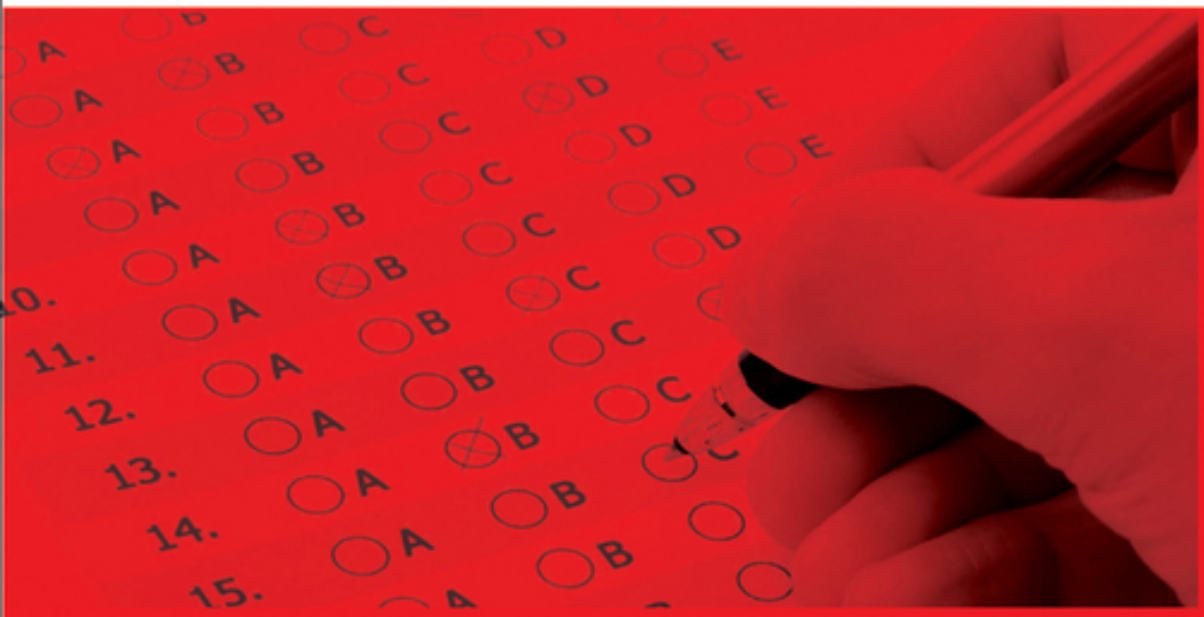
LINDA DARLING-HAMMOND
and FRANK ADAMSON



BEYOND THE BUBBLE TEST

HOW PERFORMANCE ASSESSMENTS SUPPORT 21ST CENTURY LEARNING

JOSSEY-BASS™
A Wiley Brand

Contents

List of Illustrations

List of Tables

Beyond the Bubble Test

How Performance Assessments Support 21st Century Learning

Linda Darling-Hammond

Frank Adamson

JOSSEY-BASS™
A Wiley Brand

Cover design by Wiley

Cover photograph © BrianAJackson | Thinkstock

Published by Jossey-Bass

A Wiley Brand

One Montgomery Street, Suite 1200, San Francisco, CA 94104-4594—
www.josseybass.com

Library of Congress Cataloging-in-Publication Data has been applied for and is on file with the Library of Congress.

ISBN 978-1-118-45618-7 (cloth); ISBN 978-1-118-88929-9 (ebk); ISBN 978-1-118-88932-9 (ebk)

ACKNOWLEDGMENTS

This book summarizes research and lessons learned regarding the development, implementation, and consequences of performance assessments. The volume examines experiences with and lessons from large-scale performance assessment in the United States and abroad, including technical advances, feasibility issues, teacher engagement, policy implications, uses with English Language Learners, and costs.

The work was guided by an Advisory Board of education researchers, practitioners, and policy analysts, ably chaired by Richard Shavelson. The board shaped specifications for commissioned papers that became some of these chapters and reviewed these papers upon their completion. We are grateful to Advisory Board members Eva Baker, Christopher Cross, Nicholas Donahue, Michael Feuer, Edward Haertel, Jack Jennings, Peter McWalters, Lorrie Shepard, Guillermo Solano-Flores, Brenda Welburn, and Gene Wilhoit.

The contributors to this book would like to thank all the educators and other innovators over many years who have devoted hundreds of thousands of hours to developing and implementing thoughtful curriculum and assessments that support students and teachers in their learning. We would also like to thank Barbara McKenna, who ably ushered earlier versions of these papers into production, Sonya Keller for her very helpful and thorough editorial assistance, and Samantha Brown for her help securing permissions for the entries in this volume.

A number of chapters in this volume (2–6 and 8–11) draw in part on papers that were previously published by the

Stanford Center for Opportunity Policy in Education. In addition, chapter 7 draws on a paper published by the Center for American Progress. All are included here with permission.

Research for this volume was supported by the Ford Foundation, the Hewlett Foundation, the Nellie Mae Educational Foundation, and the Sandler Foundation, to whom we are grateful. The opinions expressed in this book do not necessarily reflect the position of any of these organizations.

THE AUTHORS

Jamal Abedi, professor of education at the University of California, Davis, specializes in educational and psychological assessments. His research focus is testing for English language learners and issues concerning the technical characteristics and interpretations of these assessments. From 2010 to the present, Abedi has served as a member of the Technical Advisory Committee of the SMARTER Balanced Assessment Consortium. Before then, he served on the expert panel of the US Department of Education's LEP Partnership and he was founder and chair of AERA's Special Interest Group on Inclusion and Accommodation in Large-Scale Assessment. In 2008, the California Educational Research Association gave him its Lifetime Achievement Award. Abedi received his PhD from Vanderbilt University.

Frank Adamson, a policy and research analyst at the Stanford Center for Opportunity Policy in Education (SCOPE), currently focuses on the adoption of assessments of deeper learning and twenty-first-century skills at the state, national, and international levels. He also conducts research on educational equity and opportunities to learn and has published on teacher salary differences within labor markets in New York and California. Prior to joining SCOPE, Adamson worked at AIR and SRI International designing assessments, evaluating US education initiatives, and developing international indicators for the OECD and UNESCO. He received an MA in sociology and a PhD in international comparative education from Stanford University.

Jillian Chingos (previously Jillian Hamma) is currently a sixth-grade teacher at Alpha: Blanca Alvarado Middle

School in San Jose, California. Chingos attended Dartmouth College, where she majored in English, minored in public policy, and received her teaching credential. She previously worked at the Stanford Center for Assessment, Learning, and Equity, developing and researching performance assessments.

David T. Conley is professor of educational policy and leadership and founder and director of the Center for Educational Policy Research (CEPR) at the University of Oregon. He is also the founder, chief executive officer, and chief strategy officer of the Educational Policy Improvement Center and president of CCR Consulting Group, both in Eugene and Portland, Oregon. Through these organizations, he conducts research on a range of topics related to college readiness and other key policy issues with funding provided by grants and contracts from a range of national organizations, states, school districts, and school networks. His line of inquiry focuses on what it takes for students to succeed in postsecondary education. His latest publication, *Getting Ready for College, Careers, and the Common Core*, was recently published by Jossey Bass (for more information, see www.collegecareerready.com).

Linda Darling-Hammond is Charles E. Ducommun professor of education and faculty director of the Stanford Center for Opportunity Policy in Education at Stanford University. Darling-Hammond is a former president of the American Educational Research Association and a member of the National Academy of Education. Her research and policy work focus on issues of educational equity, teaching quality, school reform, and performance assessment. In 2008, she served as director of President Obama's education policy transition team. Her book *The Flat World and Education: How America's Commitment to Equity Will Determine Our Future* received the coveted Grawemeyer

Award in 2012. Her most recent book is *Getting Teacher Evaluation Right: What Really Matters for Effectiveness and Improvement* (2013).

Beverly Falk is professor and director of the graduate programs in early childhood education at the School of Education, City College of New York. Her areas of expertise include early childhood education, early literacy, performance assessment, school change, teacher education, and teacher research. She has served in a variety of educational roles: classroom teacher; school founder and director; district administrator; and consultant, fellow, and leader in schools, districts, states, and national organizations. Currently she is editor of the *New Educator* and senior scholar at the Stanford Center for Assessment, Learning and Equity. Falk received her EdD from Teachers College, Columbia University.

Ann Jaquith is associate director at Stanford Center for Opportunity Policy in Education. She has worked on a variety of performance assessment projects undertaken to reform schools in New York, Ohio, and California. As a former teacher and administrator, her expertise is in building the instructional and leadership capacity needed to use performance assessments to improve instruction and student learning. Her research interests include studying how instructional capacity gets built at different levels of the system and examining the practices professional development providers use that change instruction and improve student learning. She received her PhD in curriculum and teacher education from Stanford University.

Stuart Kahl is founding principal and CEO of Measured Progress as Advanced Systems in Measurement and Evaluation. A former elementary and secondary teacher, he worked for the Education Commission of the States, the University of Colorado, and RMC Research Corporation. A

frequent speaker at industry conferences, Kahl also serves as a technical consultant to various education agencies. He has been recognized for his work in the areas of standard setting for non-multiple-choice instruments and the alignment of curriculum and assessment. Kahl received his PhD from the University of Colorado.

Suzanne Lane is professor at the University of Pittsburgh. Her recent research focuses on the implications for the next generation of assessments based on the lessons from classroom instruction and achievement in the 1990s, the assessment of twenty-first-century thinking skills, and the interplay among a theory of action, validity, and consequences. Lane has been the president of the National Council on Measurement in Education (2003–2004) and vice president of Division D of the American Educational Research Association (2002–2003). She received a PhD in research methodology, measurement, and statistics from the University of Arizona.

William Montague is a student at the University of Virginia School of Law. He began his career as a high school English teacher in Roanoke Rapids, North Carolina, as a member of Teach For America. He went on to work for Independent Education, an association of independent schools in the Washington, DC, area. While there, he collaborated on a number of projects with the organization's executive director, Thomas Toch, a longtime education writer and policy analyst. Montague received his BA from the University of Virginia, where he majored in economics and history.

John Olson is senior partner of Assessment Solutions Group (ASG), which he cofounded in 2008. He is also president of the consulting business he founded in 2006, Olson Educational Measurement and Assessment Services, which provides technical assistance and support to states,

school districts, federal bodies, testing companies, researchers, and others. He has more than thirty years of experience managing and consulting on a variety of measurement and statistical issues for international, national, state, and local assessment programs through his work at Harcourt Assessment, the Council for Chief State School Officers, the American Institutes for Research, and the Education Statistics Services Institute. He served in a number of leadership roles for the National Assessment of Educational Progress at the Educational Testing Service. Olson holds a PhD in educational statistics and measurement from the University of Nebraska–Lincoln.

Margaret Owens is currently a teacher at Mission High School in San Francisco. She earned her teaching credential and MA from Stanford University. Her studies focused on new pedagogical strategies, such as complex instruction, that bring more collaboration and engagement to students historically alienated in mathematics. Prior to her teaching career, she studied political science at Stanford with a focus on American education.

Raymond Pecheone is professor of practice at Stanford University and the founder and executive director of the Stanford Center for Assessment Learning, and Equity (SCALE). Under Pecheone, SCALE focuses on the development of performance assessments and performance-based systems for students, teachers, and administrators at the school, district, and state levels. Prior to launching SCALE, Pecheone was the bureau chief for curriculum, research, and assessment in the Connecticut State Department of Education; codirector of the first Assessment Development Lab for the National Board for Professional Teaching Standards; and project director to support the redesign of the New York State Regents. Most recently, Pecheone and SCALE are developing the performance assessment specifications and tasks for the

Smarter Balanced national assessment system. He received his PhD from the University of Connecticut in measurement and evaluation.

Lawrence O. Picus is vice dean for faculty affairs and professor at the Rossier School of Education, University of Southern California. He is an expert in the area of public financing of schools, equity and adequacy of school funding, school business administration, education policy, linking school resources to student performance, and resource allocation in schools. His current research interests focus on adequacy and equity in school finance, as well as efficiency and productivity in the provision of educational programs for PreK–12 children. Picus is past president of the Association for Education Finance and Policy, has served on the EdSource board of directors for twelve years, and has consulted extensively on school finance issues in more than twenty states. He earned a PhD in public policy analysis from the RAND Graduate School, an MA in social science from the University of Chicago, and a BA in economics from Reed College.

Ed Roeber is a consultant at Assessment Solutions Group (ASG). He has served as state assessment director in the Michigan Department of Education, director of student assessment programs for the Council for Chief State School Officers, vice president of Measured Progress, and adjunct professor at Michigan State University. For ASG and the other organizations, he advises states and other organizations on student assessment-related programs and functions. Currently he is a consultant on student assessment to several organizations (Michigan Assessment Consortium, Michigan State University, and Wisconsin Center for Educational Research/University of Wisconsin). He has written extensively about educational assessment, consulted with a number of agencies and organizations, and spoken frequently about student assessment. He has a

PhD in educational measurement from the University of Michigan.

Brian Stecher is a senior social scientist and associate director of RAND Education and professor at the Pardee RAND Graduate School. His research focuses on measuring educational quality and evaluating education reforms, with an emphasis on assessment and accountability systems. During his more than twenty years at RAND, he has directed prominent national and state evaluations of No Child Left Behind, mathematics and science systemic reforms, and class size reduction. His measurement-related expertise includes test development, test validation, and the use of assessments for school improvement. Stecher has served on expert panels relating to standards, assessments, and accountability for the National Academies and is currently a member of the Board on Testing and Assessment. He received his PhD from the University of California, Los Angeles.

Thomas Toch is senior managing partner for public policy engagement at the Carnegie Foundation. He also serves as director of the Carnegie Foundation's Washington, DC, office. He is a founder and former codirector of the think tank Education Sector and a former guest scholar at the Brookings Institution, and he has taught at the Harvard Graduate School of Education. He helped launch *Education Week* in the 1980s. He spent a decade as the senior education correspondent at *US News and World Report* and has contributed to the *Atlantic*, the *New York Times*, and other national publications. His work has twice been nominated for National Magazine Awards. He is the author of two books on American education, *In the Name of Excellence* (Oxford University Press) and *High Schools on a Human Scale* (Beacon Press).

Barry Topol is managing partner of Assessment Solutions Group (ASG). He leads ASG in providing assessment cost, management, and state accountability systems analysis and consulting to states, universities, and other nonprofit institutions. Since forming ASG in 2009, Topol and ASG have worked with a number of states and the Partnership for Assessment of Readiness for College and Careers, and Smarter Balanced Assessment Consortium to assist them in designing their assessment and accountability systems to be more effective and efficient. Topol designed ASG's assessment cost model, the only model in the industry that can be used to determine the appropriate price for any assessment. Topol has a BA in economics from the University of California, Los Angeles and an MBA from the Anderson School of Management at the University of California, Los Angeles.

Laura Wentworth is director of the Stanford University/San Francisco Unified School District Partnership at Silver Giving Foundation. She focuses on supporting the link between research and practice, with special attention to the subject of assessment. As a public school teacher, she and other school leaders introduced the International Baccalaureate Primary Year Program, including the use of a portfolio assessment system for kindergarten through fifth grade. After her work as a teacher, she began researching issues in assessment, including policy issues for English learners, exit exam policies, and performance assessments. Currently she directs a university-district partnership aimed at helping practitioners use research to inform their decision making that includes several assessment projects. She received her PhD in education policy from Stanford University.

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 interactive skills that require collaborative invention and problem solving (see figure 1.1).

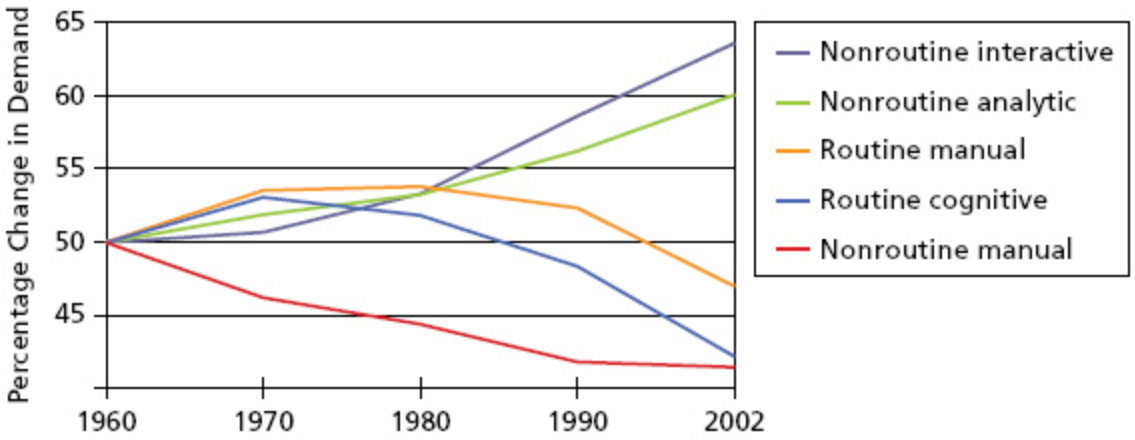


Figure 1.1 How the Demand for Skills Has Changed: Economy-Wide Measures of Routine and Nonroutine Task Input

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

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