

# MEANINGFUL MEASUREMENT

The Role of Assessments in Improving High School Education in the Twenty-First Century

June 2009



ALLIANCE FOR  
EXCELLENT EDUCATION

© 2009 Alliance for Excellent Education. All rights reserved.

**Suggested citation:**

L. M. Pinkus, ed., *Meaningful Measurement: The Role of Assessments in Improving High School Education in the Twenty-First Century* (Washington, DC: Alliance for Excellent Education, 2009).

**Ordering information:**

Copies of *Meaningful Measurement: The Role of Assessments in Improving High School Education in the Twenty-First Century* can be downloaded from the Alliance's website at [www.all4ed.org](http://www.all4ed.org). To request print copies of the report, please visit [http://www.all4ed.org/publication\\_material/order\\_form](http://www.all4ed.org/publication_material/order_form). The first copy of the report is complimentary. Additional copies are available at a charge of \$1 per copy to cover shipping and handling costs.

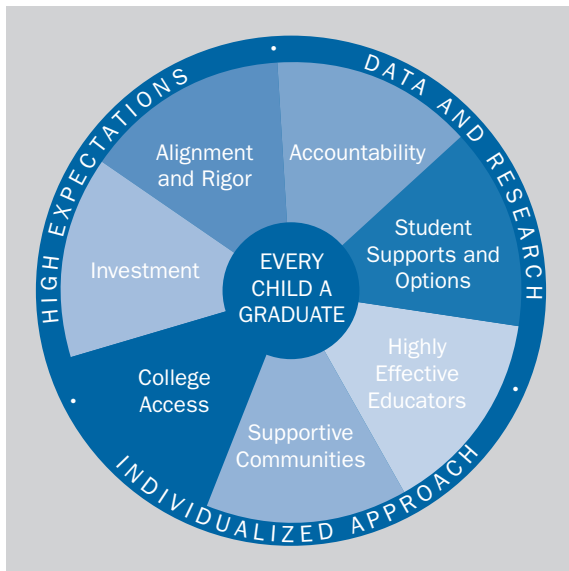
## About the Alliance for Excellent Education

The mission of the Alliance for Excellent Education is to promote high school transformation to make it possible for every child to graduate prepared for postsecondary learning and success in life.

The Alliance for Excellent Education is a national policy and advocacy organization, based in Washington, DC, working to improve national and federal policy so that all students can achieve at high academic levels and graduate high school ready for college, careers, and citizenship in the twenty-first century.

The Alliance has developed a “Framework for Action to Improve Secondary Schools” that informs a set of federal policy recommendations based on the growing consensus of researchers, practitioners, and advocates about the challenges and solutions for improving secondary student learning.

The framework, shown graphically here, encompasses seven policy areas that represent key leverage points in ensuring a comprehensive, systematic approach to improving secondary education. The framework also captures



three guiding principles that apply to all of the policy areas. Although the appropriate federal role varies from one issue area to another, they are all critically important to reducing dropouts and increasing college and career readiness.

## About the Editor

Lyndsay M. Pinkus is director of strategic initiatives at the Alliance for Excellent Education. Since joining the Alliance in January 2002, she has served in a variety of research, coordination, and advocacy roles, where her work has included managing policy and grant work on a range of issues including graduation rates, data, secondary school accountability, and secondary school improvement, and authoring a number of publications for the Alliance. Prior to rejoining the staff in January 2006, Ms. Pinkus served as a legislative associate at Washington Partners, LLC, providing government relations and policy research and analysis for a variety of clients, including the Alliance. She is a graduate of the School of Public Affairs at American University as a presidential scholar; the Public Affairs and Advocacy Institute at the Center for Congressional and Presidential Studies; and the Institute for Educational Leadership's Education Policy Fellowship program.

## Acknowledgments

The Alliance for Excellent Education is greatly appreciative of the authors for sharing their time and expertise in writing the following chapters, as well as of the multiple Alliance staff members and advisors whose dedication contributed significantly to this volume.

**The Alliance for Excellent Education is also grateful to Carnegie Corporation of New York for the financial support that made this publication possible.**

The views expressed in this volume are those of the authors and do not necessarily represent those of the Alliance for Excellent Education or the funder.

# CHAPTER

## Reframing Accountability: Using Performance Assessments to Focus Learning on Higher-Order Skills

Linda Darling-Hammond and Raymond Pecheone  
School Redesign Network, Stanford University

Over the past decade, educators, policymakers, and the public have begun to forge a consensus that our public schools must focus on better preparing all children for the demands of citizenship in the twenty-first century. This has resulted in states developing “standards-based” educational systems and assessing the success of districts and schools in meeting these standards through more systematic testing. Most of these tests are multiple-choice, standardized measures of achievement. While these assessments offer the benefits of ease of administration and inexpensive scoring, practitioners and researchers have found that they also have a number of less desirable side effects. These include narrowing of the academic curriculum and experiences of students (especially those in low-income communities); a focus on recognizing right answers to lower-level questions rather than on developing higher-order thinking, reasoning, and performance skills; and growing dissatisfaction among parents and educators with the school experience.

The sharp differences between the forms of testing used in the United States and the performance-based assessments used in other higher-achieving countries also suggest that low international rankings may be related, in part, to overreliance on these narrow conceptions of standardized testing in the United States.

In large part for cost reasons, reliance on multiple-choice tests rather than on more open-ended assessments of performance has increased in response to the annual testing requirements of the No Child Left Behind Act (NCLB), despite the fact that language in NCLB calls for “multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding.”<sup>1</sup> Changing what counts as assessment evidence, along with related changes in NCLB’s accountability structure, could contribute substantially toward school improvement.

This chapter discusses how performance assessments can help evaluate what students can actually do with what they know and encourage the teaching and learning of higher-order skills. It describes what performance assessments are and how they can benefit instruction, how they are used in policy settings in the United States and abroad, what the major challenges and considerations are that must be addressed to use performance assessments well, and how federal policy can support the development and implementation of high-quality assessments that both support and evaluate more complex knowledge and skills.

## **What Is Performance Assessment?<sup>2</sup>**

Almost every adult in the United States has experienced at least one performance assessment: the driving test that places new drivers in 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 Americans 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 the performance assessment as a real-world test of whether a person can actually handle a car on the road. Not only does the performance assessment tell us some important things about potential drivers’ skills, it

also helps improve those skills, as potential drivers practice to get better. (What parent doesn't remember the hair-raising outings with sixteen-year-olds wanting to practice taking the car out over and over again?) The test sets a standard toward which everyone must work. Without it, society would 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 are used in bar examinations for lawyers, where they must write briefs and analyze cases; in the medical boards for doctors, where they must diagnose patient cases and, in fields like psychiatry, interview patients under the watchful eye of evaluators; and in registration exams for architects, where candidates must submit a portfolio of their designs.

Performance assessments in education are similar. They are opportunities for students to show how they can apply their knowledge and skills in real-world tasks that represent the key aspects of their learning. Performance assessments may include science experiments that students design, carry out, analyze, and write up; computer programs that students create and test; or research inquiries that they pursue, seeking and assembling evidence about a question, which they may present in written and oral form.

Whether the skill or standard being measured is writing, speaking, scientific literacy, mathematical reasoning, or social science research, with a performance assessment students perform tasks involving these skills and teachers score the performance based on a set of predetermined criteria. As in our driving test example, these assessments typically consist of four parts: performance standards, a task, a scoring guide or rubric, and a set of administration guidelines. The development, administration, and scoring of these tasks requires teacher training and development to ensure quality and consistency.

Illinois's assessments provide a good example of the contrast between classroom performance assessment and a state multiple-choice test. The state's eighth-grade science learning standard 11B reads, "Technological design: Assess given test results on a prototype; analyze data and rebuild and

retest prototype as necessary.” The multiple-choice example on the state test simply asks what “Josh” should do if his first prototype sinks, with the wanted answer, “Change the design and retest his boat.” This, however, gives the assessor no idea whether Josh would have any idea *how* to change the design productively and to systematically test the design, holding some features constant while changing others.

The classroom assessment allows evaluation of these critical questions. The prompt states, “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 can explore significant physics questions such as displacement in order to understand how a ball of clay can be made to float. Such activities combine hands-on inquiry with reasoning skills, have visible real-world applications, are more engaging, and enable deeper learning. They also allow the teacher to assess student learning along multiple dimensions, including the ability to frame a problem, develop hypotheses, reflect on outcomes and make reasoned and effective changes, demonstrate scientific understanding, use scientific terminology and facts, persist in problem solving, and organize information, as well as develop sound concepts regarding the scientific principles in use.

The assessment systems of most of the highest-achieving nations in the world emphasize local in-school performance assessment throughout the elementary and middle school years. At the high school level, jurisdictions like the UK, Hong Kong, Singapore, Finland, Sweden, and Victoria, Australia, among others, use a combination of centralized assessments that use primarily open-ended and essay questions and local assessments given by teachers which are factored into the final examination scores.

The centralized assessments are often developed jointly by high school and college faculty and scored using common criteria by teachers. The classroom-based assessments—which include research papers, applied science experiments, presentations of various kinds, and projects and products that students construct—are mapped to the syllabus and the standards for the subject, and are selected because they represent critical skills, topics, and concepts. They are often suggested and outlined in the curriculum, and may be designed centrally or locally. They are administered and scored by teachers.

While not all performance assessments are locally developed—Hong Kong offers a bank of tasks teachers can draw upon, while teachers in Finland create their own—all of these systems include some rich assessment tasks at the classroom level that can be used as formative or benchmark assessments, helping teachers to gauge ongoing progress. Local scoring guided by standardized protocols allows immediate feedback to teachers and students. This enables results to be used to improve instruction and student learning immediately, something that standardized examinations with long lapses between administration and results cannot do. In addition, as teachers use and evaluate these tasks, they become more knowledgeable about the standards and how to teach to them, and about what their students' learning needs are. This process improves their teaching. Scoring is often subject to moderation, auditing, or calibration processes, as described later.

Performance assessments often provide several ways to view student learning. For example, multiple samples of actual writing taken over time can best reveal to a teacher the progress a student is making in the development of composition skills. This provides ongoing feedback to learners as well, as they see how they are developing as writers and what they have yet to master. In addition, different kinds of writing tasks—persuasive essays, research papers, journalistic reports, responses to literature—encourage students to develop the full range of their writing and thinking skills in ways that answering multiple-choice questions about writing or even writing a five-paragraph essay over and over again do not.

Locally managed performance assessments that provide multiple sources of evidence about what people can actually *do* with what they know are often characterized as “tests worth teaching to,” because they help focus effort on developing important skills. Let’s think back to the state driver’s license exam. This involves both a written test and a performance assessment on the road. Everyone knows precisely what to expect in terms of the skills to be demonstrated—for example, whether or not the applicant can manage a car safely and (at least on the East Coast) parallel-park skillfully—as the examination is not a total secret. Most performance assessments challenge students to address issues and problems in real life.<sup>3</sup> Moreover, a number of studies associate performance assessment with a positive influence over teaching and learning.<sup>4</sup>

The fact that the assessment is open and transparent is not a problem, because the point is to see whether drivers have developed these real-world abilities; this is not undermined by the drivers knowing what they need to learn to do. The performance is scored by the instructor, working from a rubric, and if the driver is sufficiently successful in all aspects of the examination (as determined by a state cutoff score), a license is conferred. The task is so well defined that instructional programs (driver's education) that include both hands-on and classroom instruction clearly demonstrate their effectiveness in preparing students to perform. (This is reflected in the reduced insurance rates granted to graduates of driver's education programs.) Imagine what life on the roads would be like if prospective drivers did not have to demonstrate what they know before taking the wheel on their own. And imagine what life in classrooms would be like if the nation *did* require students to demonstrate that they can express and defend their ideas, develop and analyze data, and apply their knowledge in problem-solving situations.

### Benefits of Performance Assessment

Research and experience have uncovered a number of benefits, challenges, and criteria for making such assessment systems successful. Among the benefits of well-designed performance assessment systems are that they can

- elevate the focus of instruction to include higher-order thinking skills;
- provide a more comprehensive assessment of what students know and can do;
- provide clearer information to parents, teachers, and the public as to student development, accomplishments, and needs;
- allow instruction to be altered in a timely fashion to meet student learning needs;
- lead to more student engagement in both the learning and assessment processes;
- invite more teacher buy-in and encourage collaborative work; and
- support standards-based instruction and improvement of teaching practices.

Considerable research suggests that performance assessments are essential tools for showing the extent to which students have developed higher-order thinking skills, such as the abilities to analyze, synthesize, and evaluate information. Studies have found that the use of such assessments has improved teaching quality and increased student achievement, especially in areas requiring complex reasoning and problem solving.<sup>5</sup> Evaluations of reading and writing portfolios in Vermont and Kentucky, for example, found that the assessments—along with the professional development opportunities associated with them—influenced instruction in positive ways, especially in encouraging much more complex mathematical tasks and more extensive and higher-quality student writing.<sup>6</sup> These assessment systems also stimulated school improvement through curriculum reforms and supports for teacher learning.

Researchers have noted that assessment systems in which teachers look at student work with other teachers and discuss standards in very explicit ways appear to help schools develop shared definitions of quality. Evaluating work collaboratively rather than grading students in isolation helps teachers make their standards explicit, gain multiple perspectives on learning, and think about how they can teach to produce the kinds of student work they want to see. Where teachers do this, studies find that changes in teaching and schooling practices almost invariably occur—especially for students who are not as consistently successful at schoolwork.<sup>7</sup>

Performance assessments are more sensitive to instruction and of more immediate use to teachers than most current standardized tests, while providing richer evidence of student learning that can be used to solve learning problems as they occur. When teachers see their students' written responses and reasoning, they can diagnose *how* students are learning and *why* they may be struggling, rather than just what they know. Typically, standardized test information is not available to schools for six to nine months after the testing date, often in the subsequent school year, and far too late and far too thin on information to provide usable data to teachers about their students' learning needs.

Perhaps the most important benefit to using performance assessments is that they assist in learning and teaching. They are *formative*, in that they

provide teachers and students with the feedback they need from authentic tasks that reveal students' mastery of content, and can guide future teaching. They can also be *summative*, in that they can serve as a final assessment of student capabilities with respect to state and local standards. As summative measures, performance assessments are useful because they organize teaching around the kinds of tasks that support the transfer of learning to new contexts, helping students learn more of what they will need to do in the world outside of school. In addition to acquiring and demonstrating in-depth knowledge of content, this may include the ability to plan an inquiry and organize their time, develop self-discipline and perseverance as well as intellectual discipline, define problems and determine strategies for how to pursue answers, organize and display data, evaluate findings, draw conclusions, and express and defend their ideas according to standards of evidence.

### Where and How Performance Assessments Are Used

As noted above, most high-achieving nations and many states in the United States—including Connecticut, Kentucky, Maine, Nebraska, New Hampshire, New Jersey, New York, Rhode Island, Vermont, and Wyoming—have developed and used state and local performance assessments as part of their testing systems. Indeed, the National Science Foundation provided millions of dollars for states to develop hands-on science and math assessments as part of its Systemic Initiative in the 1990s, and prototypes exist all over the country. Additionally, twenty-seven states use multiple approaches for high school graduation decisions, including many that combine state requirements with local performance assessments and other measures (e.g., grades, student work samples, portfolios of work, and senior projects).<sup>8</sup> In this section we briefly describe performance assessment models in both the United States and abroad.

One common model in several U.S. states and in a number of other countries is to combine an external reference exam, which includes open-ended questions that measure aspects of performance such as analysis and expression, with classroom-managed assessments that ask students to tackle more complex, extended tasks that cannot be completed in a couple of hours on a sit-down test. Some states (such as Nebraska, Rhode Island, and Wyoming) and countries (such as Finland, Scotland, and Wales, and

Queensland and ACT, Australia) rely much more heavily on school-based performance assessments. Both approaches are described below.

### *U.S. examples of performance assessment systems*

**Connecticut** developed a performance task approach during the 1990s as part of its state assessment and accountability system. Connecticut test items include a range of test formats: multiple choice, constructed responses, short essays, mini experiments, and performance tasks to measure how students can apply what they know.<sup>9</sup> Teachers are involved in all areas of test development, including task development, scoring, and standard setting. At the high school level, the Connecticut Academic Performance Test (CAPT), administered in the tenth grade, reports on student performance in four areas: mathematics, reading across the disciplines (focusing on response to literature and reading for information), writing across the disciplines, and science. The CAPT uses classroom-embedded tasks as part of its statewide assessment system. For example, students design and conduct science experiments that are embedded in the science curriculum around a unit of study on specific topics. Students are asked to formulate hypotheses, conduct the experiment, analyze the data, and report their results to prove their ability to engage in scientific reasoning. They also critique experiments and evaluate the soundness of findings and are tested on their findings as part of the CAPT on-demand science assessment.

While the CAPT is required of all public high schools students in Connecticut, the state legislature specifies that the test cannot be used as the sole basis for graduation or promotion. As part of its official policy (2000), the state board of education stated that “the CAPT results alone do not provide a comprehensive picture of student accomplishment. There is a danger that overemphasizing state test scores to evaluate a student’s school or district performance can result in an inappropriate narrowing of the curriculum and inappropriate classroom instructional practice.”<sup>10</sup> As a consequence, districts are required to use the CAPT assessment in combination with local assessments, which must include performance assessments.

**Maine, Vermont, New Hampshire, and Rhode Island** also have developed performance assessment components as part of their accountability systems,

but with more participation on the part of the state in helping local districts implement their assessments. These New England states combine a jointly constructed reference exam—the New England Common Assessment Program (NECAP)—with locally developed assessments that provide evidence of student work from performance tasks and portfolios.

**Vermont** was an early leader, developing in the late 1980s and early '90s both on-demand performance tasks and portfolios that are used throughout the school year, so teachers and students can learn from the results of the assessments and continually improve their work. The writing and mathematics portfolios, developed by the state department of education with the engagement of teachers, include both common tasks to be completed by all students and locally selected work samples that reflect particular kinds of work to be represented in the portfolios.

As the system was phased in, teachers learned how to develop and evaluate assessments and how to teach toward the standards through support networks that sponsored professional development sessions and summer institutes across the state. Teachers from different schools convened to score assessment tasks together, moderating their scoring to gain consistency. While evaluations found that the early, nonstandardized portfolios were not scored very reliably, revisions brought common structures to the portfolios and performance assessments, which resulted in much higher levels of reliability, comparable to those achieved on AP exams.<sup>11</sup>

The state's involvement of large numbers of teachers in designing and scoring the assessments created substantial focus on the quality of student work, providing a powerful form of professional development. Harvard professor Richard Murnane described the conversations of Vermont teachers who gathered in the summer to evaluate portfolios: "Often heated, the discussions focused on what constitutes good communication and problem-solving skills, how first rate work differs from less adequate work, and what types of problems elicit the best student work."<sup>12</sup>

For more than a decade, the Vermont portfolios were the primary assessments for support and accountability in the state. They are now a voluntary adjunct to the annual standardized tests at each grade level

required by NCLB, and many districts and schools continue to use them to obtain a comprehensive assessment of student learning.

**Maine's** assessment system was designed to include the use of the NECAP reference exam and the Maine Education Assessment, both of which include many open-ended items and a writing assessment, plus locally developed performance assessments. The local assessments are organized around Maine's Learning Results in eight areas (English language arts, mathematics, science, social studies, health/physical education, career preparation, modern and classical language, and visual and performing arts). With extensive professional development provided by the state, local districts developed common performance tasks, classroom-based portfolios, observations, and exhibitions of student work. With the advent of NCLB, which introduced new standardized tests at each grade level, the performance components are now used voluntarily by districts to support instructional decisions.

As part of a state high school redesign initiative, **Rhode Island** has also developed a performance-based graduation system. Starting in 2008, all Rhode Island graduates had to show evidence of success across three elements of the performance-based graduation requirement: a standardized reference exam, course performance, and state-approved performance assessments such as portfolios, senior projects, and/or end-of-course exams. The performance outcomes for each of these data elements must be authentic and aligned to state standards, and must demonstrate meaningful content knowledge. Commissioner Peter McWalters emphasized that there are three non-negotiables in this work: "We have to educate every child; we have to hold high standards; and we have to provide differentiated learning and instruction." In its first year of implementation, the new system was reported to engender greater student engagement and participation in school, with graduation rates increasing from 70 percent to 74 percent, rather than declining, as is frequently the case when new state graduation assessments are introduced.<sup>13</sup>

**New Hampshire** is moving to a competency-based system for graduation that will no longer use Carnegie units. The state will base graduation on a competency-based credit system using a "mastery of learning" approach to

assess student learning, which relies on performance assessments to evaluate mastery of content and skills and allows students to earn credits both in school and during out-of-school time. The state has already introduced a technology portfolio, which all students must complete to demonstrate their technology competence in high school.

**Ohio** is also developing a set of standards-based performance tasks measuring core knowledge and skills in the content areas of math, English, science, and history to become part of the state's high school assessment system. These tasks represent the skills of disciplinary inquiry necessary for college readiness and success in the workplace and will support instructional decisions as well as accountability reporting.

**Nebraska** utilizes a system of performance assessments created and scored by local educators trained to score reliably. These systems are peer reviewed by measurement and assessment experts and include a check on validity through the use of a statewide writing examination and the administration of one norm-referenced test. **Wyoming** uses a "body of evidence" approach that is locally developed in order to determine whether students have mastered standards required for graduation. **Oregon** uses both online diagnostic assessments and performance assessments in multiple subject areas that are state designed and locally scored using state rubrics as the basis for a Certificate of Mastery.

Some well-developed performance assessment systems were created and are used by consortia of local schools and/or districts. In New York, for example, the **New York Performance Assessment Consortium** is a network of forty-seven schools in the state that rely upon a set of performance tasks assembled in a portfolio to determine graduation. These include a major task in each disciplinary area: a scientific investigation, a historical research paper, a literary response, an applied mathematical problem or model, an arts exhibition, and an analysis of an internship experience. These are defended before a panel that includes outside experts as well as teachers and parents and scored according to common rubrics. Because of the quality of their work, the consortium schools have a state waiver from some of the Regents Examinations. Research shows that New York City students who graduate from these schools (which have a much higher graduation rate

than the City even though they serve more low-income students, students of color, and recent immigrants) are more successful in college than students with a traditional Regents diploma, which relies upon standardized tests.

Among other notable performance-based systems under development nationally is the **College Readiness Performance Assessment System (C-PAS)**, developed by David Conley at the University of Oregon. The C-PAS is designed to track the development of five generic cognitive strategies that represent the thinking skills necessary for college readiness and success: problem solving, research, interpretation, reasoning, and precision. The C-PAS assessment is a series of performance tasks that teachers administer and score with a common scoring guide.

The **Collegiate Learning Assessment (CLA)**, developed by Richard Shavelson at Stanford University, Stephen Klein at RAND, and colleagues, is a collegiate assessment that is being adapted for secondary schools. The CLA uses real-world performance tasks that elicit critical thinking, analytic reasoning, problem solving, and communication skills. Students are typically faced with a problem that requires them to collect and evaluate evidence, then frame and defend a solution. They may use a variety of documents and resources provided in an “in-basket” to learn about aspects of the problem that is posed. The CLA uses a matrix sampling approach to assess student performance at the beginning and the end of college (not all students perform all tasks), and develops institutional reports focusing on students’ college-level competencies.

### *Performance assessments abroad*

School-based performance assessment is the dominant mode of assessment in most high-achieving countries.<sup>14</sup> (See Table 1.) At the high school level, a number of countries use a blended approach that combines school-based tasks that measure specific subject-matter concepts and skills with a common examination, often developed by teachers in collaboration with university faculty, featuring primarily open-ended questions requiring written or oral responses.

**Table 1: Summary of International Assessment Systems Using Performance Assessment**

Country/ organization	What assessments are used?	Who grades the assessments?	Who designs the assessments?	How are the results used?
Victoria, Australia	<b>School-based</b> • Projects, labs, papers, essays, presentations	<b>School-based</b> • 50%+ of grade • Graded by the teacher	<b>School-based</b> • Teacher designed based on state syllabi curriculum	Use scores to guide admission to a university and workplace apprenticeship programs
	<b>National</b> • Multiple choice, short answer, essays, oral exams	<b>National</b> • Administered by the teacher • Graded by the teacher	<b>National</b> • Designed by teachers, professors through VCAA	
Sweden	<b>School-based</b> • Coursework, projects, essays, test	<b>School-based</b> • To 30% of grade • Graded by the teacher	<b>School-based</b> • Teacher designed based on national curriculum	Uses scores to compare coursework grades and local assessment results to national standards
	<b>National</b> • National syllabi, open-ended questions, material given in advance	<b>National</b> • Included in grading, but not the sole factor in grading	<b>National</b> • Educational research institution designed, teacher input	
Finland	<b>National</b> • Short problems that ask students to apply their thinking	<b>National</b> • Graded by teachers and re-checked by the board of education	<b>National</b> • Originally developed by University of Helsinki	Uses scores to inform instruction and student self-reflection, and in some cases used for placement in a university
	<b>School-based</b> • Presentations, plays, demos	<b>School-based</b> • Graded by the teacher	<b>School-based</b> • Teachers design with national themes	

United Kingdom: England	<b>School-based</b>	<ul style="list-style-type: none"> <li>Coursework, tests, projects, essays</li> </ul>	<b>School-based</b>	<ul style="list-style-type: none"> <li>To 30% of grade</li> <li>Graded by the teacher</li> </ul>	<b>School-based</b>	<ul style="list-style-type: none"> <li>Teacher designed based on national curriculum</li> </ul>	Use scores to select upper-secondary coursework and gain admission to a university
	<b>National</b>	<ul style="list-style-type: none"> <li>Essays</li> </ul>	<b>National</b>	<ul style="list-style-type: none"> <li>To 80% of grade</li> <li>By exam group</li> </ul>	<b>National</b>	<ul style="list-style-type: none"> <li>Designed by examining group</li> </ul>	
United Kingdom: Wales	<b>School-based</b>	<ul style="list-style-type: none"> <li>Student investigations, presentations</li> </ul>	<b>School-based</b>	<ul style="list-style-type: none"> <li>Graded by the teacher</li> </ul>	<b>School-based</b>	<ul style="list-style-type: none"> <li>Teacher designed based on national curriculum</li> </ul>	Reported to parents and government, meant to encourage better teaching, more student engagement
	<b>National</b>	<ul style="list-style-type: none"> <li>Essays (only upper secondary)</li> </ul>	<b>National</b>	<ul style="list-style-type: none"> <li>By exam group</li> </ul>	<b>National</b>	<ul style="list-style-type: none"> <li>Designed by examining group</li> </ul>	
International Baccalaureate	<b>School-based</b>	<ul style="list-style-type: none"> <li>Speeches, projects, portfolio, presents, investigates, labs</li> </ul>	<b>School-based</b>	<ul style="list-style-type: none"> <li>20–50% of grade</li> <li>Graded by the teacher</li> </ul>	<b>School-based</b>	<ul style="list-style-type: none"> <li>Designed by the classroom teacher</li> </ul>	Used for awarding IB Diploma; giving college credit in some cases
	<b>National</b>	<ul style="list-style-type: none"> <li>Multiple choice, essay, short answer</li> </ul>	<b>National</b>	<ul style="list-style-type: none"> <li>Administered and graded by IB</li> </ul>	<b>External</b>	<ul style="list-style-type: none"> <li>Designed by IB</li> </ul>	
Hong Kong	<b>School-based</b>	<ul style="list-style-type: none"> <li>Closely aligned with national assessments</li> </ul>	<b>School-based</b>	<ul style="list-style-type: none"> <li>Graded by the teacher</li> </ul>	<b>School-based</b>	<ul style="list-style-type: none"> <li>Designed by the classroom teacher</li> </ul>	Uses scores to judge whether students may advance to the next level
	<b>National</b>	<ul style="list-style-type: none"> <li>Projects, portfolio, observations, exam</li> </ul>	<b>National</b>	<ul style="list-style-type: none"> <li>Graded by the teacher</li> </ul>	<b>National</b>	<ul style="list-style-type: none"> <li>Designed by teachers</li> </ul>	

The **International Baccalaureate (IB)** program, which enrolls 650,000 worldwide, including a growing number in the United States, exemplifies the syllabus-based approach to classroom assessment used in many countries in Europe and Asia. Designed for students in grades eleven and twelve, it assesses student learning using school-based performance assessments and external exams at the end of each course. Both types of assessments measure students' performance on the objectives specified in the "subject outlines" written by the IB organization. School-based performance assessments—such as oral exercises in language subjects, projects, student portfolios, practical laboratory work, mathematical investigations, and artistic performances—contribute 30 to 50 percent of the final examination grade. The external exam consists largely of essays, constructed responses, and data response questions, case study questions, and text response questions, with a limited use of multiple-choice items. A typical essay question students might choose among several options on the exam would be the following:

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.

This blended approach also characterizes the **General Certificate of Secondary Education (GCSE)** examinations in **Great Britain**, as well as the high school examinations in **Finland, Sweden, Hong Kong, Singapore, and Victoria, Australia**. (Most of these countries now use primarily local performance assessments in the elementary and middle school years.) The school-based performance assessments typically comprise from 30 to 50 percent of the total examination score in these assessment systems.

In **Sweden**, schools offer nationally approved examinations in the upper-secondary years in several subjects.<sup>15</sup> Teachers work with university faculty to help design the tasks and questions, and they weight information from these exams, their own assessments, and classroom work to assign a grade reflecting how well students have met the objectives of the syllabus.<sup>16</sup> Regional education officials and schools provide time for teachers to calibrate their grading practices to minimize variation across the schools and across the region.<sup>17</sup> Toward the end of their upper-secondary schooling,

Swedish students receive a final grade or “learning certificate” in each area that acts as a compilation of all of these sources of evidence, including projects completed by the student as well as grades awarded for courses.

In **Victoria, Australia**, the Victoria Curriculum and Assessment Authority (VCAA) establishes courses in a wide range of studies, develops the external examinations, and ensures the quality of the school-assessed component of the Victorian Certificate of Education (VCE). The VCAA conceptualizes assessment as “of,” “for,” and “as” learning. Teachers are involved in developing assessments, along with university faculty in the subject area, and all prior-year assessments are public, in an attempt to make the standards and means of measuring them as transparent as possible. Before the external examinations are given to students, teachers and academics take the exams themselves, as if they were students. The external subject-specific examinations, given in grades eleven and twelve, include written, oral, and performance elements scored by classroom teachers.

In addition, at least 50 percent of the total examination score is comprised of classroom-based tasks that are given throughout the school year. These required assignments and assessments—lab experiments and investigations on central topics as well as research papers and presentations—are designed by teachers in response to syllabus expectations. These required classroom tasks ensure that students are getting the kind of learning opportunities that prepare them for the assessments they will later take, that they are getting feedback they need to improve, and that they will be prepared to succeed not only on these very challenging tests but in college and in life, where they will have to apply knowledge in these ways.

As in Victoria, assessments in **Great Britain** use a combination of external and school-based tasks based on the national curriculum and course syllabi. Throughout the school years, classroom-based tasks scored by teachers are used to evaluate student achievement of curriculum goals. A mandatory set of assessments at year nine (age fourteen) includes both teacher-created and -administered assessments and, for students who have reached a certain level of achievement, national exams and tasks.<sup>18</sup>

While not mandatory, most students take a set of exams at grade eleven (age sixteen) to achieve their GCSE. Students choose which tests they will take based on their interests and areas of expertise. Most GCSE items are essay questions. The math exam includes questions that ask students to show the reasoning behind their answers, and foreign-language exams require oral presentations. About 25 to 30 percent of the final examination score is based on class work, coursework, and assessments developed and graded by teachers. In many subjects, students also complete a project worked on in class that is specified in the syllabus.

**Hong Kong** has typically used the British A- and O-Level exams for students in high school. In collaboration with educators from Australia, the United Kingdom, and other nations, Hong Kong's assessment system is evolving from a centralized examination structure to one that increasingly emphasizes school-based formative assessments that expect students to analyze issues and solve problems. The government has decided to gradually replace the Hong Kong Certificate of Education Examinations, which most students sit for at the end of their five-year secondary education, with a new Hong Kong Diploma of Secondary Education that will feature school-based assessments.

In addition, the Hong Kong Territory-wide System Assessment (TSA), which assesses lower-grade student performance in Chinese, English, and mathematics, is developing an online bank of assessment tasks to enable schools to assess their students and receive feedback on their performance on their own timeframes. The formal TSA assessments, which include both written and oral components, occur at primary grades three and six and secondary grade three (the equivalent of grade nine in the United States).

As outlined in Hong Kong's "Learning to Learn" reform plan, the goal of the reforms is to shape curriculum and instruction around critical thinking, problem solving, self-management skills, and collaboration. A particular concern is to develop metacognitive thinking skills, so students may identify their strengths and areas needing additional work.<sup>19</sup> By 2007, Curriculum and Assessment Guides were published for four core subjects and twenty elective subjects, and assessments in the first two subjects—Chinese language and English language—were revised. These became criterion-

referenced, performance-based assessments featuring not only the kinds of essays previously used on the exams, but also new speaking and listening components, the composition of written papers testing integrated skills, and a school-based component that factors into the examination score. Although the existing assessments already use open-ended responses, the proportion of such responses will increase in the revised assessments. Like the existing assessments, the new assessments are developed by teachers with the participation of higher education faculty, and they are scored by teachers who are trained as assessors.

In **Queensland, Australia**, there has been no assessment system external to schools for forty years. Until the early 1970s, a traditional “postcolonial” examination system controlled the curriculum. When it was eliminated—about the same time as in Finland and Sweden—all assessments became school based. School-based assessments are developed, administered, and scored by teachers in relation to the national curriculum guidelines and state syllabi (also developed by teachers), and are moderated by panels that include teachers from other schools as well as professors from the university system.

The syllabi spell out a small number of key concepts and/or skills to be learned in each course, and what kinds of projects or activities (including minimum assessment requirements) students should be engaged in. Each school designs its program to fit the needs and experiences of its own students, choosing specific texts and topics with this in mind. At the end of the year, teachers collect a portfolio of each student’s work, which includes the specific assessment tasks, and grade it on a five-point grading scale. To calibrate these grades, teachers put together a selection of portfolios from each grade level—one from each of the five score levels, plus borderline cases—and send these to a regional panel for moderation. A panel of five teachers re-scores the portfolios and confers about whether the grade is warranted, making a judgment on the spread. A state panel also looks at portfolios across schools. Based on these moderation processes, the school is given instructions to adjust grades so they are comparable to others.

## Summary

The use of curriculum-embedded assessments in these performance-based systems allows for the testing of more complex skills that cannot be measured in a two-hour test on a single day. They 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 content standards and how to teach to them, as well as about their own students and how they learn. The process of using these assessments can improve their teaching and their students' learning. The processes of collective scoring and moderation that many nations use to ensure reliability in scoring also prove educative for teachers, who learn to calibrate their understanding of the standards to common benchmarks. In these ways, as part of a balanced assessment approach, performance assessments can help ensure that students are ready for college and the workplace.

## Challenges and Considerations in Scaling Up Performance Assessments

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

**Calibration of scoring:** Perhaps the most complex question surrounding these assessments when they are locally developed or scored is how to ensure comparability. Many of the systems described above, both in the United States and abroad, use common scoring guides and extensive scorer training to achieve consistency in the use of these rubrics. In addition, they use auditing, moderation, and calibration systems of several kinds to maintain the quality of the system over time.

In Victoria, Australia, the quality of the tasks assigned by teachers, the work done by students, and the appropriateness of the grades and feedback given to students are audited through an inspection system, and schools are given feedback on all of these elements. In addition, the VCAA uses statistical moderation to ensure that the same assessment standards are applied to

students across schools. The external exams are used as the basis for this moderation, which adjusts the level and spread of each school's assessments of its students to match the level and spread of the same students' scores on the common external test score. The result is a rich curriculum for students with extensive teacher participation and a comparable means for examining student learning.

In Hong Kong, tests are allocated randomly to scorers, and essay responses are typically rated by two independent scorers.<sup>20</sup> Results of the new school-based assessments are statistically moderated to ensure comparability within the province. The assessments are internationally benchmarked, through the evaluation of sample student papers, to peg the results to those in other countries. Many of the new assessments are also to be scored online, which the Examinations Authority notes is now the common practice in twenty of China's mainland provinces, as well as in the United Kingdom.

Queensland's system, like those in a number of countries, also employs "moderation," a process of bringing samples from different schools to be re-scored, with results sent back to the originating schools. This process leads to stronger comparability across schools and is part of building a strong performance assessment system. Nebraska also supplements extensive scorer training on common rubrics with external validation checks such as comparisons with the statewide writing assessment, the ACT, and other commonly administered standardized tests. Each district's assessment system is evaluated and approved through a review process conducted by measurement experts.

**Costs and scoring models:** Appropriate, affordable, and educationally supportive scoring models must be developed. Although some methods of managing performance assessments can cost more than machine scoring of multiple-choice tests (i.e., when such assessments are treated as traditional external tests and shipped out to separately paid scorers), the cost calculus changes when assessment is understood as part of teachers' work and learning—built into teaching and professional development time. Much evidence suggests that developing and scoring these assessments is a high-yield investment in teacher learning and a good use of professional development resources.

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

**Professional development:** Extensive professional development is necessary for educators to learn to build, use, and score assessments that will inform and guide their teaching. Many systems have demonstrated that teachers can develop this knowledge rapidly when given the support. In successful systems, teachers are engaged in curriculum alignment, performance task development, scoring processes, and data analysis so that they understand the system and can teach productively to the standards. The processes include a peer review or moderation system that provides a feedback loop, checks on quality, and directions for staff development. Teachers often report that some of the best professional development of their careers occurs when they have opportunities to examine, score, and discuss student work. Importantly, international assessments have strategically "captured" teacher professional development time to evaluate and validate student work. Capitalizing on this time can both lower costs and establish a common language around curriculum standards and assessment.

**Administrative support:** Education agency officials and legislators at the state and federal levels must develop targeted assistance to teachers, administrators, and school systems that allows their effective participation in these systems and leverages improvements in teaching. In addition to professional development, this will include widespread information, extensive training in both use and scoring, the redesign of curriculum materials to ensure alignment with and support of new assessments, and the redesign of school schedules to provide in-class time for more in-depth work on the part of students and out-of-class time for teachers' planning, analysis, and scoring of student work, as is common in other countries.

**Quality of tasks:** Careful attention must be paid to the quality of performance tasks. They should be developed around important disciplinary content so that they measure core concepts and abilities with strong validity, and they should be developed in response to criteria that establish the technical quality of assessments (including checking for bias and fairness), high proficiency standards, consistent administration of assessment (including clear criteria that would certify the quality of an assessment task), and opportunity to learn what is assessed. They should also be constructed to allow students with special needs and those who are learning English opportunities to demonstrate their knowledge appropriately.

**Proper use:** Productive use of performance assessments, like proper use of standardized tests, should be aimed at revealing areas needing improvement and should lead to curriculum and professional learning supports that can result in powerful learning outcomes for all students. Additionally, tools and protocols, including technological tools, are needed to support the design and use of performance assessment. For example, tools such as task blueprints, rubric specifications, and training and scoring protocols should be developed to support proper use of performance assessments. Finally, as other countries have found, using assessments for information rather than sanctions allows the development of more ambitious tasks aimed at higher standards, and less corruption of the assessment system. This framework for assessment has driven stronger learning and higher achievement in many nations abroad.

Many nations have developed strategies to monitor and improve assessment quality. In Hong Kong, for example, to guide the process of assessment reform, the Education Bureau has implemented a School Development and Accountability Framework which emphasizes school self-evaluation, plus external peer evaluation, using a set of performance indicators. The bureau promotes the use of multiple forms of assessment in schools including projects, portfolios, observations, and examinations, and looks for the variety of assessments in the performance indicators used for school evaluation.<sup>21</sup> For example, the performance indicators ask, “Is the school able to adopt varied modes of assessment and effectively assess students’ performance in respect of knowledge, skills, and attitude?” and “How does the school make use of curriculum evaluation data to inform curriculum

planning?”<sup>22</sup> This practice of examining school practices and the quality of assessments through an inspection or peer review process is also used in Australia and Great Britain to improve teaching by using standards as a tool for sharing knowledge and reflecting on practice.

## Federal Policy Recommendations

Performance assessment is a key component in a balanced assessment system that responds to fast-paced changes placing greater demands on education and knowledge development in the United States and around the world. Images of what students will need to *do* with their knowledge should help shape formulations of curriculum, instruction, and assessment policy at the national, state, and local levels. As a starting point for the development of the next generation of assessments, we must begin with a vision of our young people as lifelong learners who deeply understand core concepts and modes of inquiry within the disciplines, and who can also work across disciplines to evaluate evidence, frame and solve problems, express and defend their ideas, and create new ideas, technologies, and solutions.

Many efforts are under way to refine standards for learning at the state level and by consortia of states collaborating under the auspices of the Council for Chief State School Officers and Achieve, Inc., a national organization of governors, business leaders, and education leaders. These efforts seek to ensure that standards are internationally benchmarked and are fewer, higher, and deeper. It is critical that new assessments be developed in the context of new standards and in relation to curriculum frameworks that ensure the content and skills can be taught coherently and well. To accomplish this federal policy should:

**Fund an intensive development effort** that enables states and consortia of states, in collaboration with development experts in federal labs, centers, nonprofit organizations, and universities, to

- develop, validate, and test high-quality performance assessments that are part of balanced assessment systems which are guided by thoughtful, coherent standards and curriculum frameworks;
- train the field of practitioners—ranging from psychometricians to a new generation of state and local curriculum and assessment

specialists to teachers—who can be skillfully involved in the development, administration, and scoring of these assessments in valid and reliable ways; and

- conduct high-quality research on the validity, reliability, instructional consequences, and equity consequences of these assessments.

**Encourage improvements in federal, state, and local assessment practice in the following ways:**

- Provide incentives and funding for states to refine their existing state assessments and introduce related high-quality locally administered performance assessments that evaluate critical thinking and applied skills. Support states in making such assessments reliable, valid, and practically feasible through teacher professional development and scorer training, moderated and audited scoring systems, and calibration systems, as well as research.
- As part of these efforts, develop more appropriate assessments and accommodations for special education students and English language learners by underwriting efforts to strengthen the validity and reliability of existing performance assessments for these populations, properly adapt new assessments under construction, and create, as needed, new assessments of performance in the content areas for these students, based on professional testing standards that consider principles of universal design as well as specific needs for valid assessment of students in these groups.
- To model high-quality items and better measure the standards, support the further development and implementation of the new blueprints, already under way, for the National Assessment of Educational Practice (NAEP), which include more performance-oriented items that evaluate students' abilities to evaluate evidence, solve problems, and explain and defend their ideas. These kinds of tasks were part of NAEP when it was first launched in the 1960s, and are common in other nations' large-scale assessments, as well as in PISA. Their introduction would need to be incorporated carefully

over time in a planful fashion that maintains existing trend data and continues to enable comparisons among states and over time.

**Enable the incorporation of new assessments into the NCLB accountability system in the following ways:**

- Replace the current “status model” for measuring school progress with a Continuous Progress Index that sets expectations for schools—and groups of students within them—to show progress on an index of measures that include multiple assessments of student learning, including performance measures, as well as school progression and graduation rates. In such an index, which reports information on multiple indicators and then combines them for tracking overall progress, states could choose to include subject areas beyond reading and mathematics, such as writing, science, and history—which are important in their own right and essential to encourage and evaluate students’ literacy skills as they are applied in the content areas. Within a given subject, the index could accommodate assessments of student learning that capture a wider array of skills—including the more complex inquiry and problem-solving skills demanded by twenty-first-century jobs and colleges. Such an index would reduce incentives to narrow the curriculum. It would evaluate students’ growth over time across the entire achievement continuum, thus focusing attention on progress in all students’ learning, not just those who fall at the so-called “proficiency bubble,” reducing ceiling effects, and recognizing schools’ gains with students who score well below and above a single cut score. The CPI would also encourage greater inclusion and more appropriate measurement of gains for special education students and English language learners by tracking gains at all points along the continuum and by incorporating the results of appropriate measures.

## **Conclusion**

Current accountability reforms are based on the idea that standards can serve as a catalyst for states to be explicit about learning goals, and the act of measuring progress toward meeting these standards is an important force

toward developing high levels of achievement for all students. However, an on-demand test taken in a limited period of time on a single day cannot measure all that is important for students to know and be able to do. As described by Achieve, Inc., the limitation of traditional on-demand tests is that they cannot measure many of the skills that matter most for success in the worlds of work and higher education:

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.<sup>23</sup>

Balanced systems of assessment that include performance assessments have the potential to strengthen curriculum and instruction by evaluating the full range of standards in valid and appropriate ways, providing rich information about student learning that is useful to classroom teachers, and providing diverse means for students to demonstrate their learning. Developed carefully and used properly, such assessments can stimulate more thoughtful teaching, become an engine for ongoing improvement and professional development, and create a commitment to standards that shape more powerful learning.

The views expressed in this chapter are those of the authors and do not necessarily represent those of the Alliance for Excellent Education.

### About the Authors

**Linda Darling-Hammond** is Charles E. Ducommun Professor of Education at Stanford University, where she has launched the School Redesign Network and the Stanford Center for Opportunity Policy in Education (SCOPE). Her research, teaching, and policy work focus on issues of school restructuring, teacher quality, and educational equity. Dr. Darling-Hammond has served as faculty sponsor for the Stanford Teacher Education Program, where she helped to introduce performance-based portfolio assessments for pre-service teachers, and cofounded the Performance Assessment for California Teachers with colleagues from eleven other universities. Previously, she served on the

National Board for Professional Teaching Standards and supported its development of a new model of performance assessment for accomplished teachers, and she chaired the standards drafting committee of the Interstate New Teacher Assessment and Support Consortium as it developed new standards for beginning teachers and piloted performance assessments to evaluate the standards. She also chaired the New York State Council on Curriculum and Assessment as it redesigned the state standards and introduced new performance elements to the Regents testing system. Dr. Darling-Hammond is a former president of the American Educational Research Association and a member of the National Academy of Education. Among her more than three hundred publications are *Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able to Do*, *The Right to Learn: A Blueprint for Schools That Work*, and *Authentic Assessment in Action: Studies of Schools and Students at Work*.

**Raymond L. Pecheone** is the co-executive director of the Stanford School Redesign Network LEADS Network. LEADS is an executive educational leadership program that builds partnerships between schools of business and education to bring interdisciplinary perspectives and knowledge bases to the work of K–12 district and school reformers. Dr. Pecheone also serves as the director of the Performance Assessment for California Teachers (PACT) program. PACT is a consortium of thirty-two California universities that have joined together to develop a reliable and valid licensure assessment of pre-service teaching. He also leads and directs a performance-based student assessment project that is aligned to college- and workplace-readiness skills and includes work in California (the Stanford Bay Area Assessment Consortium), and nationally in partnerships with the Asia Society and the State of Ohio. Prior to Stanford, Dr. Pecheone was the Connecticut bureau chief for curriculum, research, and assessment. In this role, he directed the First Assessment Development Laboratory for the National Board for Professional Teaching Standards and cofounded the Interstate New Teacher Assessment and Support Consortium, housed at the Council of Chief State School Officers. He supported the redesign of New York State's Regents Examinations, and served as a consultant to ETS in the development and validation of a national performance-based assessment test for school administrators.

---

<sup>1</sup> No Child Left Behind Act of 2001, Public Law 107-110, 107th Cong., 1st sess., Sec. 1111, b, I, vi.

<sup>2</sup> This section draws from L. Darling-Hammond and G. H. Wood, *Refocusing Accountability: Using Performance Assessments to Enhance Teaching and Learning* (Washington, DC: Forum for Education & Democracy, 2008).

<sup>3</sup> S. Messick, "Validity," in *Educational Measurement*, ed. R. L. Linn, 13–103 (Washington, DC: National Council of Measurement in Education and the American Council on Measurement in Education, 1989).

<sup>4</sup> L. A. Shepard, R. J. Flexer, E. H. Heibert, S. F. Marion, V. Mayfield, and T. J. Weston, *Effects of Introducing Classroom Performance Assessments on Student Learning*, CSE Technical Report 394 (Los Angeles: National Center for Research on Evaluation, Standards, and Students Testing [CRESSST], Graduate School of Education & Information Studies, University of California, Los Angeles, 1995).

<sup>5</sup> For a summary see L. Darling-Hammond and E. Rustique-Forrester, "The Consequences of Student Testing for Teaching and Teacher Quality," in *The Uses and Misuses of Data in Accountability Testing*, ed. Joan Herman and Edward Haertel, 289–319 (Malden, MA: Blackwell Publishing, 2005).

<sup>6</sup> Appalachia Educational Laboratory, "Five Years of Reform in Rural Kentucky," *Notes from the Field: Educational Reform in Rural Kentucky* 5, no. 1 (February 1996); Charleston, WV: Author; B. M. Stecher, S. Barron, T. Kaganoff, and J. Goodwin, *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, 1998); B. L. Whitford and K. Jones, *Accountability, Assessment, and Teacher Commitment: Lessons from Kentucky's Reform Efforts* (Albany: State University of New York, 2000); D. Koretz, B. Stecher, and E. Deibert, *The Vermont Portfolio Program: Interim Report on Implementation and Impact, 1991–92 School Year* (Santa Monica, CA: RAND, 1992); Shepard et al., *Effects of Introducing Classroom Performance Assessments*.

<sup>7</sup> L. Darling-Hammond, J. Ancess, and B. Falk, *Authentic Assessment in Action* (New York: Teachers College Press, 1995); M. Kornhaber and H. Gardner, *Varieties of Student Excellence* (New York: National Center for Restructuring Education, Schools, and Teaching, Teachers College, Columbia University, 1993).

<sup>8</sup> L. Darling-Hammond, E. Rustique-Forrester, and R. Pecheone, *Multiple Measures Approaches to High School Graduation* (Stanford: Stanford University, School Redesign Network, 2005).

<sup>9</sup> R. Mitchell, *Testing for Learning* (New York: Free Press, 1992).

<sup>10</sup> Connecticut State Board of Education, 2000.

<sup>11</sup> Koretz, Stecher, and Deibert, *The Vermont Portfolio Program*.

<sup>12</sup> R. Murnane and F. Levy, *Teaching the New Basic Skills* (New York: Free Press, 1996).

<sup>13</sup> J. D. Jordan, "R.I. Graduation Rate Up: What Happened to the 13,163 Ninth Graders of 2004?" *Providence Journal* (March 21, 2009).

<sup>14</sup> This section draws on L. Darling-Hammond and L. McCloskey, "Assessment for Learning Around the World: What Would It Mean to Be Internationally Competitive?" *Phi Delta Kappan* 90, no. 4 (2008): 263.

<sup>15</sup> Swedish National Agency for Education, *The Swedish School System: Compulsory School*, 2005, <http://www.skolverket.se/sb/d/354/a/959> (accessed May 31, 2008).

<sup>16</sup> M. A. Eckstein and H. J. Noah, *Secondary School Examinations: International Perspectives on Policies and Practice* (New Haven: Yale University Press, 1993); S. O'Donnell, *International Review of Curriculum and Assessment Frameworks, Comparative Tables and Factual Summaries—2004* (London: Qualifications and Curriculum Authority, 2004).

<sup>17</sup> Eckstein and Noah, *Secondary School Examinations*, p. 230.

<sup>18</sup> Qualifications and Curriculum Authority, "England: Assessment Arrangements," 2008, <http://www.inca.org.uk/1315> (accessed May 27, 2008); <http://education.qld.gov.au/corporate/newbasics/html/richtasks/richtasks.html> (accessed April 1, 2009).

<sup>19</sup> J. K. Chan, K. J. Kennedy, F. W. Yu, and P. Fok, "Assessment Policy in Hong Kong: Implementation Issues for New Forms of Assessment," *Hong Kong Institute of Education*, 2008, <http://www.iaea.info/papers.aspx?id=68> (accessed September 12, 2008).

<sup>20</sup> M. Dowling, "Examining the Exams," [http://www.hkeaa.edu.hk/files/pdf/markdowling\\_e.pdf](http://www.hkeaa.edu.hk/files/pdf/markdowling_e.pdf) (accessed September 14, 2008).

<sup>21</sup> Chan et al., "Assessment Policy in Hong Kong"; "Performance Indicators for Hong Kong Schools, 2008 with Evidence of Performance," 2008, [http://www.edb.gov.hk/FileManager/EN/Content\\_6456/pi2008%20eng%205\\_5.pdf](http://www.edb.gov.hk/FileManager/EN/Content_6456/pi2008%20eng%205_5.pdf) (accessed September 12, 2008).

<sup>22</sup> Quality Assurance Division of the Education Bureau, "Performance Indicators for Hong Kong Schools."

<sup>23</sup> Achieve, Inc., *Do Graduation Tests Measure Up? A Closer Look at State High School Exit Exams*, Executive Summary (Washington, DC: Achieve, Inc., 2004).