 2002, the amount of new information produced in the world exceeded the amount produced in the entire history of the world previously.[2] The amount of new technical information was doubling every two years at that time. It is now doubling annually.[3]

As a consequence, a successful education can no longer be organized by dividing a set of facts into the twelve 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 their 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, its education minister at the time, 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."[4]

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.[5] 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

During the 1990s, the advent of standards-based reform intended to move the education system toward twenty-first-century skills led many states and districts to develop systems that included open-ended assessments reflecting central ideas and performances in the disciplines.[6] These products—essays, mathematics tasks, research papers, scientific investigations, literary analyses, artistic exhibitions—were scored by teachers trained to evaluate the responses reliably. Studies found that these assignments improved the quality of instruction in states ranging from

California to Kentucky, Maine, Maryland, Vermont, and Washington[7] and improved achievement on both traditional standardized tests and more complex performance measures.[8]

However, performance assessments encountered rocky shoals in the United States as a function of implementation challenges, scoring costs, and conflicts with the requirements of No Child Left Behind (NCLB), the federal education law launched in 2002.[9] While NCLB introduced laudable goals for improving and equalizing school achievement, its approach to test-based accountability ultimately reduced the quality of assessments used by nearly all states.

Many states discontinued the assessments they had developed in the 1990s that required writing, research, and extended problem solving and replaced them with multiple-choice and short-answer tests. States abandoned performance assessments in part because of constraints placed on the types of tests approved by the US Department of Education as it reviewed state plans following NCLB passage, despite language supporting complex assessments in the law.

In addition, states had to adjust to strict NCLB testing time lines and a dramatic increase in costs with the law's requirement for every-child, every-year testing. Together these forces led to tests that rely heavily on multiple-choice questions that can be scored rapidly and inexpensively by machine. By their very nature, such tests 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 US Congress, reported in 2009 that states' reliance on multiple-choice testing increased sharply in the NCLB era to achieve inexpensive scoring within tight time frames. Meanwhile, state education officials "reported facing trade-offs between efforts to assess highly complex content and to accommodate cost and time pressures."[10]

Indeed, a 2012 study by the RAND Corporation found that fewer than 2 percent of mathematics items and only about 20 percent of English language arts items on current state tests measure higher-order skills.[11] These skills, such as the abilities to analyze, synthesize, compare, connect, critique, hypothesize, prove, or explain ideas, are, in testing parlance, those represented at depth of knowledge levels 3 and 4 in the Webb taxonomy

that classifies cognitive demand.[12] Levels 1 and 2 represent lower-level skills of recall, recognition, and use of routine procedures.

This study, which echoes the findings of other research,[13] is particularly worrisome, since these states were selected because their standards were viewed as especially rigorous. The RAND study found that the level of cognitive demand was severely constrained by the dominance of multiple-choice questions, which are rarely able to measure higher-order skills. Thus, the ambitious expectations found in state standards documents are frequently left unmeasured.

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 that knowledge in different ways.[14] 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,[15] not only on other tests, but, more important, in the real world.

As Brian Stecher has noted, multiple-choice tests do not reflect the nature of performance in the real world, which rarely presents people with structured choices.[16] With the possible exception of a few game shows, we demonstrate our ability in the real world by applying knowledge and skills in settings where there are no predetermined 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 this person's 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. Rarely does a citizen or a student have to choose among four distinct alternatives.

This would not be a major problem if tests were not used for more and more high-stakes decisions. Currently federal, state, and local governments have created policies that use test scores to determine student promotion from grade to grade, program placements, and graduation;

teacher tenure, continuation, and compensation; and school rewards and sanctions, including loss of funds and closure.

A long line of research has shown that—for good or ill—tests used for decision-making purposes can drive curriculum and instruction in ways that mimic both the content and the format of tests.[17] 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.[18]

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."[19]

The NCLB school accountability model and the standardized testing that undergirds it have not catalyzed the law's pursuit of twenty-first-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.[20] Yet as the Center on Education Policy has documented, NCLB has narrowed the curriculum for many students, encouraging teachers to

focus not only on the content but also the format of the tests at the expense of other essential kinds of learning.[21]

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.[22]

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.[23]

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.[24] Ironically, states that test more topics in a grade level may encourage more superficial coverage, leading to less solid learning.

It is therefore not surprising that while student scores have risen sharply on the state tests used for accountability purposes under NCLB, scores have not increased commensurately on tests that gauge students' ability to apply knowledge to novel problems, such as the Program for International Student Assessment (PISA). US scores on PISA changed little between 2000 and 2012. After nearly a decade of test-based accountability, in 2012 the United States ranked thirty-second among member countries of the Organization for Economic Cooperation and Development in mathematics, twenty-third in science, and twenty-first in reading. Furthermore, US students scored lowest on the problem-solving tasks.[25]

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.

By contrast, with a few exceptions, testing in most US states has been less focused on higher-order skills during the NCLB era than it was in the 1990s, even though it has increasingly functioned as a primary influence on curriculum and classroom instruction. Thus, while students in high-achieving nations are engaged in the kind of learning aimed at preparing to succeed in college and in the modern workplace, students in the United States have been drilling for multiple-choice tests that encourage recognition of simple right answers rather than the production of ideas.

EMERGING OPPORTUNITIES FOR BETTER ASSESSMENT

In addition to the new Common Core State Standards (CCSS) in English language arts and mathematics, a consortium of states has developed a set of Next Generation Science Standards (NGSS) that also aim for more intellectually ambitious learning and teaching. These new standards offer an opportunity to address the fundamental misalignment between our aspirations for students and the assessments we use to measure whether they are achieving those goals. The United States has a chance to create a new generation of assessments that build on NCLB's commitment to improve the education of traditionally underserved groups of students, while measuring a wider range of skills and expanding instruction 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 reasoning, 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.

The recently released report of the Gordon Commission (2013), sponsored by the Educational Testing Service and written by the nation's leading experts in curriculum, teaching, and assessment, described the most critical objectives this way:[26]

> To be helpful in achieving the learning goals laid out in the Common Core, assessments must fully represent the competencies that the increasingly complex and changing world demands. The best assessments can accelerate the acquisition of these competencies if they guide the actions of teachers and enable students to gauge their progress. To do so, the tasks

> and activities in the assessments must be models worthy of the attention and energy of teachers and students. The Commission calls on policy makers at all levels to actively promote this badly needed transformation in current assessment practice. . . . The assessment systems [must] be robust enough to drive the instructional changes required to meet the standards . . . and provide evidence of student learning useful to teachers.

At least a modest step in this direction is intended by the two multistate consortia creating assessments to evaluate the CCSS—the Partnership for Assessing Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC). PARCC and SBAC assessments, to be launched in 2014–2015, will increase the use of constructed response items and performance tasks.

The plans for the new consortia assessments suggest they will increase cognitive demand, offering more tasks that require students to analyze, critique, evaluate, and apply knowledge. An analysis of the Content Specifications for the Smarter Balanced Assessment Consortium found, for example, that 68 percent of the assessment targets in English language arts and 70 percent of those in mathematics intend to tap these higher-level skills (Herman & Linn, 2013).[27] The sample tasks released by the two consortia include performance tasks that encourage instruction aimed at helping students acquire and use knowledge in more complex ways. (See exhibits 1.1 and 1.2.)

Exhibit 1.1 Mathematics Performance Tasks

SBAC Sixth-Grade Task: Planning a Field Trip

Classroom Activity: The teacher introduces the topic and activates students' prior knowledge of planning field trips by:

- Leading students in a whole class discussion about where they have previously been on field trips or other outings, with their school, youth group, or family.

- Creating a chart showing the class's preferences by having students' first list and then vote on the places they would most like to go on a field trip, followed by whole class discussion on the top choices.

Student Task: Individual students:

- Recommend where their class should go on a field trip, based on their analysis of the class vote.
- Determine the per-student cost of going on a field trip to three different locations, based on a chart showing the distance and entrance fees for each option, plus formula for bus charges.
- Use information from the cost chart to evaluate a hypothetical student's recommendation about going to the zoo.
- Write a note to their teacher recommending and justifying which field trip the class should take, based on an analysis of all available information.

PARCC High School Task: Golf Balls in Water

Part A: Students analyze data from an experiment involving the effect on the water level of adding golf balls to a glass of water in which they:

- Explore approximately linear relationships by identifying the average rate of change.
- Use a symbolic representation to model the relationship.

Part B: Students suggest modifications to the experiment to increase the rate of change.

Part C: Students interpret linear functions using both parameters by examining how results change when a glass with a smaller radius is used by:

- Explaining how the *y*-intercepts of two graphs will be different.
- Explaining how the rate of change differs between two experiments.
- Using a table, equation, or other representation to justify how many golf balls should be used.

Source: Herman, J. L., & Linn, R. L. (2013). *On the road to assessing deeper learning: The status of Smarter Balanced and PARCC assessment consortia* (CRESST Report No. 823). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. See also http://ccsstoolbox.agilemind.com/parcc/about_highschool_3834.html and http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/performance-tasks/fieldtrip.pdf.

Exhibit 1.2 English Language Arts Performance Tasks

PARCC Seventh-Grade Task: Evaluating Amelia Earhart's Life

Summary Essay: Using textual evidence from the *Biography of Amelia Earhart*, students write an essay to summarize and explain the challenges Amelia Earhart faced throughout her life.

Reading/Pre-Writing: After reading *Earhart's Final Resting Place Believed Found*, students:

- Use textual evidence to determine which of three given claims about Earhart and her navigator, Noonan, is the most relevant to the reading.
- Select two facts from the text to support the claim selected.

Analytic Essay: Students:

- Read a third text called *Amelia Earhart's Life and Disappearance*.
- Analyze the evidence presented in all three texts concerning Amelia Earhart's bravery.
- Write an essay, using textual evidence, analyzing the strength of the arguments presented about Amelia Earhart's bravery in at least two of the texts.

SBAC Eleventh-Grade Task: Nuclear Power—Friend or Foe?

Classroom Activity: Using stimuli such as a chart and photos, the teacher prepares students for Part 1 of the assessment by leading students in a discussion of the use of nuclear power. Through discussion:

- Students share prior knowledge about nuclear power.
- Students discuss the use and controversies involving nuclear power.

Part 1: Students complete reading and pre-writing activities in which they:

- Read and take notes on a series of Internet sources about the pros and cons of nuclear power.
- Respond to two constructed-response questions that ask students to analyze and evaluate the credibility of the arguments in favor and in opposition to nuclear power.

Part 2: Students individually compose a full-length, argumentative report for their congressperson in which they use textual evidence to justify the position they take pro or con on whether a nuclear power plant should be built in their state.

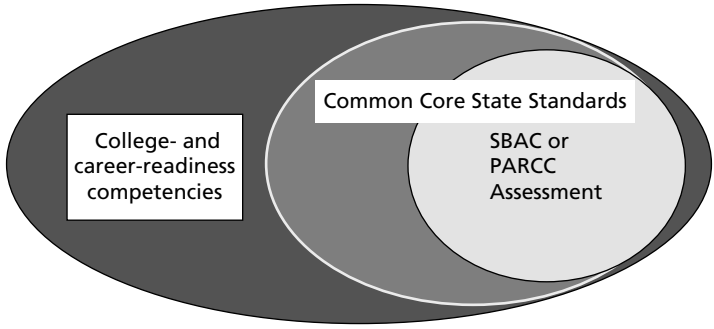
Source: Herman & Linn (2013). See also http://www.parcconline.org/samples/english-language-artsliteracy/grade-7-elaliteracy. http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/performance-tasks/nuclear.pdf

Even these more ambitious assessment tasks, each conducted in a one- or two-day session, will not measure all of the CCSS skills—such as extended writing and research, oral communications, collaboration, and uses of technology for investigation, modeling solutions to complex problems, and multimedia presentation. Furthermore, as shown in figure 1.2, the CCSC represent only a subset of the full range of college- and career-readiness expectations, including knowledge and skills in content areas beyond English language arts and math, traits such as resourcefulness and perseverance, key learning skills and cognitive strategies, and transition knowledge and skills.

These skills are evaluated in a growing number of countries, which require students to design and conduct complex projects that may take many days or weeks to complete and require considerable student planning, perseverance, and problem solving. The products of this work are evaluated by teachers, using a moderation process that trains them to score reliably, and are included in examination results.

Some states and districts, such as those belonging to the Innovation Lab Network, coordinated by the Council for Chief State School Officers, plan to introduce even more extensive performance assessments to complement the consortium tests. These may include longer-term tasks that require students to undertake investigations over multiple weeks and could result in a range of products (engineering designs, built objects, spreadsheets,

Figure 1.2
Competencies to Be Developed and Assessed



research reports) presented in a variety of forms, including oral, graphic, and multimedia presentations.

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 over 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 recent advances in testing technology. 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 book is to provide an analysis of the prospects and challenges of sustaining performance assessment on a large scale. We describe the history and current uses of performance assessments in the United States and abroad and summarize the results of decades of research on advances in, effects of, and costs of performance assessments. We hope that this work will inform the efforts of policymakers, practitioners, and researchers seeking more productive assessment and accountability models for education—models that can encourage and assess the advanced knowledge and skills that have become critically important for students and can support educators' capacity to develop them.