chapter
1

Introduction

The Rationale and Context for Performance Assessment

Linda Darling-Hammond

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

Over the past decade, the effects of US test-driven accountability practices have been the focus of intense debate. Disappointment about the performance of US students on international tests, concern about the nation's global competitiveness, and questions about our students' readiness to enter college and the workforce have led to another wave of efforts to significantly reform American education.

A recurring theme in the public debate among educators, business leaders, elected officials, and community members is the need for schools to focus on a new and expanded skill set in order for American students to compete in a digital

age. The discourse centers on the need to measure the core knowledge and higher-order skills critical to postsecondary learning and career success. In particular, growing emphasis on critical thinking, analytical reasoning, and communication skills has led to calls for a more balanced assessment system that includes authentic measures of student performance.

The United States is not alone in this pursuit. Reform of educational standards and assessments has been a constant theme in nations around the globe. New curriculum approaches and assessments have recently been adopted in Singapore, Hong Kong, and the United Kingdom, among many others. For example, as Singapore prepared to overhaul its assessment system, its education minister at that 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" (Ng, 2008).

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 the Common Core State Standards in English language arts and mathematics that aim to outline internationally benchmarked concepts and skills needed for success in today's world. The standards, adopted by forty-five states and three territories, intend to create "fewer, higher, and deeper" curriculum goals that ensure that students are college and career-ready (http://www.corestandards.org).

This goal has profound implications for teaching and testing. Genuine readiness for college and 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.

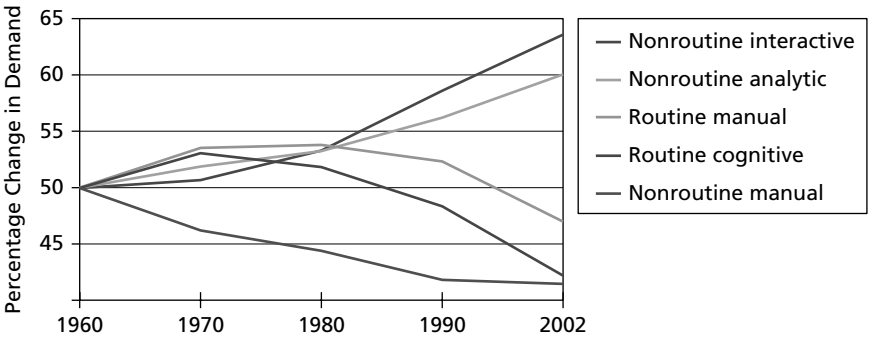
The recently released report of the Gordon Commission on Future Assessment in Education (2013), sponsored by the Educational Testing Service and written by the nation's leading experts in curriculum, teaching, and assessment, described the most critical objectives this way:

> 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.
>
> New assessments must advance competencies that are matched to the era in which we live. Contemporary students must be able to evaluate the validity and relevance of disparate pieces of information and draw conclusions from them. They need to use what they know to make conjectures and seek evidence to test them, come up with new ideas, and contribute productively to their networks, whether on the job or in their communities. As the world grows increasingly complex and interconnected, people need to be able to recognize patterns, make comparisons, resolve contradictions, and understand causes and effects. They need to learn to be comfortable with ambiguity and recognize that perspective shapes information and the meanings we draw from it. At the most general level, the emphasis in our educational systems needs to be on helping individuals make sense out of the world and how to operate effectively within it. Finally, it is also important that assessments do more than document what students are capable of and what they know. To be as useful as possible, assessments should provide clues as to why students think the way they do and how they are learning as well as the reasons for misunderstandings. (p. 7)

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 Richard Murnane and Frank Levy (1996) 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 nonroutine

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



Source: Murnane and Levy (1996).
Organization for Economic Cooperation and Development (2012), *Lessons from PISA for Japan, Strong Performers and Successful Reformers in Education, OECD Publishing,* http://dx.doi.org/10.1787/9789264118539-en

interactive skills that require 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, 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 (Lyman & Varian, 2003). The amount of new technical information was doubling every two years at the turn of the century (McCain & Jukes, 2001) and is now doubling every year.

As a consequence, a successful education can no longer be organized by dividing a set of static facts into the twelve years of schooling, to be doled out to students 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.

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 to solve complex problems (Silva, 2008). Remembering pieces of knowledge is no

longer the highest priority for learning; what counts is what students can do with the knowledge they acquire.

THE NEED FOR PERFORMANCE ASSESSMENTS

In order to encourage and measure this kind of learning, performance assessments that reflect how students acquire and use knowledge to solve real-world problems are increasingly needed. Many high-achieving nations have developed national or state curriculum guidance that incorporates performance assessments that require students to solve complex real-world problems and defend their ideas orally and in writing. These assessments—which include research projects, science investigations, mathematical and computer models, and other products—are mapped to the syllabus and the standards for the subject and are selected because they represent critical skills, topics, and concepts. They are generally designed, administered, and scored by teachers in local schools.

These nations recognize that classroom-embedded performance tasks allow the development and assessment of more complex skills that cannot be measured in a two-hour test on a single day. Such assessment systems shape the curriculum in ways that ensure stronger learning opportunities. They give teachers timely, formative information they need to help students improve—something that standardized examinations with long lapses between administration and results cannot do. And they help teachers become more knowledgeable about the standards and how to teach to them, as well as about their own students and how they learn. The process of using these assessments improves their teaching and their students' learning. The processes of collective scoring and moderation that many nations or states use to ensure reliability in scoring also prove educative for teachers, who learn to calibrate their sense of the standards to common benchmarks.

During the 1990s, many US states developed systems that featured state and locally administered performance assessments. These states included Connecticut, Kentucky, Maine, Maryland, Nebraska, New Hampshire, New Jersey, New York, Oregon, Vermont, Rhode Island, Washington, Wisconsin, and Wyoming, among others. In addition, some districts and consortia of schools have constructed well-developed performance assessment systems that engage students in developing high-quality products designed to measure central understandings and performances in disciplinary areas. Often these products—scientific investigations, social science research papers, literary analyses, artistic exhibitions, mathematical

models, technology applications—are presented to a jury of assessors who press for understanding in the questions they pose and the judgments they make about whether the work meets specific standards.

Research suggests that these assignments improved the quality of instruction in states ranging from California to Kentucky, Maine, Maryland, Vermont, and Washington (for a review, see Darling-Hammond & Rustique-Forrester, 2005). Other studies have found increases in achievement on both traditional standardized tests and performance measures for students in classrooms that offer a problem-oriented curriculum that regularly features performance assessment (see Newmann, Marks, & Gamoran, 1996; Lee, Smith, & Croninger, 1995).

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, the federal education law launched in 2002.[1] Many states discontinued the assessments they had developed in the 1990s, which required writing, research, and extended problem solving, and replaced them with multiple-choice and short-answer tests. States abandoned performance assessments because of costs and the constraints on the types of tests that were approved. As a consequence, testing in most states is less focused on higher-order skills than it was in the 1990s, even though it now functions as the 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 production of ideas.

For example, a recent RAND Corporation study found that on tests in seventeen states, fewer than 2 percent of mathematics items and only 21 percent of English language arts items reached the higher levels that ask students to analyze, synthesize, compare, connect, critique, hypothesize, prove, or explain their ideas (Yuan & Le, 2012). In testing parlance, these are the skills measured at levels 3 and 4 in the Webb Depth of Knowledge framework that classifies cognitive demand (Webb, 2002). Levels 1 and 2 represent lower-level skills of recall, recognition, and use of routine procedures.

This study echoes the findings of other studies (see Polikoff, Porter, & Smithson, 2011) and is even more worrisome, since these states were selected because their standards and tests were viewed as more rigorous than those of other states. The

RAND study found that the level of cognitive demand was severely constrained by the dominance of multiple-choice questions, which they found were rarely able to measure higher-order skills. Thus, the ambitious expectations found in state standards documents are frequently left unmeasured.

What and how tests measure matters, because when they are used for decision making, they determine much of what happens in classrooms. In the United States, students are tested far more frequently than in any other industrialized country, and test scores are used for more decisions about students, teachers, and schools. No Child Left Behind created a requirement for "every child, every year" testing in grades 3 through 8, plus once in high school. It also constrained the types of tests that could be used. By contrast, most countries test students at most once or twice before high school, and some, like Finland, do not have any external tests before the twelfth grade other than tests that sample a small subset of students at a couple of grade levels.

Finally, also in contrast with other countries, US tests are often used to determine whether students are promoted or graduated; whether teachers are tenured, continued, or fired; and whether schools are rewarded or sanctioned, even reconstituted or closed. With scores used to determine so many decisions, the incentives for teachers to teach to the test have become increasingly intense (Amrein & Berliner, 2002). In most other countries, tests are used to inform curriculum improvement and professional development and student pathways after middle or high school, not to serve as arbiters of graduation, personnel decisions, or school sanctions and survival. Tests are taken seriously, but there is much more room for school-based assessment, scored by teachers, that counts in the system and enables richer performance tasks.

High-performing jurisdictions have been moving assertively to increase their teaching and assessment of inquiry and problem solving. Their educational investment strategies, which have yielded higher and more equitable levels of performance and rapidly increasing levels of educational attainment, are intended to support career and college readiness, and they appear to do so. Where instruction focuses on assessment content, it is of paramount importance that tests actually test students on the deeper learning skills that they require now. As a recent report from the National Research Council noted, "The extent to which [deeper learning] goals are realized in educational settings will be strongly influenced by their inclusion in district, state, and national assessments,

because of the strong influence of assessment on instruction in the United States" (Pellegrino & Hilton, 2012).

THE RETURN OF PERFORMANCE ASSESSMENT

It is clear that such assessments will return to the educational landscape shortly as part of the tests created by two multistate assessment consortia designed to evaluate the Common Core Standard Standards (CCSS)—the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC)—and as part of other state and local initiatives. PARCC and SBAC assessments, to be launched in 2014–2015, will increase the use of constructed-response items and performance tasks.

Whereas items on current state tests represent mainly recall and recognition, the new Common Core assessments under development will have many more tasks that require students to analyze, critique, evaluate, and apply knowledge. The plans for the new consortia assessments will increase cognitive expectations by many orders of magnitude. An analysis of the content specifications for the SBAC 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).

It seems clear from the following sample tasks that have been released by the two consortia that the new tests will include performance tasks that encourage instruction aimed at helping students acquire and use knowledge in more complex ways:

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 a 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 and Linn (2013). 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.

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.

(*Continued*)

Reading/Prewriting: 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.

Analytical Essay: Students:

- Read a third text, *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 them 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 prewriting 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 and 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 assessments, conducted in a one- or two-day session, do 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. These skills, evaluated in a growing number of other countries, require students to design and complete 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 produces reliable scoring (see chapter 4). The scores are then included as part of examination results. Some states and districts, such as those belonging to the Innovation Lab Network, coordinated by 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, research reports) presented in a variety of forms, including oral, graphic, and multimedia presentations.

THE CHALLENGES FOR NEW ASSESSMENTS

The challenge ahead will be for states and districts to prepare to implement new assessments given the many changes they will entail. On the one hand, there is substantial consensus that US assessments must evolve to meet the new expectations for student learning. On the other hand, there are countervailing pressures regarding funding, time, and traditions that could stand in the way of assessment changes. In this time of extensive change, it is critical that we learn from our prior experiences and the work of other nations so that we do not repeat the mistakes of the past or ignore new research and development that could solve important problems.

THE PURPOSE OF THIS BOOK

Although there is substantial research and much experience with performance assessments in the United States and abroad, very little of this information is available to policymakers and researchers in a readily accessible form. This book, aimed at researchers, assessment developers, and policymakers, summarizes much of this research, examining under what conditions approaches to performance assessment can be feasible and worthwhile and what policy supports are important.

For our purposes in this book, the term *performance assessment* includes authentic assessments that require students to develop a product, response, analysis, or problem solution that reflects the kind of reasoning or performance required beyond the classroom setting. We consider assessments that are used for both formative and summative purposes and are both curriculum embedded (developed in the context of classroom work) and managed as a more centralized, externally determined testing event.

A major goal of this book is situating the current discussion of performance assessment within a historical context, in both the United States and abroad, and harvesting lessons learned from work in many other countries. At the same time, the chapters offer new perspectives based on recent research that offers an up-to-date analysis of the possibilities for adopting and implementing performance assessments at the state and local levels.

Originally developed as a collection of monographs, the chapters treat historical and contemporary research from the United States and abroad (chapters 2 to 4 in part 1); advances that influence the technical quality, accessibility, and instructional usefulness of performance assessments (chapters 5 to 7 in part 2); and issues of system development, including costs, benefits, and system design (chapters 8 to 10 in part 3), concluding with recommendations that pull these elements together (chapter 11).

In chapter 2, Brian Stecher introduces performance assessment within the larger context of large-scale testing and reviews recent testing history in the United States. He examines claims supporting performance assessment as well as research on the quality, impact, and burden of these assessments in K–12 education. He also situates performance assessment within the context of contemporary standards-based educational accountability, offering recommendations to support its effective use.

In chapter 3, Raymond Pecheone and Stuart Kahl, both experienced assessment developers, outline lessons learned from the performance assessment systems implemented within the United States in the past two decades. Pecheone and Kahl, with assistance from Jillian Chingos and Ann Jaquith, discuss performance systems in Connecticut, Kentucky, New Jersey, New York, Ohio, Rhode Island, Vermont, and Washington and identify how their experiences provide important knowledge that can inform a state or multistate performance assessment system.

Chapter 4 examines performance assessments around the world. Linda Darling-Hammond, with the assistance of Laura Wentworth, examines approaches

to performance assessment currently employed in countries that have been intensifying their efforts to enable students to develop twenty-first-century skills: Australia, Finland, Hong Kong, Singapore, Sweden, and the United Kingdom. These countries integrate performance assessments into curriculum to create stronger learning for both students and teachers, resulting in higher and more equitable achievement.

Chapter 5 summarizes contemporary research on the design and scoring of performance assessments, as well as psychometric advances that can better capture student performance and support their use in large-scale assessment programs. Suzanne Lane describes the important learning outcomes measured by performance assessments from the perspective of cognitive theory and then outlines critical considerations for validity and fairness.

Additional technical considerations have to do with the accessibility of text-intensive assessments that call for students to read and write much more than traditional test items. In chapter 6, Jamal Abedi addresses the research on the use of such assessments with the growing population of English language learners (ELLs). Abedi finds that multiple-choice achievement tests often employ "distractor" responses that confuse ELL students. Performance assessments, however, when designed carefully to avoid unnecessary linguistic complexity, can allow ELL students to present a more comprehensive picture of what they know. The chapter closes with recommendations for accessibility when using performance assessments for ELL students.

In chapter 7, Linda Darling-Hammond and Beverly Falk examine the ways in which teacher participation in developing, using, and scoring performance assessments—as well as reflecting on the student work they produce—can help teachers better understand standards, curriculum, instruction, and their students. The outcomes include improved learning for teachers and, ultimately, those they teach.

Finally, in part 3, we address systems issues associated with performance assessments, beginning with costs and benefits, a high-priority policy issue. One of the concerns frequently raised about performance assessments is the cost of scoring more extended, open-ended items in relation to the costs of machine-scored multiple-choice tests. However, many states and nations have maintained performance assessment systems that are manageable and affordable. Lawrence Picus, Frank Adamson, William Montague, and Margaret Owens present a new conceptual framework for analyzing the costs of performance assessment in

chapter 8. The authors discuss an array of cost considerations in play for assessment and provide expenditure data from previous performance assessment systems to provide a frame of reference. Chapter 8 then presents a new cost-benefit framework that delineates and compares the costs, the opportunity costs, and the benefits of multiple-choice testing and performance assessment.

This analysis is continued in chapter 9, in which Barry Topol, John Olson, Ed Roeber, Linda Darling-Hammond, and Frank Adamson provide financial estimates for implementing performance assessments using data that account for the myriad decisions that states face when implementing assessments. They show that by incorporating a range of cost savings measures, a consortium of states can offer tests that include performance assessments at a lower cost than most are now already spending for interim and summative tests of lower quality.

Chapter 10, by David Conley and Linda Darling-Hammond, illustrates how states can develop systems of assessment that use a mix of carefully considered measures for different purposes—distinguishing among the needs for state accountability, teaching and learning guidance, and information for colleges and employers.

Finally, Linda Darling-Hammond, Frank Adamson, and Thomas Toch conclude the book by synthesizing the research from the preceding chapters and offering policy recommendations for creating next-generation assessments that can be sustained and improved over time.