



Performance Assessment in an Era of Standards-Based Educational Accountability

Brian Stecher
RAND Corporation

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

© 2010 Stanford Center for Opportunity Policy in Education. All rights reserved.

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

Citation: Stecher, B. (2010). *Performance Assessment in an Era of Standards-Based Educational Accountability*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Stanford Center for Opportunity Policy in Education

Barnum Center, 505 Lasuen Mall

Stanford, California 94305

Phone: 650.725.8600

scope@stanford.edu

<http://edpolicy.stanford.edu>



Table of Contents

Preface and Acknowledgements	i
Introduction	1
Defining Performance Assessment	2
Primer on Large-Scale Testing	9
Recent History of Performance Assessment in Large-Scale Testing	11
Research Findings	20
Performance Assessment in the Context of Standards-Based Accountability	34
Recommendations	37
References	39

Preface and Acknowledgements

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

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

Christopher Cross, Chairman, Cross & Jofus, LLC

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

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

Edward Haertel, Jacks Family Professor of Education, Stanford University

Jack Jennings, President and CEO, Center on Education Policy

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

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

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

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

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

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

The papers listed below examine experiences with and lessons from large-scale performance assessment in the United States and abroad, including technical advances, feasibility issues, policy implications, uses with English language learners, and costs.

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

An overview of all these papers has also been written and is available in electronic and print format:

- ~ Linda Darling-Hammond and Frank Adamson, *Beyond Basic skills: The Role of Performance Assessment in Achieving 21st Century Standards of Learning*.

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

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

Introduction

Performance assessment—judging student achievement on the basis of relatively unconstrained responses to relatively rich stimulus materials—gained increasing favor in the United States in the late 1980s and 1990s. At least one national commission advocated the replacement of multiple-choice tests with performance assessments (National Commission on Testing and Public Policy, 1990); the National Assessment of Educational Progress conducted extensive pilot testing of hands-on assessment tasks in science and mathematics (Educational Testing Service, 1987); and performance assessments were adopted by a number of state testing programs, including Vermont, Kentucky, Maryland, Washington, and California. Yet, despite this enthusiasm, performance assessment has almost disappeared from large-scale K-12 testing in the U.S. in the intervening years (Council of Chief School Officers, 2009). A number of factors account for the failure of performance assessment to capture a large role in achievement testing in the U.S., and this history can inform educators and education policymakers looking for better ways to test students and schools in an era of standards-based accountability.

The paper is organized as follows. First, it presents a definition of performance assessment and suggests ways to classify different types of performance tasks. Then background information on large-scale testing is provided to familiarize readers with key terms and concepts. Third, there is a review of the recent history of performance assessments in the U.S. and the claims supporting this approach to measuring student performance. Following that, the paper summarizes research on the quality, impact, and burden of performance assessments used in large-scale K-12 achievement testing. The paper concludes with a discussion of the relevance of performance assessment to contemporary standards-based educational accountability and offers recommendations to support effective use of this form of assessment.

Defining Performance Assessment

In this paper, the terms *test* and *assessment* are used interchangeably, although “multiple-choice” is paired with “test” and “performance” with “assessment.” Individual multiple-choice questions are called “items,” and individual performance activities are called “tasks.”

Constructed Versus Selected Response

For many educators, performance assessment is most easily defined by what it is *not*; specifically, it is not multiple-choice testing. In performance assessment, rather than choosing among predetermined options, the examinee must either construct or supply an answer, produce a product, or perform an activity (Madaus & O’Dwyer, 1999). From this perspective, performance assessment encompasses a very wide range of activities, from completing a sentence with a few words (short-answer) to writing a thorough analysis (essay) to conducting a laboratory investigation and writing a descriptive analysis of the process (hands-on). Given this range, it is surprising how often people make general claims about performance assessment without differentiating among types.

Multiple-choice tests are an easier target for generalization. Advocates of performance assessment argue that the fixed set of responses in multiple-choice tests (and their cousins, true-false tests and matching tests), are inauthentic. That is, the tests do not reflect the nature of performance in the real world, which rarely presents people with structured choices. With the possible exception of a few game shows, one demonstrates his or her ability in the real world by applying knowledge and skills in settings where there are no pre-determined options. A person balances his or her checkbook; buys ingredients and cooks a meal; reads a new article in the paper and frames an opinion of the rightness or wrongness of the argument; assesses a customer’s worthiness for a mortgage; interviews a patient, then orders tests and diagnoses the nature of his or her disease; listens to a noisy engine running at low and high RPM and identifies the likely cause; etc. Even in the context of school, the typical learning activity involves a mix of skills and culminates in a complex performance: a homework assignment, a persuasive letter, a group project, a research paper, a first-down, a band recital, a sketch, etc. Rarely does a citizen or a student have to choose among four distinct alternatives.

Multiple-choice tests are also criticized because the stimulus materials tend to be very limited (usually short passages or problem statements with simple figures or diagrams) and because the demands of the format encourage test developers to focus on declarative knowledge (knowing that) or procedural knowledge (knowing how) rather than schematic knowledge (knowing why) or strategic knowledge (knowing when, where, and how our knowledge applies) (Shavelson, Ruiz-Primo & Wiley, 2005).

Advocates of performance assessment argue that the limits of multiple-choice tests can

be overcome, in part, by replacing them with tests in which respondents have to construct responses instead of selecting responses from a pre-determined set. Such situations are common in schools, suggesting that they are a more authentic way to judge what students have learned. In almost every subject, homework includes open-ended tasks: mathematics problems where students work out the solution (even earning partial credit for partial solutions); English assignments where students reflect on a text from their own point of view; chemistry laboratory where students record their observations while conducting an experiment; choral music ensembles where students sing a score, etc. All of these situations are forms of performance assessment. Once students leave school, the tasks they will encounter in their lives will be even less structured, both in terms of the assignments and the nature of the responses. For this reason, many educators use the term “authentic assessment” to emphasize the similarity of the task to the real world.

If performance events are so pervasive in our lives and performance assessments have such advantages over multiple-choice tests, why do we rely almost exclusively on multiple-choice tests when making important judgments about students (promotion to the next grade level, graduation), schools (making adequate yearly progress), and, recently, teachers (value-added judgments about effectiveness)? Should educators and policymakers be trying to overturn the “tyranny of multiple-choice testing” that exists in educational accountability systems?

Definition of Performance Task and Performance Assessment

In the literature, different authors use the term “performance assessment” to mean different things. Some emphasize the cognitive processes demanded of the students, some the format of the desired response, and others the nature and content of the actual response (Palm, 2008). These differences in emphasis underscore one of the lingering problems facing performance assessment, which is that different educators and policymakers have different implicit meanings for the term.

For the purposes of this paper, I define a performance assessment primarily in terms of the performances required of test takers involved in large-scale assessments:

A performance task is a structured situation in which stimulus materials and a request for information or action are presented to an individual, who generates a response that can be rated for quality using explicit standards. The standards may apply to the final product or to the process of creating it. A performance assessment is a collection of performance tasks.

This definition has four important elements. First, each task must occur in a “structured situation,” meaning the task is constrained with respect to time, space, access to materials, etc. The standardized structure makes it possible to replicate the conditions, so the same assessment can be administered to different people and their performances

can be compared. The requirement that there be structure with respect to administrative conditions does not exclude from consideration complex, extended tasks, such as conducting a scientific experiment and reporting the results; instead, structure insures that tasks can be replicated at different times and in different places. The requirement for structure does exclude from consideration a number of familiar incidents of assessment, including oral examinations in which the examiner asks a different question of each student; ratings of behaviors in natural settings, such as ratings of athletes during competition (Nadeau & Godbour, 2008) or observations of students' literacy behaviors (Meisels, Xue & Shablott, 2008); and assessments that attempt to compare change in performance over time (Wentworth et al., 2009).

Second, each performance task contains some kind of stimulus material or information that serves as the basis for the response. In this respect, performance tasks can be very similar to multiple-choice items. They might begin with a mathematical problem, a text passage, a graph, or figure. However, because the response options are unconstrained, the stimulus material does not have to lead to four specific choices. That freedom permits performance tasks to include more varied, complex, and novel stimuli than are typically used in multiple choice assessments.¹

Third, the task must have directions indicating the nature of the desired response. The directions can be part of the stimulus materials (e.g., How many 3-inch by 4-inch tiles would it take to cover a circular dance floor 8 feet in diameter?) or separate from them (e.g., Read these descriptions of the Battle of the Bulge written by soldiers on opposing sides and write an essay to support or refute the claim that the Allied and Axis soldiers were more similar than different.) The key feature of these directions is that they be explicit enough so two test takers would have similar understanding of what they were asked to do. Because the responses can be much broader than the responses for multiple-choice tests, vague queries (e.g., What do you think about that?) that can be widely interpreted must be avoided.

Fourth, the task must prompt responses that can be scored according to a clear set of standards. It is usually the case that the standards are fully developed before the task is given: If the task developer does not know what constitutes a good response, it is unlikely that the task is measuring something clearly enough to be useful. In some cases, however, the scoring rubrics will need to be elaborated on the basis of responses received. For example, students may come up with legitimate responses the developers did not anticipate.

Types of Performance Assessments

The definition we gave above admits a wide range of performance assessments, and it is helpful to have a system for classifying them into categories for discussion.

1. These stimuli can be used in multiple-choice questions, but they are less common, and complex stimuli are infrequent in large-scale, multiple choice testing.

Classifying Based on Stimulus Materials and Response Options. We suggest a two-way classification scheme based on the structural characteristics of the task, particularly the nature of the stimulus materials and the nature of the response options. (This scheme is inspired by the work of Baxter and Glaser (1998) discussed subsequently.) The stimulus materials can be classified in terms of complexity along a dimension that runs from simple to complex. A math task that asks the student to solve an equation for X represents a relatively simple stimulus. In contrast, a language arts task that asks the student to read an essay and a poem and to look at a reproduction of a painting as the basis for a comparative task presents a relatively complex set of stimulus materials. Similarly, the response options can be classified in terms of freedom along a dimension that runs from constrained to open. A short-answer history question requiring the test taker to fill in a short phrase to correctly place an event offers a relatively constrained response space. In comparison, a science task in which students are given a set of leaves to observe and are asked to create at least two different classification schemes and arrange the leaves into groups based on each scheme offers a relatively open range of responses.

By crossing the stimulus and response dimensions, we create four quadrants that can be used to classify all performance tasks. A written, short-answer (fill in the blank) question is an example of a relatively simple, relatively constrained task. A math word problem that requires setting up equations, use of a graphing calculator, and other calculations is an example of a relatively simple, relatively open task. Such tasks can also be found in many international examinations, such as the A-level examinations used in England, in which relatively simple prompts require extended, open-ended responses demonstrating the application of knowledge in real-world settings (Darling-Hammond, 2010).

The aforementioned language arts task in which an essay, a poem, and a piece of art serve as prompts for a comparative essay on the similarities and differences in the artists' visions of the subject is an example of a relatively complex, relatively open task. The College and Work Readiness Assessment (CWRA) is another example; it includes complex, open tasks in which students must review a set of materials that might include graphs, interviews, testimonials, or memoranda, extract information from each, assess the credibility of the information, and synthesize it to respond to problem situation (Hersh, 2009). An interesting feature of CWRA is that it can be administered online with both prompts and responses transmitted over the Internet.

At this point, the savvy reader should object that the stimulus-response classification scheme focuses on surface features of the task and ignores more important cognitive and performance aspects of the activity. We agree, as the next paragraph will demonstrate. Yet, it turns out that this simple classification will be useful when thinking about some of the practical aspects of performance assessment, including feasibility, burden, and costs.

Classifying Based on Content Knowledge and Process Skills. Baxter and Glaser (1998) suggest a way to classify science performance tasks by their cognitive complexity, and this approach could be used more generally. They divide the science assessment space into four quadrants depending on whether the process skills demanded are open or constrained, and whether the content knowledge demanded is lean or rich. They provide examples of science tasks corresponding to each quadrant. For example, “Exploring the Maple Copter” is an example of a content-rich-process open task. In this task, high school physics students are asked to design and conduct experiments with a maple seed and develop an explanation of its flight for someone who does not know physics. The two-way content-process classification is helpful for characterizing the cognitive complexity of performance tasks. This distinction is useful when thinking about the inferences that are appropriate to make from scores on performance assessments and the kinds of information that would be needed to validate those inferences.

Classifying Based on Subject Field. The examples above suggest that subject field may be another useful way to classify performance tasks. Expertise in one subject is demonstrated very differently than expertise in another, and performance assessment lets these distinctive styles of thinking and performing come to the fore. For example, “doing science” involves observing events, designing experiments, imposing controls, collecting information, analyzing data, building theories, etc., and rich performance tasks in the sciences incorporate these kinds of skills and behaviors. In contrast, “doing mathematics” is somewhat more abstract. Although mathematics begins with observations of objects in space, the discipline focuses more on manipulating numbers, building facility with operations, translating situations into representations, explicating assumptions, testing theories, etc. Performance assessment in mathematics is more likely to involve solving problems with pencil, paper, and calculators, representing relationships in different forms such as graphs, etc. Similarly, performance assessment in the arts involves performing—music, dance, drawing, etc. In language arts, performance assessments are likely to focus on the written word, reading, comprehending, interpreting, comparing, producing text that describes, explains, persuades, etc. Because of these disciplinary ways of thinking and acting, educators in different fields may be thinking about different very kinds of activities when they refer to performance assessments.

Finally, some performance assessments are designed to be less discipline bound. For example, the Collegiate Learning Assessment includes tasks that are design to measure the integration of skills across disciplines (Klein et al., 2007). Similarly, Queensland in Australia has developed a bank of Rich Tasks that call for students to demonstrate “transdisciplinary learnings” (Darling-Hammond, 2010). Other performance tasks are designed to measure transferable reasoning skills that can be applied in many contexts.

Portfolios

A portfolio is a collection of work, often with personal commentary or self-analysis, assembled by an individual over time as a cumulative record of accomplishment. In most cases, the individual selects the work that goes into the portfolio, making each one unique. This was the case with the Vermont and Kentucky portfolio assessments in the early 1990s, although both became more standardized in the definition of tasks over time, and it is the case with most of the currently popular portfolios in teacher education programs.

According to the definition of performance assessment given above, non-standardized individual collections of work are not performance assessments because each portfolio contains different performance tasks. One student might include a persuasive essay or the solution to a math word problem that was omitted from the portfolio of another student. It is not just narrow-mindedness that leads us to exclude free-choice portfolios from the realm of performance assessment; rather, the lack of standardization actually undermines the value of such collections of work as assessment tools.

The initial Vermont experience is a case in point. Teachers and students jointly selected student work to include in each student's mathematics and writing portfolios. Thus, two portfolios selected from a given teacher's class contained some common pieces and some unique choices. This variation made the portfolios difficult to score. Scoring criteria must be general enough to be applicable to different collections of student work, but such general criteria are difficult to apply consistently to any specific piece of work. The challenge is exacerbated when portfolios from many classes are sent to a common site for scoring. The proportion of common pieces is smaller and the proportion of unique pieces is larger. Researchers found that teachers could not score the Vermont portfolios consistently enough for the scores to be valid for the purpose of comparing schools (Koretz et al., 1994). The lack of standardization for the portfolios made it impossible to develop scoring rubrics that were both general enough to accommodate differences in content and specific enough to enable raters to make similar judgments. Researchers have also reported that teachers have difficulty scoring non-traditional tasks when the tasks contain unfamiliar content (Meier, Rich, & Cady, 2006).

Studies have reported much higher rates of scoring consistency for more standardized portfolios featuring common task expectations and analytic rubrics, like those that were ultimately developed in Kentucky (Measured Progress, 2009).

Another problem with portfolios containing schoolwork is not knowing "whose work is it?" i.e., who contributed to the final product (Gearhart et al., 1993). Many people may contribute to a student's work—the teacher offers comments on a draft, another student reads it and offers suggestions, the student prepares an initial version and then revises it so it looks more polished. In the end, work produced as part of the normal teaching

and learning process often benefits from the contributions of others, so it is difficult to attribute the product to a single individual. An attempt in the Dallas Public Schools to develop a reading/language arts portfolio in the primary grades was criticized for problems with both reliability of scoring and validity of contents (Shapley & Bush, 1999).

For these reasons, free-choice portfolios should not be considered as performance assessments for use in high-stakes large-scale testing. Such portfolios can play a valuable role in the learning process, providing increased opportunities for students to reflect on their work and improve. They can also be helpful as classroom assessment, where the teacher understands the context in which the work was produced and can interpret the student's role appropriately. Portfolios can also provide a cumulative record that can be used to reflect on a students' growth over time. For portfolios to be useful as performance assessments, however, they must be standardized; that is, all students collect the same work products, and those work products are produced under similar conditions. In theory, it is easy to meet the former criterion but more difficult to achieve the latter.

Primer on Large-Scale Testing

Readers reasonably familiar with large-scale testing should feel free to skip the remainder of this section, which is a short primer on testing to introduce ideas that will be important later.

What is a Test?

In simple terms, a test (or assessment) is a sample of knowledge or behaviors selected from a larger *domain* of knowledge. We hope that results on the test allow us to make an *inference* about likely mastery of the domain. Thus scores on a 35-item test of fourth grade mathematics are interpreted as indicators of proficiency on the whole body of fourth grade mathematics.² In some instances, advocates for performance assessment have argued that certain kinds of very rich activities are valuable in their own right as demonstrations of a set of core understandings and abilities, and are not just proxies that allow generalizations about a larger construct (Haertel, 1999). These demonstrations—which combine many kinds of knowledge and skills into a major undertaking that represents the way in which work in that domain is generally done—might include performances like designing and completing an independent science investigation, an independently-designed computer program, or a doctoral dissertation. Others have supported efforts to infuse the curriculum with rich performance tasks because they reveal more about students’ thinking and are useful for instructional planning and classroom assessment. This paper focuses on large-scale testing in which items, including performance tasks, represent a more bounded set of concepts or skills and are viewed as samples from a domain.

In a standards-based system, state *academic content standards* provide descriptions of the domains that have been formally endorsed by policymakers, educators, and the public. Standards writing committees consult scientists, researchers, teachers, and others to craft descriptions of the content domain that serve as the basis for curriculum and assessments. Broader domains, such as mathematics, are usually divided into *sub-domains*, e.g., number and operations, measurement, and geometry. And the sub-domains are further partitioned into more detailed statements about expected knowledge, skills, and/or procedures. One might think that educators long ago reached agreement on the content of the public school curriculum, but, despite many efforts to codify content in the disciplines (e.g., National Research Council, 1996; National Council of Teachers of Mathematics, 2000), state standards still vary considerably in terms of breadth, depth, coverage, and format.

2. In some situations, our interest may be focused on responses to specific test question more than generalization to a broader domain. For example, the teacher who gives a test of spelling words assigned for this week may want to know students ability to spell those particular words. A spelling test at the end of the year might have a broader purpose.

How Are Tests Developed?

The test development process usually begins with decisions about the length of the test and the *format* of the test items. These decisions define the broad parameters for the test. Using the content standards as a guide, developers create *test specifications* that indicate how test items will be distributed across sub-domains and how items will be written in terms of cognitive complexity, i.e., how many will assess recall of facts, application of principles or procedures, synthesis of ideas, etc. The reason for creating test specifications is to make the test as *representative* of the domain as possible, even though it measures only a small portion of it.

Working from the test specifications, item writers create *prompts* and, in the case of multiple-choice items, *response options* that are used to elicit student choices. These format choices reduce the generality of the test results to some degree because they represent only one of many possible ways that knowledge or skills might be demonstrated. For example, one might assess knowledge of grammar by asking students to find an error in a mostly correct passage or asking them to find the one instance of correct usage in an error-filled passage. Both skills—finding errors and recognizing appropriate usage—are relevant to mastery of English prose, but tests tend to ask questions in one format only. Such preferences on the part of the test developer are often unrecognized, but they limit what test scores tell us about student understanding of the domain of interest. Furthermore, in some cases these incidental features of test items have become the focus of test preparation, further eroding the meaning of the test scores (Koretz, McCaffrey, & Hamilton, 2001).

It should be noted that expert item writers are able to write multiple-choice items to measure a wide range of skills, including complex reasoning. However, there are some behaviors that can only be weakly approximated in this format. For example, you cannot fully measure a person's ability to write a persuasive essay, conduct a scientific investigation, or perform a somersault³ using multiple-choice items. Therefore, if a test uses only multiple-choice items, some aspects of the domain may be excluded from the test specifications. Furthermore, novice item writers may find that the format limits the kinds of skills they can measure and the ways they can measure them.

In general, the greater the distance between the specifications and items that constitute the test and the academic content standards that describe the domain—in terms of content, cognitive demands, and format—the less confidence we can have that the test score will reflect understanding of the domain. One of the potential advantages of performance tasks over multiple-choice items is that they offer broader windows onto student understanding. By using more complex prompts and permitting more unconstrained responses, performance assessments can represent many domains in more complete ways.

3. Many states have standards for physical education.

Recent History of Performance Assessment in Large-Scale Testing

In 1990, eight states were using some form of performance assessment in math and/or science, and another six were developing or piloting alternative assessments in math, science, reading, and/or writing. An additional 10 states were exploring the possibility of or developing plans for various forms of performance assessment. In total, 24 states were interested in, developing, or using performance assessment (Aschbacher, 1991). Twenty years later, the use of performance assessment has been scaled back significantly, although it has certainly not disappeared.⁴ No Child Left Behind was a factor in some state decisions. For example, the requirement that all students in grades 3 through 8 receive individual scores in reading and math presented a major obstacle for states like Maryland that were using matrix sampling and reporting scores only at the school level. Concerns about technical quality, costs, and politics contributed to changing assessment practices in other states. In this section we recap some of this history to explore why the enthusiasm of the 1990s was tempered in the 2000s.

Promise of Performance Assessment

The use of performance assessment can be traced back at least two millennia to the Han Dynasty in China, and its history makes fascinating reading (Madaus & O'Dwyer, 1999). For our purposes, it is sufficient to look back two or three decades to the educational reform movement of the 1980s. This period was marked by an increased use of standardized tests for the purposes of accountability with consequences for schools and/or students (Hamilton & Koretz, 2002). Minimum competency testing programs in the states gave way to accountability systems, in which test results determined, variously, student grade-to-grade promotion or graduation, and school rankings, financial rewards, or interventions. In this period educators also began to subscribe to the idea that tests could be used to drive educational reform. The term “measurement-driven instruction” was coined to describe the purposeful use of high-stakes testing to try to change school and classroom behaviors (Popham et al., 1985).

By the end of the decade, educators began to recognize a number of problems associated with high-stakes multiple-choice testing, including degraded instruction (e.g., narrowing of the curriculum to tested topics, excessive class time devoted to test preparation) that led to inflated test scores (Hamilton & Koretz, 2002). They worried about persistent differences in performance between demographic groups, which many incorrectly attributed to the multiple-choice testing format. There were also pointed criticisms from specific content fields. Science educators, for example, complained that multiple-choice

4. There are performance tasks on the New England Common Assessment Program (NECAP) used for accountability in Vermont, New Hampshire, and Rhode Island. Open-ended writing assessments are used in a number of states, and many state end-of-course exams contain performance tasks.

tests emphasized factual knowledge rather than procedural knowledge (Frederiksen, 1984).

Not wanting to give up the power of measurement-driven instruction to shape teacher and student behavior, many educators began to call for a new generation of “tests worth teaching to.” Under the banner of “what you test is what you get” (WYTIWYG) (Resnick & Resnick, 1992), advocates of performance assessment thought they could bring about improvements in curriculum, instruction, and outcomes by incorporating more performance assessments into state testing programs. They also recognized that performance assessments could more easily be designed to tap higher-order skills, including problem solving and critical thinking (Raizen et al., 1989). This enthusiasm led many states to incorporate forms of performance assessment into their large-scale testing programs. We briefly summarize some of the more notable efforts in the following paragraphs.

Vermont Portfolio Assessment Program

Educators in Vermont began to develop the Vermont Portfolio Assessment Program in 1988. They had twin goals: to provide high-quality data about student achievement (sufficient to permit comparisons of schools or districts) and to improve instruction. The centerpiece of the program was portfolios of student work in writing and mathematics collected jointly by students and teachers over the course of the school year. Teachers and students had nearly unconstrained choice in selecting tasks to be included in the portfolios. In writing, students were expected to identify a single best piece and a number of other pieces of specified types. In mathematics, students and teachers reviewed each student’s work and submitted the five to seven best pieces. The portfolios were complemented by on-demand “uniform tests” in writing (a single, standardized prompt) and mathematics (primarily multiple choice). The program was implemented in grades 4 and 8 as a pilot in 1990-91, and statewide in 1991-92 and 1992-93. Early evaluation studies, however, raised concerns with the reliability of the scoring and the overall validity of the portfolio system (Koretz et al., 1994).

Largely due to these problems, the portfolio assessment program was replaced for accountability purposes in the late 1990’s with the New Standards Reference Exam (NSRE) (Rohten et al., 2003), which included some on-demand, performance tasks but not portfolios. Most local districts continue to use the portfolios for their own purposes, but they are not used for state-level reporting. More recently, Vermont joined with New Hampshire and Rhode Island to develop the New England Common Assessment Program (NECAP), which includes multiple-choice and short constructed-response items. In 2009-10, the NECAP reading and math assessments will be administered to all students in grades 3 through 8 and grade 11; the writing assessment to grades 5, 8, and 11; and the science assessment to grades 4, 8, and 11.

In addition, as part of the Vermont Developmental Reading Assessment (VT-DRA)⁵ for second grade, students are asked to read short books and retell the story in their own words. Teachers score the student's oral reading for accuracy and their retelling for comprehension. To ensure reliability, teachers who administer the assessment first have their scoring calibrated through an online process. In addition, the results of the DRA are reviewed annually at the Summer Auditing Institute. Note that the Vermont accountability system does not have high stakes for students; student promotion and high school graduation do not depend on test scores (Rohten et al., 2003).

Kentucky Instructional Results Information System (KIRIS)

In response to a 1989 decision by the Kentucky Supreme Court declaring the state's education system to be unconstitutional, the state legislature passed the Kentucky Education Reform Act of 1990. This law brought about sweeping changes to Kentucky's public school system, including changes to school and district accountability for student performance. The Kentucky Instructional Results Information System (KIRIS) was a performance-based assessment system implemented for the first time in the spring of 1992.⁶ KIRIS tested students in grades 4, 8, and 11 in a three-part assessment that included multiple-choice and short-essay questions, performance "events" requiring students to solve practical and applied problems, and portfolios in writing and mathematics in which students presented the "best" examples of classroom work collected throughout the school year. Students were assessed in seven areas: reading, writing, social science, science, mathematics, arts and humanities, and practical living/vocational studies (U.S. Department of Education, 1995).

KIRIS was designed as a school-level accountability system, and schools received rewards or sanctions based on the aggregate performance of all their students.⁷ School ratings were based on a combination of cognitive and non-cognitive indicators (including drop out rates, retention rates, and attendance rates). A school accountability index combined cognitive and non-cognitive indicators and was reported in biennial cycles. Schools were expected to have all their students at the proficient level, on average, within 20 years, and their annual improvement target was based on a straight-line projection toward this goal. Every two years, schools that exceeded their improvement goals received funds that could be used for salary bonuses, professional development, or as school improvement funds. In 1994-95, about \$26 million was awarded, with awards of about \$2,000 per teacher in eligible schools. The state also devoted resources to support and improve low performing schools, including assigning "distinguished educators" to advise on school operations.

5. Adapted from the Developmental Reading Assessment published by Celebration Press.

6. KIRIS was modified in many small ways during the initial years (e.g., testing was moved from grade 12 to grade 11, mathematics portfolios were moved from fourth to fifth grade); we do not recount all the changes here.

7. Including students with disabilities.

By the late 1990s, two independent panels studied the research evidence on KIRIS and reported serious flaws in the program (Hambleton et al., 1995; Catterall et al., 1997). Perhaps as a result of these criticisms, many parents and educators questioned the validity of the system (Fenster, 1996). The Kentucky legislature voted in 1998 to replace KIRIS with the Commonwealth Accountability Testing System (CATS) (White, 1999), which incorporated some of the components (performance tasks, the writing portfolio) that comprised KIRIS but eliminated the mathematics portfolios. Many factors contributed to this decision, including philosophical disagreements over the “valued outcomes” adopted for education, disputes about the correct way to teach mathematics and literacy, and a switch in the political balance in the legislature (Gong, 2009). Recently Kentucky switched to a criterion-referenced test for No Child Left Behind reporting, the Kentucky Core Content Test (KCCT) for math (grades 3 through 8 and 11), English language arts (grades 3 through 8 and 10), and sciences (grades 4, 7, and 11), which includes some constructed response items.

The KCCT continues to assess student achievement in writing, however, using the Writing Portfolio in grades 4, 7, and 12 and the On-Demand Writing Assessment in grades 5, 8, and 12. A four-piece portfolio is required in grade 12, and a three-piece portfolio is required in grades 4 and 7. The required content includes samples of reflective writing, personal expressive writing/literary writing, transactive writing, and (in 12 grade only) transactive writing with an analytical or technical focus. The On-Demand Writing Assessment provides students in grades 5 and 8 with the choice of two writing tasks that include a narrative writing prompt and a persuasive writing prompt; students in grade 12 are given one common writing task and the choice of one of two additional writing tasks (Kentucky Department of Education, 2009).

As in Vermont, there was initial concern about the reliability of portfolio scoring procedures. As a consequence, tasks were more clearly specified, analytical rubrics were developed, and training was strengthened. By 1996, scoring reliability for the writing portfolio had increased significantly. An independent review of 6,592 portfolios from 100 randomly selected schools found an agreement rate of 77% between independent readers and the ratings given at the school level. By 2008, the agreement rate (exact or adjacent scoring) for independent readers involved in auditing school-level scores was over 90% (Commonwealth of Kentucky, 2009, p. 92).

After using portfolios and writing prompts for 15 years, the Kentucky Department of Education (KDE) published a fact sheet in 2008 called “Considering Myths Surrounding Writing Instruction and Assessment in Kentucky” to address the continued concerns of parents and other groups (KDE, 2008). Among the issues addressed were the perceived “burden” of assembling a portfolio and the possibility of bias and subjectivity in scoring.

Maryland School Performance Assessment System (MSPAP)

The Maryland School Performance Assessment System (MSPAP) was created in the late 1980s and early 1990s to assess progress towards the state's educational reform goals. The MSPAP, first administered in 1991, assessed reading, writing, language usage, mathematics, science, and social science in grades 3, 5, and 8. All of the MSPAP tasks were performance based, ranging from short-answer responses to more complex, multistage responses to data, experiences, or text. As a result, human raters scored all responses. MSPAP tasks were innovative in several ways. Activities frequently integrated skills from several subject areas, some tasks were administered as group activities, some were hands-on tasks involving the use of equipment, and some tasks had pre-assessment activities that were not scored. MSPAP items were matrix sampled, i.e., every student took a portion of the exam in each subject. As a result, there was insufficient representation of content on each test form to permit reporting of student-level scores. MSPAP was designed to measure school performance, and standards-based scores (percentage achieving various levels) were reported at the school and district levels. Schools were rewarded or sanctioned depending on their performance on the MSPAP (Pearson et al., 2002).

Many of the features of MSPAP were unusual for state testing programs, and some stakeholders raised concerns about the quality of MSPAP school results. A technical review committee commissioned by the Abell Foundation in 2000 reported generally positive findings with respect to the psychometric aspects of MSPAP (Hambleton et al., 2000). They also suggested changes to remove some of the more troublesome aspects of the program, like the group-based, pre-assessment activities. The technical review committee also criticized the content of the tests, and some objected to the Maryland Learning Outcomes on which the test was based (Ferrara, 2009). According to the *Washington Post* (Schulte, 2002), MSPAP school-level scores fluctuated widely from year to year, leading the superintendent of one of Maryland's largest districts to demand the delay of the release of the test scores until the fluctuations could be explained. Some prominent supporters of the program turned against it. Partially due to concerns with scoring, and partially due to a desire (and an NCLB requirement) to have individual student scores, the Maryland School Assessment (MSA) replaced MSPAP in 2002 (Hoff, 2002). The MSA tests reading and mathematics in grades 3 through 8 and science in grades 5 and 8 using both multiple-choice and brief constructed response items.

Washington Assessment of Student Learning (WASL)

In 1993, the Washington Legislature passed the Basic Education Reform Act, including the Essential Academic Learning Requirements (EALRs) for Washington students. The EALRs defined learning goals in reading; writing; communication; mathematics; social, physical, and life sciences; civics and history; geography; arts; and health and fitness. The Washington Assessment of Student Learning (WASL) was developed to assess

student mastery of these standards. WASL included a combination of multiple-choice, short-answer, essay, and problem-solving tasks. In addition, the Washington assessment system included classroom-based assessments in subjects not included in WASL.

WASL was implemented in fourth grade in 1996 and in other grades subsequently. Eventually, WASL was administered in reading (grades 3 through 8 and 10), writing (grades 4, 7, and 10), mathematics (grades 3 through 8 and 10), and science (grades 5, 8, and 10). Test results were reported in terms of levels of accomplishment for individuals, and the percentages of students at each level of accomplishment was reported for schools and districts. Initially, listening was assessed as part of the WASL, but this test was discontinued in 2004 as part of a legislative package of changes in anticipation of WASL's use as the high school exit exam starting with the class of 2008.

The use of WASL as the state's high school exit exam was controversial because of the low pass rates of 10th graders, especially in mathematics (Queary, 2004). Other concerns included considerable variation in student performance (percent proficient) in reading and mathematics from year to year, and even greater variation at the strand level (Washington State Institute for Public Policy, 2006). In 2007, the governor delayed the use of the math and science sections, and in 2008 he mandated that scores for the math portion of the WASL not be used.

The WASL will be replaced in 2009-10 with the Measurements of Student Progress (MSP) in grades 3 through 8 and the High School Proficiency Exam (HSPE) in grades 10 through 12. The MSP and HSPE tests include multiple-choice and short-answer questions; the essay questions have been eliminated from the reading, math, and science tests.

Interestingly, Washington uses classroom-based assessments, including performance assessments, to gauge student understanding of the EALR learning standards in social studies, the arts, and health/fitness. Districts must report to the state that they are implementing the assessments/strategies in those content areas, but individual student scores are not reported.

California Learning Assessment System (CLAS)

The California Learning Assessment System (CLAS) was designed in 1991 to align the testing program with the state's curricular content, to measure students' attainment of that content using performance-based assessment, and to provide performance assessments for both students and schools (Kirst & Mazzeo, 1996). First administered in 1993, CLAS assessed students' learning abilities in reading, writing, and mathematics in grades 4, 8, and 10. In reading and writing, CLAS used group activities, essays, and short stories to measure students' critical thinking. In math, students were asked to show how they arrived at their answers. The performance assessment was based not only on the annual exams, but also on portfolios of student work.

Controversy over CLAS arose shortly after the first round of testing, when some school groups and parents claimed that the test items were too subjective, that they encouraged children to think about controversial topics, or asked about the students' feelings, which some parents said was a violation of their student's civil rights (McDonnell, 2004; Kirst & Mazzeo, 1996). In addition, the debate in California highlighted fundamental conflict about the role of assessment in education, with policymakers, testing experts, and the public often voicing very different expectations and standards of judgment (McDonnell, 1994). The California Department of Education did not help matters when it initially declined to release sample items from the exams, citing the cost of developing new items. There were a series of newspaper articles and state level committee reports critical of the test's sampling procedures and of the objectivity of the scoring. In 1994, the legislature reauthorized CLAS in a bill that increased the number of multiple-choice and short answer questions to complement the performance tasks, but this change came too late to save the program; CLAS was administered for the last time later that year.

After a four-year hiatus from statewide achievement testing, the Standardized Testing and Reporting (STAR) exams began in 1998. STAR uses multiple-choice questions to measure the achievement of California content standards in English language arts, mathematics, science, and history-social science (in grades 2 through 11). Initially, the STAR program used the Stanford Achievement Test (Ninth Edition); however, beginning in 2001 the state began to substitute the California Standards Tests, which are largely multiple-choice tests aligned to the California standards, with a writing component at specific grade levels.

Connecticut Academic Performance Test

Connecticut has two assessment programs: the Connecticut Mastery Test (CMT), which assesses reading, writing, and mathematics of students in grades 3 through 8, and the Connecticut Academic Performance Test (CAPT), which assesses reading, writing, mathematics, and science in grade 10 (Connecticut Department of Education, 2009). The CAPT test includes performance-based components in each subject area. For example, the Reading Across the Disciplines section of CAPT has two tests that assess students' reading abilities: Response to Literature and Reading for Information. The Response to Literature test asks students to read a short story and respond to a series of essay questions requiring them to describe, interpret, connect to, and evaluate the story. The Reading for Information test requires students to read several nonfiction articles taken from magazines, newspapers, and journals and answer a combination of multiple-choice and short-answer questions about the meaning of the article and the way the author wrote the article.

The CAPT mathematics section uses questions requiring written responses to assess students' abilities to solve problems, communicate, compute, and estimate in four major content areas (number quantity; measurement and geometry; statistics, probability and discrete mathematics; and algebra and functions).

The Writing Across the Disciplines section contains two tests that assess students' writing abilities: Interdisciplinary Writing and Editing and Revising. In the Interdisciplinary Writing test, students are given a set of source materials (e.g. newspaper and magazine articles, editorials, charts, and graphs) representing different perspectives on an issue. They are asked to read the materials and use the information to write a persuasive piece, such as a letter to a congressperson or a letter to the editor of their local newspaper, which supports their own position on the issue. Students are required to take two interdisciplinary writing tests about separate issues. In the Editing and Revising test, students read passages with embedded errors and answer multiple-choice questions to indicate corrections.

The CAPT science section currently uses a combination of multiple-choice questions and open-ended written responses. Prior to 2007, this section included a performance task that assessed experimentation using a hands-on laboratory activity, but, starting in 2007, the laboratory activity was removed as part of the on-demand testing program and shifted to classroom testing. Beginning in 2007, Connecticut provided schools with detailed materials for curriculum-embedded assessment tasks in five new content strands for grades 9 and 10. The five suggested tasks include a laboratory activity and a Science, Technology and Society (STS) investigation. Schools and teachers are encouraged to use the curriculum-embedded tasks within the classroom during the course of the normal instructional day. The open-ended items on the next generation of the written portion of the CAPT Science Assessment assess inquiry and communication skills in the same context as the five curriculum-embedded tasks (CSDE, 2007).

NAEP Higher-Order Thinking Skills Assessment Pilot

In 1985-86, the National Science Foundation funded the National Assessment of Educational Progress to conduct a pilot test of techniques to study higher-order thinking skills in mathematics and science. Adapting tasks that had been used in the United Kingdom, NAEP developed prototype assessment activities in a variety of formats, including pencil and paper tasks, demonstrations, computer-administered tasks, and hands-on tasks. In all, 30 tasks were developed and piloted with about 1,000 students in grades 3, 7, and 11. The hands-on tasks were designed to assess classifying, observing and making inferences, formulating hypotheses, interpreting data, designing an experiment, and conducting a complete experiment. For example, in *Classifying Vertebrae*, an individual hands-on task, students were asked to sort 11 small animal vertebrae into three groups based on similarities they observed, record their groups on paper, and provide written descriptions of the features of each group. In *Triathlon*, a group pencil-and-paper activity, students were given information about the performances of five children on three events (frisbee toss, weight lift, and 50-yard dash), asked to decide which child would be the all around winner, and write an explanation of their reasoning. According to NAEP, the results were promising; students “responded well to the tasks and in some cases, did quite well” (NAEP, 1987, p.7). Older students did better than younger

students, and across grade levels students did better on tasks involving sorting and classifying than those that required determining relationships and conducting experiments. The researchers also concluded that conducting hands-on assessments was both feasible and worthwhile, although they found it to be “costly, time-consuming, and demanding on the schools and exercise administrators” (Blumberg et al., 1986). Perhaps for these reasons, the “hands-on” items were not used in the 1990 NAEP assessment in science.

Summary

These seven examples were among the more ambitious attempts to use performance assessment on a large scale in the United States during the past two decades, but there were many other states that incorporated performance assessments in some form in their testing programs and many that continue to use performance assessments today (Kahl & Pecheone, 2010). The examples were selected because they were pioneering efforts, because performance assessment played such a prominent role in each system, and because they offer lessons regarding technical quality, impact, and burden associated with performance assessment that continue to be relevant today. This brief history should not be interpreted to mean that performance assessment has no future. The demise of the Vermont Portfolio Assessment system, KIRIS, MSPAP, CLAS, etc., was the result of a confluence of factors unique to each time and setting. While there are lessons to be learned from these histories (as the rest of the paper will discuss), it would be incorrect to infer from these cases that large-scale performance assessment is infeasible or impractical. Many states are using performance assessment successfully today for classroom assessment, end-of-course testing, and on-demand assessment. Instead, the history highlights the kinds of challenges that have to be addressed if performance assessments are to be used successfully on a large scale.

Research Findings

Researchers studied many of the state performance-assessment initiatives to understand how these highly-touted reforms operated in practice. In addition, a number of researchers undertook research and development efforts of their own. These efforts produced a rich literature on the technical quality of the assessments, their impact on practice, and their feasibility for use in large-scale assessment.

Technical Quality

Research on the technical quality of performance assessments provides information about agreement among raters (reliability of the rating process), the reliability of student scores, the fairness of performance assessment for different population groups, and the validity of scores for particular inferences. In reviewing the evidence it is important to remember that the research was conducted in many different contexts—mathematics portfolios, hands-on science investigations, writing tasks, music performances—and the body of evidence may not be complete with respect to technical quality for any specific type of performance assessment.

Agreement Among Raters. When students construct rather than select answers, human judgment must be applied to assign a score to their responses. As performance tasks become more complex, i.e., as process skills become richer and content knowledge more open, it becomes more difficult to develop scoring criteria that fully reflect the quality of student thinking (Baxter & Glaser, 1999). For example, a review of nine studies of direct writing assessment reported rater consistency estimates that ranged from 0.33 to 0.91 (Dunbar, Koretz, & Hoover, 1991). The authors speculated that rater consistency was affected by many factors, including the number of score levels in the rubric and the administrative conditions under which ratings are obtained.

As a broad generalization, in most cases it is possible to train qualified raters to score well-constructed, standardized performance tasks with acceptable levels of consistency using thoughtful rating criteria. Of course, the adjectives “qualified,” “well constructed,” and “thoughtful” are not insignificant obstacles. The keys to achieving consistency among raters on performance tasks seem to be:

1. Selecting raters who have sufficient knowledge of the skills being measured and the rating criteria being applied,
2. Designing tasks with a clear idea of what constitutes poor and good performance,
3. Developing scoring guides that minimize the level of inference raters must make to apply the criteria to the student work,
4. Providing sufficient training for teachers to learn how to apply the criteria to real examples of student work, and

5. Monitoring the scoring process to maintain calibration over time.
When all these elements are in place, it is usually possible to obtain acceptable levels of agreement among raters.

One way to achieve the third goal is to develop “analytic” scoring guides, which tell raters exactly what elements to look for and what score to assign to each type of response. However, success has also been achieved using “holistic” rules, which call for overall judgments against more global standards. Klein et al. (1998) compared analytic and holistic scoring of hands-on science tasks and found that the analytic scoring took longer but led to greater inter-reader consistency; however, when scores were averaged over all the questions in a task the two methods were equally reliable. On the other hand, not all methods are interchangeable. While item-by-item (analytic) scoring and holistic scoring yielded similar scores on mathematics performance assessments, “trait” scoring for conceptual understanding and communication were sensitive to different aspects of student performance (Taylor, 1998).

Non-standardized portfolios present tougher challenges for raters. For example, in Vermont, rater consistency was very low for both the reading portfolios and the mathematics portfolios (piece-level correlations among raters in reading and mathematics averaged about 0.40 during the first two years). Even when aggregated across pieces and dimensions to produce student scores, the correlations between raters were never more than moderate (0.60 in writing in the second year and 0.75 in mathematics in the second year). Initially, Vermont was not able to achieve agreement among raters that was high enough for the scores to be useful for accountability purposes (Koretz et al., 1994). The difficulty in scoring was attributed to a number of factors including the quality of the rubrics, the fact that the portfolios were not standardized (so raters had to apply common rubrics to very different pieces), and the large number of readers who had to be trained.

Over time, rater reliability improved (by 1995 rater reliability for total student scores averaged 0.65 in writing and 0.85 in mathematics across the grades),⁸ suggesting that insufficient rater familiarity and training may have played a large role in unreliability in the early years. In Kentucky, raters assigned a single score for the portfolio as a whole (not a separate score for each piece). In the early years, rater reliability for these overall scores was comparable to reliability for total scores in Vermont, i.e., 0.67 for grade 4 writing portfolios and 0.70 for grade 8 writing portfolios.⁹ While reasonably high, there was concern because, on average, students received higher scores from their own teachers than from independent raters of their portfolios (Hambleton et al, 1995). In later years, scoring reliabilities improved.

The results from the National Assessment of Educational Progress writing portfolio pilot were somewhat better, but rater consistency was still problematic (National Cen-

8. Daniel Koretz, personal communication, 2009

9. Mathematics portfolio scores were not included in the accountability index in the early years.

ter for Education Statistics, 1995). To facilitate comparison of performance, students were asked to include in their portfolios pieces representing particular genres (e.g., persuasive writing, descriptive writing). When readers reviewed the portfolios they first classified each piece as to genre, and then they scored those pieces that fell into genres for which scoring rubrics had been developed. (The remainder of the portfolio was not scored.) Even with this simplification, the level of inter-rater consistency was only moderate (from 0.41 for persuasive writing in grade 4 to 0.68 for informative writing in grade 8).

Reliability of Student Scores. For accountability purposes, the results from multiple performance tasks are combined to produce an overall score for each student. (This is analogous to combining results from many multiple-choice items to produce a total student score.) In an accountability context, the reliability of these student-level scores is of greater importance than the consistency of ratings, although score reliability depends in part on rater consistency. Unfortunately, research suggests that student performance can vary considerably from one performance task to the next, due to unique features of the task and the interaction of those features with student knowledge and experience. This “task sampling variability” means that it takes a moderate to large number of performance tasks to produce a reliable score for a student (Shavelson, Baxter, & Gao, 1993; Dunbar, Koretz, & Hoover, 1991; Linn et al., 1996). In addition, researchers have found that performance on complex tasks can vary by occasion, further complicating interpretation of student performance (Webb, Schlackman, & Sugrue, 2000).

The number of tasks needed to obtain a reliable score for a student is probably a function of the complexity of each task, the similarity among tasks,¹⁰ and the specific task-related knowledge and experiences of the student. As a result, researchers working in different contexts have reported estimates of the minimum number of performance tasks needed for reliable student score that range from two tasks per student to well over 20 tasks per student. For example, the number of writing tasks required to obtain a score reliability of 0.8 ranged from 2 to 10 in six studies reviewed by Koretz, Dunbar, and Hoover (1991). Three class periods of hands-on science tasks were required to produce score reliability of 0.8 (Stecher & Klein, 1997). To produce a student score with reliability of 0.85 as many as 25 pieces would have to be included in the student’s mathematics portfolio in Vermont (Klein et al., 1995). Over 20 mathematics performance tasks that were relatively similar would be needed to produce reliable student scores, and many more would be needed if dissimilar mathematics performance tasks were used (McBee & Barnes, 1998).

Thus, there is no simple answer to the question of how many performance tasks are needed to produce reliable scores. It might be possible to produce more consistent

10. It may be difficult to determine whether two tasks are “similar.” Klein et al., found that scores on two science tasks generated from the same template or task shell correlated no more highly than scores on two tasks generated from different shells (Klein & Stecher, 1991).

results by studying separately tasks with different levels of complexity and different response characteristics. In theory, score reliability could also be improved by developing tests that combined performance tasks with multiple-choice items, assuming the items were assembled to represent a domain in a conceptually sound manner.

Fairness. Some advocates of performance assessment hope that the tasks will reduce score differences between population groups that are commonly reported on multiple-choice tests. They interpret persistent group differences as evidence of inherent bias in multiple-choice tests, although researchers admonish that mean group differences are not *prima facie* evidence of bias, and, over the years, bias reviews have removed items that show differences that are not associated with overall ability. Nevertheless, many hope that performance assessments would reduce traditional group differences. Research does not find that the use of performance assessments changes the relative performance of racial/ethnic groups, however (Linn, Baker, & Dunbar, 1991). For example, differences in scores among racial/ethnic groups on hands-on science tasks were comparable to differences on multiple-choice tests of science in grades 5, 6, and 9 (Klein et al., 1997). Similar results were obtained on NAEP mathematics assessments, which included both short, constructed-response tasks and extended-response items (Peng, Wright, & Hill, 1995).

Validity of Inferences from Performance Assessment Scores. Validity is not a quality of tests, per se, but of the inferences made on the basis of test scores. Researchers gather a variety of types of evidence to assess the validity of inferences from a given measure, including such things as expert judgments about the content of the measure, comparisons among scores from similar and dissimilar measures, patterns of correlations among elements that go into the total score, comparisons with concurrent or future external criteria, and sensitivity to relevant instruction. In fact, Miller and Linn (2000) identify six aspects of validity that are relevant for performance assessments. None of the performance assessments described in this paper has been subject to a thorough validity study encompassing all these elements. In many cases, the assessments were developed for research purposes and were not part of an operational testing program where the intended uses of the scores would be clear, but few operational programs examine validity as thoroughly as they might.

One of the challenges in establishing validity for performance assessments is lack of clarity about precisely what the assessments are intended to measure and what relationships ought to be found with other measures of related concepts. For example, would we expect to find high or low correlations between student scores on an on-demand writing task and scores on their writing portfolios? Both are measures of writing, but they are obtained in different situations. Are they measuring the same or different writing skills?¹¹ In many situations where performance tasks have been used, students also complete multiple-choice items covering related content (this is generally the case with

11. In the 1992 NAEP writing portfolio trial, the correlation between these two writing scores was essentially chance.

NAEP, for example). Yet, there is often little theoretical or empirical justification for predicting how strongly the scores of performance tasks and multiple-choice items of the same overall subject should be related or how strongly they should be related to scores on performance tasks and multiple-choice items measuring a different subject.

For example, in the early days in Vermont, mathematics portfolio scores correlated as highly with score on the uniform test of writing as they did with scores on the uniform test of mathematics, and researchers concluded that portfolio scores were not of sufficient quality to be useful for accountability purposes (Koretz et al., 1994). Similarly, in Kentucky, scores on KIRIS were improving while comparable scores on NAEP and on the American College Testing program were not (Koretz & Barron, 1998); yet, the state standards were generally consistent with the NAEP standards. Kentucky teachers were more likely to report that score gains were the result of familiarity, practice tests, and test preparation than broad gains in knowledge and skills, which would have appeared on other tests of the same content.

Overall, there is insufficient evidence to warrant overall claims about the validity of performance assessments as a class. Recall that one of the primary justifications for using performance assessment is to learn things about student knowledge and skills that cannot be learned from multiple-choice tests. Yet, few would argue that there is no relationship between skills measured by performance assessments and those measured by multiple-choice tests in the same subject. Thus, psychometricians generally look for some relationship between the two measures, but would not expect an extremely high correlation. This ambiguity about predicted relationships makes it difficult to establish a simple concurrent validity argument for a given performance assessment. As a result, performance assessments are often validated primarily on the basis of expert judgment about the extent to which the tasks appears to represent the construct(s) of interest. Even here there are complications (Crocker, 1997). As Baxter and Glaser note, it can be difficult to design performance assessment to measure complex understanding; as a corollary, it can be just as difficult to interpret evidence from complex performance assessments.

Impact

Tests used for standards-based accountability send signals to educators (as well as students and parents) about the specific content, styles of learning, and styles of performing that are valued. An abundance of research suggests that teachers respond accordingly, emphasizing in their lessons the content, styles of learning, and performing that are manifest on the tests.¹² In reviewing this literature, Stecher (2002) concluded that, “large-scale high-stakes testing has been a relatively potent policy in terms of bringing about changes within schools and classrooms.” On the positive side, high-stakes test-

12. Haertel (1999) points out that there are other factors, in addition to testing, that contributed to the emphasis on basic, component skills in the curriculum, including a behaviorist educational philosophy that calls for breaking complex skills into their component parts.

ing is associated with more content-focused instruction and greater effort on the part of teachers and students. Performance assessment, in particular, has been found to lead to greater emphasis on problem solving and communication in mathematics, and to more extended writing in language arts. For example, researchers in Vermont reported that the portfolio assessment program had a powerful positive effect on instruction, leading to changes that were consistent with the goals of the developers. Mathematics teachers reported devoting more time to problem solving and communication in mathematics; similarly they spent more time having students work in pairs or small groups (Stecher & Mitchell, 1995).

Likewise, researchers studying the Kentucky reforms found considerable evidence that teachers were changing their classroom practices to support the reform (e.g., to support problem solving and communicating in mathematics and writing) (Koretz, Barron, Mitchell, & Stecher, 1996). Similarly, researchers in Maryland found that statewide, most mathematics teaching activities were aligned with the state standards and performance assessments (although classroom assessments were less consistent with state assessments) (Parke & Lane, 2008). Teachers in Maryland reported making positive changes in instruction as a result of MSPAP, and schools in which teachers reported the most changes saw the greatest score gains (Lane, Parke, & Stone, 2002). In general, these instructional effects are not a function of the format of the test—they occur both with multiple choice tests and performance assessments—but of attaching consequence to measured student outcomes. However, Kentucky teachers were more likely to report that open-response items and portfolios had an effect on practice than multiple choice items or performance events, adding credence to the impact of “tests worth teaching to.”

On the negative side, Stecher (2002) concluded, “Many of these changes appear to diminish students’ exposure to curriculum...” This conclusion was drawn primarily from research in which teachers reported that they changed instruction in ways that “narrowed” the curriculum to those topics covered by the tests (Shepard & Dougherty, 1991). In addition, researchers documented substantial shifts in instructional time from non-tested to tested subjects and, within subjects, from non-tested to tested topics. For example, teachers increased coverage of basic math skills, paper-and pencil computations and topics included in the tests and decreased coverage of extended project work, work with calculators, and topics not included in the test (Romberg, Zarinia, & Williams, 1989).

More recently, researchers have documented the phenomenon of “educational triage,” where teachers focus resources on students near the cut-off point for proficient at the expense of other students (Booher-Jennings, 2005). Although these studies were conducted in the context of multiple-choice testing, it seems fair to predict that similar effects would be observed with high-stakes performance assessments if they focused on some parts of the curriculum or some students more than others. In fact, the curriculum-narrowing problem might be exacerbated with the use of performance assessments

because each task is more “memorable” than a corresponding multiple-choice item, increasing the likelihood that teachers might focus on task-specific features rather than broader skills.

Finally, the research documents instructional changes that can be associated with the format of the high-stakes test. For example, teachers engaged in excessive test preparation, in which students practiced taking multiple-choice tests between one and four weeks per year and up to 100 hours per class (Herman & Golan, nd; Smith, 1994). Others noted instances of coaching that focused on incidental aspects of the test (e.g., the orientation of the polygons) that were irrelevant to the skills that were supposed to be measured (Koretz, McCaffrey, & Hamilton, 2001). Other researchers found that teachers had students engage in activities that mimicked the format of the tests. For example, teachers had students find mistakes in written work rather than producing writing of their own (Shepard & Dougherty, 1991).

Using performance assessments rather than multiple-choice tests might reduce the prevalence of these effects because the tasks are more representative of the reasoning embodied in the standards. But performance assessment is not immune from negative effects when used in a high-stakes context. For example, teachers in Vermont were found to engage in “rubric driven instruction,” in which they emphasized the aspects of problem solving that led to higher scores on the state rubric rather than problem-solving in a larger sense (Stecher & Mitchell, 1995).

Burden

Large-scale testing takes student and teacher time away from teaching and learning; it imposes additional administrative burdens on schools (storage, preparation, proctoring, shipping, etc.); and it commands financial resources (for test booklets, scoring services, and reporting). Policymakers—and most educators—accept these practical constraints as a necessary part of having a standardized testing program. Incorporating performance assessments into standards-based accountability will probably make the testing system more burdensome than it currently is. As a general rule, performance assessments require more classroom time, place greater administrative burdens on staff, and are more expensive than multiple-choice tests for a similar amount of testing time, or for scores with similar levels of reliability.

However, to compare the burdens and costs of these two modes of assessment, we must hold something constant. For example, we might focus on a particular set of skills and compare a performance assessment measuring those skills to a multiple-choice test measuring the same skills, or we might hold testing time constant and compare a performance assessment and a multiple-choice test lasting the same length of time, or we might focus on score reliability and compare a performance assessment and a multiple-choice test of similar reliability, etc. Each approach is found in the literature, but seldom are they all used in any single study. Thus, we have to amalgamate bits of cost informa-

tion from different sources. Furthermore, many of these estimates are quite old, and it is reasonable to assume that test development costs have declined as states and contractors learn from the past efforts. The following paragraphs explore the costs of developing, administering, and scoring performance assessments relative to multiple-choice tests as reported in the literature.

In general, it is more difficult and more costly to develop high-quality, open-response tasks than high-quality, multiple-choice items. Test developers have more freedom to craft interesting stimulus materials and prompt for thoughtful student responses. However, this freedom can easily undermine the goal of producing a task that is understood in the same way by all test takers. Test developers are less able to anticipate the way students will approach a task, and more extensive pilot testing and revision is often necessary (Hamilton, 1994). In fact, in the early 1990s, test developers who did not devote sufficient time and energy to the process produced a number of weak performance tasks (Baxter & Glazer, 1988).

As tasks become more complex, (e.g., involving equipment and materials), more iterations may be necessary to write instructions that are clear to test takers. For example, in one research study, the cost to develop hands-on science tasks lasting 30 to 40 minutes ranged from \$84,000 to \$135,000 (including enough materials for 1,000 students) (Stecher & Klein, 1997). Assuming those tasks were eventually used with 100,000 students, the per-student cost was estimated to be \$22 to \$45. This is the same order of magnitude as the cost of the State Collaborative on Assessment and Student Standards (SCASS) written science performance assessments, which ranged from \$11 to \$14 per student per class period (Doolittle, 1995).

Another study of science assessments estimated multiple-choice tests to be the least expensive in terms of development, scorer training, and scoring, and the researchers found that it cost “80 times as much for an open ended item, 300 times as much for a content station, and 500 times as much for a full investigation item” (Lawrenz, Huffman, & Welch, 2000, page 623). Commercial test development is also costly; the cost to develop the Iowa Writing Assessment, which does not involve equipment, was estimated to be \$370,000 prior to publication or marketing (Hoover & Bray, 1995).

In the early 1990s, a federal agency estimated that performance assessment was 3 to 10 times as expensive as multiple choice testing (Office of Technology Assessment, 1992). According to a recent analysis using current cost data, performance assessments can be developed and implemented as part of a larger assessment system at costs closely comparable to those of traditional tests through strategic uses of technology, teacher scoring, and economies of scale achieved by states working in consortia together (Topol, Olson & Roeber, 2010).

Nevertheless, there are now commercially available achievement test batteries, such as the TerraNova CTBS (from CTB McGraw-Hill), which include constructed-response

items, at a cost that is competitive with multiple choice testing. Similarly, the New England Common Assessment Program (NECAP), developed by and used in Vermont, New Hampshire, and Rhode Island, includes a variety of performance assessments. By pooling resources these small states could afford to develop the test for their annual state assessments.

It should be noted that researchers are experimenting with tools to make the task development process more systematic and routine, which might reduce costs substantially. For example, researchers have experimented with “shells” that could be used to generate multiple versions of a task (Solano-Flores et al., 2001) and model-based approaches (Baker, 1997) to formalize what was often a hit-and-miss endeavor. Others are exploring alternative formats for assessing complex understanding, including approaches such as knowledge mapping (Herl et al., 1999) and dynamic evaluation of enhanced problem-solving (DEEP) (Spector, 2006). These approaches, and others that have been developed by researchers and test publishers, have the potential to reduce the development and scoring costs and expand the skills that can be assessed effectively (Lane, 2010).

As performance tasks become more complex, involving extensive source materials, equipment, and apparatus, disposable supplies, etc., the cost of shipping the materials, storing them at the school site, arranging facilities for administration (flat tables rather than slanted desks), clean up, packaging, and returning the materials increases. Other than shipping, these responsibilities may not involve actual expenditures by schools, but they do represent opportunity costs, as staff time must be reallocated from other duties. We have not found a good estimate of the relative costs of administration for performance and multiple-choice assessments, but we know from experience the complexity of administering tasks that involve equipment and disposable materials.

As a general rule, performance assessments are more expensive to score than multiple-choice tests. However, there is considerable variation in scoring costs based on the nature of the performance and the nature of the score(s) to be assigned. For example, the cost of scoring student written essays was estimated to be between \$1.47 and \$5.88 per student (Hardy, 1995). Similarly, the cost for scoring the Iowa writing assessment was estimated to be about \$5 per student (Hoover et al.). On the other hand, scoring costs for the hands-on science tasks mentioned above were two or three times as much, ranging from \$9 to \$15 per student (Stecher & Klein, 1997).

Scoring costs for performance assessments may be reduced in the future through the use of computerized scoring procedures. In fact, computers are currently used for scoring some large-scale student writing assessments. In most cases, one human reader is still used with the computer replacing the second reader. However, computerized scoring is accurate enough that it may soon replace human scoring in cases where the unit of analysis is the school (Klein, 2008).

It is also important to consider score reliability when thinking about the relative burden of performance assessment. Because performance assessments require more time to complete than multiple-choice items, and because task sampling variability is greater with performance assessments, it takes more time and more performance tasks to yield a reliable student score. Hands-on experimental science may represent the extreme case because of the complexity of the tasks and the time required to complete them, but the cost of producing a reliable science score for an individual student was estimated to be 60 times greater using hands-on tasks than using multiple-choice items (Stecher & Klein, 1997). The cost was due, in part, to the need for three class periods of hands-on testing. If one were satisfied with a classroom-level score or a school-level score, the testing time could be reduced to a single class period, cutting the cost by two-thirds, but this would still be 20 times the cost of multiple-choice test.

Ultimately, policymakers would like to know whether the benefits of performance assessment (in terms of more valid measurement of student performance, positive impact on classroom practice, etc.) justify the burdens (in terms of development costs, classroom time, scoring costs, etc.). This review suggests that the expenditures and administrative burdens associated with performance assessments, particularly portfolios and extended hands-on science tasks, were high relative to multiple-choice tests. Yet, that is not the end of the story.

First, the benefits may justify the burdens from the perspective of education. Vermont teachers and principals thought their state's portfolio assessment program was a "worthwhile burden." In fact, in the first years, many schools expanded their use of portfolios to include other subjects (Koretz et al., 1994), and even in recent years, most Vermont districts have continued the use of the writing and mathematics portfolios, even though they are not used for state accountability purposes. Similarly, Kentucky principals reported that although they found KIRIS to be burdensome, the benefits outweighed the burdens (Koretz, Barron, Mitchell, & Stecher, 1996). Second, the costs associated with performance assessments have probably declined over the past decade, making it more attractive to incorporate some degree of performance assessment into state testing programs. In spite of its complicated history, the future for performance assessment looks promising.

Current examples of large-scale performance assessments

Despite the technical and practical challenges that confront large-scale use of performance assessment, there are testing programs in operation that rely on performance assessments.

Collegiate Learning Assessment (CLA). The Council for Aid to Education created the Collegiate Learning Assessment (CLA) in 2000 to help postsecondary faculty improve teaching and learning in higher education institutions (Benjamin et al., 2009).

The intent of the test is to provide an assessment of the value added by the school's instructional and other programs with respect to desired learning outcomes (Klein et al., 2007).

CLA is entirely performance based and uses two types of tasks, Performance Tasks and Analytic Writing Tasks. The Performance Tasks present students with problems that simulate real world issues and give an assortment of relevant documents, including letters, memos, summaries of research reports, newspaper articles, maps, photographs, diagrams, tables, charts, interview notes, or transcripts. Students have 90 minutes to review the materials and prepare their answers. Performance Tasks often require students to marshal evidence from different sources; distinguish rational from emotional arguments and fact from opinion; understand data in tables and figures; deal with inadequate, ambiguous, and/or conflicting information; spot deception and holes in the arguments made by others; recognize information that is and is not relevant to the task at hand; identify additional information that would help to resolve issues; and weigh, organize, and synthesize information from several sources. Students' written responses to the problems are evaluated to assess their abilities to think critically, reason analytically, solve problems, and communicate clearly and cogently.

The Analytic Writing Tasks ask students to write answers to two types of essay prompts: a "Make-an-Argument" question that asks them to support or reject a position on some issue; and a "Critique-an-Argument" question that asks them to evaluate the validity of an argument made by someone else.

The CLA is computer administered over the Internet, with all supporting documents contained in an online document library. The online delivery permits the performance assessments to be administered, scored, analyzed, and reported to the students and their institutions more quickly and inexpensively. Initially, trained readers scored the tasks using standardized scoring rubrics. Starting in fall 2008, a combination of machine and human scoring was used. Scores are aggregated by institution and are not reported at the individual student level (Collegiate Learning Assessment, 2009). A recent study of the machine scoring has suggested that the correlations between hand and machine scoring are so high that, when the institution is the unit of analysis, machine scores alone can be relied on (Klein, 2008). Even at the student level, the correlation between machine scoring and human scorers is 0.86 (Klein et al., 2007).

Program for International Student Assessment (PISA). The Programme for International Student Assessment (PISA) is a triennial survey of the reading, mathematics, and science literacy of 15-year-olds across the globe (Organisation for Economic Cooperation and Development, 2009). The test is the product of collaboration between participating countries and economies under the auspices of the Organisation for Economic Cooperation and Development (OECD). In 2009, there will be 67 participating countries, including the United States. The assessment focuses on young people's ability to use their knowledge and skills to meet real-life challenges, not

just the extent to which they have mastered a specific school curriculum. The results are presented at the country level.

Tests are typically administered to between 4,500 and 40,000 students in each participating country. The PISA questions are grouped into units that consist of stimulus material such as texts, tables, and/or graphs, followed by questions on various aspects of the material. The questions use different formats: Some are multiple choice, most require a short answer, and some a longer constructed response. The reading unit consists of material, which could include a graph and/or text, a short story, an excerpt from a play, etc. Students read and then respond to a set of multiple-choice and short answer questions. The mathematics questions are predominately open-ended and ask the student to show their work. The science unit provides the student with graphs, data tables, and/or text describing a problem, procedure, or observation, and asks the student to respond to a series of multiple-choice and open-ended questions.

National Assessment of Educational Progress (NAEP). The National Assessment of Educational Progress (NAEP) conducts periodic assessments in mathematics, reading, science, writing, the arts, civics, economics, geography, and U.S. history. NAEP results are based on representative samples of students at grades 4, 8, and 12 for the main assessments, or samples of students at ages 9, 13, or 17 for the long-term trend assessments (National Assessment of Educational Progress, 2009a). In all of the subject areas but writing, the NAEP items are a combination of multiple-choice and constructed-response items, which require short or extended written responses. In the three subject areas discussed below, NAEP also uses other forms of performance-based tasks.

Science Assessment Hands-On Experiments. In the 2009 Science Assessment, administered to students in grades 4, 8 and 12, a sample of students performed hands-on experiments, manipulating selected physical objects to solve a scientific problem. In addition, one-half of the students in each participating school received one of three hands-on tasks and related questions. These performance tasks required students to conduct actual experiments using materials provided to them, and to record their observations and conclusions in their test booklets by responding to both multiple-choice and constructed-response questions. For example, students at grade 12 might be given a bag containing three different metals, sand, and salt and be asked to separate them using a magnet, sieve, filter paper, funnel, spoon, and water and to document the steps they used to do so.

Arts Assessment Creative Tasks. NAEP has used performance-based assessments of the arts since 1972 (music) and 1975 (visual arts). Both music and visual arts were assessed in 1997 and most recently in 2008; the next assessment is planned for 2016. In 1997, NAEP assessed students in four arts disciplines: dance, music, theater, and visual arts; the 2008 assessment included music and visual arts only because of budget constraints and the small percentage of schools with dance and theater programs.

The 2008 arts assessment of a sample of eighth grade students used a combination of “responding” tasks (written tasks, multiple-choice items) and “creative” performance-based tasks. The music portion of the assessment was composed of responding questions only, such as listening to pieces of music and then analyzing, interpreting, critiquing, and placing the pieces in historical context. The visual arts assessment also included creative response tasks in which, for example, students were asked to create a self-portrait that was scored for identifying detail, compositional elements, and use of materials (Keiper et al., 2009). Responding questions asked students to analyze and describe works of art and design; For example, the test asked students to describe specific differences in how certain parts of an artist’s self-portrait were drawn.

Writing Assessment. In 2007, NAEP assessed writing in a nationwide sample of eighth and twelfth graders (Salahu-Din, Persky, & Miller, 2008). Students were provided with narrative, informative, and persuasive writing prompts.

Currently, NAEP scans all open-ended responses and the scanned responses are sent to appropriately trained human readers for scoring. In 2005, NAEP examined whether the mathematics and writing exams, including the written constructed responses, could be automatically or machine scored (Sandene et al., 2005). In the Mathematics exam, eight of the nine constructed-response items included in the computer test in grades 4 and 8 were scored automatically. For both grades, the automated scores for the items requiring simple numeric entry or short text responses generally agreed as highly with the scores assigned by two human raters as the raters agreed with each other.

Questions requiring more extended text entry had less agreement between the automatic scores and the scores assigned by two human raters. The Writing Assessment presented two essay questions to eighth graders. The results showed that the automated scoring of essay responses did not agree with the scores awarded by human readers; the automated scoring produced mean scores that were significantly higher than the mean scores awarded by human readers. Furthermore, the automated scores agreed less frequently with the readers in level than the readers agreed with each other, and the automated scores agreed less with the readers in rank order than the readers agreed with one another.

The 2011 draft NAEP Writing Framework calls for assessment of “computer-based writing” using word processing software with commonly available tools in grades 8 and 12 (National Assessment of Educational Progress, 2009b). However, NAEP still plans to score the 2011 writing assessment using human readers.

The previous examples all have low-stakes for students, but performance assessments are also being used in cases where student performance has important consequences. The National Board for Professional Teaching Standards uses a variety of assessment center exercises as well as an annotated portfolio to certify teachers with advanced teaching skills (which often qualifies them for salary bonuses) (National Research

Council, 2008). California requires teacher preparation programs to use a teaching performance assessment like the Performance Assessment for California Teachers (PACT) as one component in their credentialing decision; PACT has been found to be a valid measure for this purpose (Pecheone & Chung, 2006).

National Occupational Competency Testing Institute (NOCTI). NOCTI is a non-profit organization founded in the early 1970s to coordinate and lead the efforts of the states in developing competency tests for occupational programs (National Occupational Competency Testing Institute, 2009). Today, NOCTI provides “Job Ready” assessments to measure the skills of an entry-level worker or a student in secondary or post-secondary career and technical programs. Most Job Ready assessments include both multiple-choice and performance components. For example the Culinary Arts Cook II Assessment includes as a performance assessment preparing a chicken with sauce recipe; the task is scored on organization, knife skills, use of tools and equipment, preparation of chicken and sauce, safety and sanitation procedures, appearance and taste of finished product. The Business Information Processing Assessment includes a performance task requiring the student to create a spreadsheet; the task includes header and placement, spreadsheet and column headings, data entry, formula entry, computation of totals, use of functions, formatting, creating a pie chart, saving spreadsheet, printing material, and overall timeliness of job completion. More than 70 Job Ready Assessments are available in 16 industry and occupational categories. The performance assessments may be scheduled over one to three days (NOCTI, 2009) and, depending on the subject area and the test site, range in cost from less than \$100 to as much as \$700 per student.

Performance Assessment in the Context of Standards-Based Accountability

This review suggests that large-scale testing for accountability in the United States could be enhanced by the thoughtful incorporation of standardized performance assessment. The enhancements would come from better representation of academic content standards, particularly those describing higher-order, cognitively demanding performance; from clearer signals to teachers about the kinds of student performances that are valued; and from reduced pressures to mimic the multiple-choice frame-of-mind in classroom instruction.

The appropriate role for performance assessments should be determined, in part, by an analysis of content standards. Such an analysis should reveal which standards are served well by which types of assessments. To the extent that the standards call for mastery of higher-order, strategic skills, they may favor the use of performance assessment. Perhaps more importantly, to the extent that standards are silent about the nature of performance expected of students, they abrogate responsibility to others for these decisions. Thus, it may be important to revisit standards documents to make sure they provide adequate guidance with respect to desired student performance.

The research reminds us that subject domains are different, and mastery of each domain is manifest in unique ways. Rich, thoughtful, integrated writing can be observed under different circumstances and in different ways than rich, thoughtful, integrated scientific inquiry or rich, thoughtful, integrated musical performance. When the Mikado sings, “Let the punishment fit the crime” (Gilbert & Sullivan, 1885), the educator should reply “Let the assessment fit the domain.”

There are costs and benefits associated with testing for accountability in whatever form that testing takes. We are used to the current high-stakes, multiple-choice model, but that does not mean it is cost free or benefit rich. Adopting performance assessments for some or all accountability testing will have trade-offs, and we are more likely to make wise decisions if we understand these trade-offs better. Unfortunately, performance assessments themselves differ; the costs and benefits associated with short answer, fill-in-the-blank items are not the same as those associated with prompted writing tasks, equipment-rich investigations, or judged real-time performances.

In general, the addition of performance tasks would increase the overall cost of assessment, and the more complex the tasks the greater the additional costs. The size of the differential is uncertain; we suspect it has fallen over the past decade as development techniques have improved, but we do not believe it will ever approach zero because the process is inherently more complicated than the process of developing multiple-choice items. Costs could be controlled by deciding to use the scores from performance tasks as indicators of achievement at the school-level rather than the individual-level. This

approach is consistent with current NCLB and state accountability systems in which the primary unit of accountability is the school.¹³

If this situation changes—for example, if incentives are assigned to individual teachers (as they are in “pay for performance” schemes) it would still be possible to use a matrix sampling approach within classrooms to offset some of the cost increases associated with performance assessment. Student level accountability policies (such as promotional gates testing) would probably require census use of performance tasks; this would be more costly overall, although not more complicated logistically. Further analysis will be necessary to estimate the magnitude of the additional costs under different performance assessment scenarios. If states adopt common core standards, as many are now contemplating for Algebra I, this could reduce costs by permitting the wider use of tasks.

Improvements in artificial intelligence, which would allow computers to take on larger roles in scoring open-ended responses, could reduce costs in the future. Zelinsky and Sireci (2002, page 337) believe “there appears to be vast potential for expanding the use of more computerized constructed-response type items in a variety of testing contexts.” If these advances come to fruition, they could reduce burden as well as costs.

Messick (1994) distinguishes between task-centered performance assessment, which begins with a specific activity that may be valued in its own right (e.g., an artistic performance) or from which one can score particular knowledge or skills, and construct-centered performance assessment, which begins with a particular construct or competency to be measured and creates a task in which it can be revealed. Research suggests it would be more productive to concentrate performance assessment for accountability on construct-oriented tasks derived from academic content standards, and leave for classroom use more task-defined activities that may be engaging and stimulate student learning but do not represent a clear, state-adopted learning expectation. In addition, it would be wise to eschew the use of unstructured portfolios in large-scale assessment both because they are difficult to score reliably and because it is difficult to interpret the scores once obtained.

It would also be wise to remember that Campbell’s Law applies to performance assessments as well as multiple-choice tests (Campbell, 1979). When you attach stakes to the scores, teachers will feel pressure to focus narrowly on improving performance on specific tasks, which will undermine the interpretability of scores from those tasks. While performance assessments may be more “worth teaching to” than multiple choice tests, performance tasks still represent just a sample of behaviors from a larger domain. Strategies that are used to reduce corruption with multiple-choice tests, including changing test forms regularly and varying the format of tasks and the representation of topics, will be equally useful to reduce corruption with performance assessments.

13. However, this approach would not satisfy the current NCLB requirement for reporting individual scores for students.

The recent history of performance assessment at the state level raises some concerns about using these tasks in high stakes contexts. States at the forefront of the performance assessment movement often found that it was difficult to garner and sustain public support for these “new” forms of testing. While some of the problems states encountered were due to difficulties with scoring, reliability, and validity, others came from energized stakeholder groups who objected to aspects of the assessments or the manner in which they were implemented. In some states, people objected because the assessments were unfamiliar and stretched the boundaries of traditional testing. In others, the assessments were implemented in ways that did not adequately answer parents’ questions and did not always respect parents’ opinions.

McDonnell (2009) characterized these problems as disputes about “the cultural and curricular values embodied in the standards and assessments” (p. 422). Conflicts over values are not easily resolved, but better communication and dissemination of information might help to forestall them. Educators and policymakers may underestimate the need for efforts to inform the public. For example, despite the endless discussion of No Child Left Behind in the education community since 2001, a majority of the general public reports that it is not very familiar with the law (Bushaw & Gallup, 2008). History suggests that educators would be wise to clearly delimit the role of performance assessments and make extra efforts to educate parents and the general public about changes in the testing program before they are adopted.

The successful use of structured performance assessments on a large scale in low-stakes contexts such as PISA and NAEP suggests that practical and logistical problems can be overcome, and that performance tasks can enhance our understanding of student learning.

In their 1999 review, Madaus and O’Dwyer concluded that “the prognosis for the feasibility of deploying a predominantly performance-assessment oriented system for high-stakes decisions about large numbers of individual students is not very promising in light of the historical, practical, technical, ideological, instructional and financial issues described above” (p. 694). While it is still the case today that a standards-based accountability system relying primarily on performance assessment seems unlikely, we can be more optimistic about the possibility of using performance assessment as part of a large-scale testing program in combination with multiple-choice items.

Finally, this review suggests the need for additional research and development. It is unfortunate that relatively little R&D has occurred in the past decade. Conceivably, a more active program of research might have fulfilled the need that Baxter and Glaser (1994) identified for a “foundation for assessment design that is grounded in a theory of cognition and learning and methodological strategies for establishing the validity of score use and interpretation in terms of the quality and nature of the observed performance” (p. 44). Such work may need to be done in each of the major disciplines to be sensitive to discipline-specific ways of knowing, thinking, and acting.

Recommendations

We offer the following recommendations to educators and policymakers considering the future role of performance assessment in large-scale testing in the United States.

1. Set reasonable expectations. Performance assessment is not a panacea for the ills of American education, but it can improve our understanding of what students know and can do, and can help educators focus their effort to bolster critical skills among American youth. Performance assessments are more likely to be successful operating in tandem with multiple-choice items than replacing them completely.
2. Let the standards inform the assessments. The use of performance assessments should be linked clearly to state academic contents standards, so they have a strong warrant for inclusion and a clear reference for inferences. As Linn (2000, p. 8) observed, "...content standards can, and should, if they are to be more than window dressing, influence both the choice of constructs to be measured and the ways in which they are eventually measured." This is not an unreasonable demand, and there is evidence that researchers and practitioners can work together toward this goal. For example, Niemi et al. (2007) describe a seven-year collaboration with a large school district to develop and implement performance assessment connected to explicit learning goals and standards.
3. Revise standards so they better support decisions about assessments, and revise test specifications accordingly. Rewrite state standards to include descriptions of how knowledge and skills would be manifest in student performance. In its review of assessment for K-12 science, the National Research Council (2006) recommended that effective standards should "describe performance expectations and identify proficiency levels." After such modifications are made in state standards, states should revise their test specifications to clearly delineate the role of performance assessment.
4. Clearly delimit the role of performance assessments in ways that help the public understand their relevance and value in making judgments about student performance. Provide adequate information (including sample items) to educate parents about the nature of performance tasks, their role in testing, and the way the results should be interpreted.
5. Invest in the development of a new generation of performance tasks. Previous efforts demonstrated the creativity of researchers and test developers but they were not well integrated with standards-based systems. One goal of such efforts should be to develop multiple approaches to measuring particular skills. Develop more than one format for measuring each construct to avoid focused test preparation on incidental aspects of task

format. This work might be facilitated by encouraging states to pool efforts to develop performance tasks. Joint development efforts will reduce unit costs, broaden the applicability of the tasks, and provide information across a larger universe of students. A natural place to begin would be subjects where common standards are under development, such as Algebra I.

6. Provide instructional support materials for teachers. When performance assessment are included in statewide testing, it is important to develop and make available support materials for teachers, including descriptions of skills assessed, sample lessons for teaching those skills, and sample tasks to use locally to judge student performance. As NAEP found, “Teachers need the political, financial and administrative support that will allow them to concentrate on developing ideas and building up the process skills necessary for students to learn to solve problems and accomplish complex tasks” (National Assessment of Educational Progress, 1987, p. 7).

7. Support research and development to advance the science of performance assessment. This should include efforts to develop performance assessment models to facilitate new task development and research into automated delivery and scoring to reduce costs. A relatively simple but important task is to develop clearer terminology. Having a clearer vocabulary to differentiate among performance tasks with respect to format, cognitive demand, etc. will facilitate thoughtful discussion and policymaking and avoid misapplication of lessons from the past.

References

- Aschbacher, P. (1991). Performance assessment: State activity, interest and concerns. *Applied Measurement in Education*, 4 (1), 275–288.
- Baker, E. L. (1997). Model-based performance assessments. *Theory Into Practice*, 36 (4), 247-254.
- Baxter, G., & Glaser, N. (1998, Fall). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, 37-45.
- Benjamin, R., Chun, M., Hardison, C, Hong, E., Jackson, C., Kugelmass, H., Nemeth, A., & Shavelson, R. (Monograph, 2009). *Returning to learning in an age of assessment: Introducing the rationale of the collegiate learning assessment*.
- Blumberg, F, Epstein, M., MacDonald, W., & Mullis, I. (1986, November). *A pilot study of higher-order thinking skills assessment techniques in science and mathematics*. Final Report – Part I. Princeton, NJ: National Assessment of Educational Progress.
- Booher-Jennings, J. (2005). Below the bubble: “Educational triage” and the Texas accountability system. *American Educational Research Journal*, 42 (2), 231-268.
- Bushaw, W. J., & Gallup, A. M. (2008, September). Americans speak out—Are educators and policy makers listening? The 40th annual Phi Delta Kappa/Gallup Poll of the public’s attitudes toward the public schools. *Phi Delta Kappan*, 90 (10), 8-20.
- Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2, 67-90.
- Catterall, J., Mehrens, W., Flores, R. G., & Rubin, P. (1998, January). *The Kentucky instructional results information system: A technical review*. Frankfort, KY: KY Legislative Research Commission.
- Collegiate Learning Assessment. (2009). Retrieved September 9, 2009, from <http://www.collegiatelearningassessment.org/>
- Connecticut State Department of Education (2009). *Student Assessment*. Retrieved September 7, 2009 from, <http://www.csde.state.ct.us/public/cedar/assessment/index.htm>
- Connecticut State Department of Education (2007). *CAPT high school science assessment handbook – 3rd generation*. Retrieved September 7, 2009, from <http://www.sde.ct.gov/sde/cwp/view.asp?a=2618&q=320890>
- Council of Chief State School Officers (2009). *Statewide student assessment 2007-08 SY: Math, ELA, science*. Retrieved September 4, 2009, from http://www.ccsso.org/content/pdfs/2007-08_Math-ELAR-Sci_Assessments.pdf
- Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education*, 10 (1), 83-95.
- Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

- Doolittle, A. (1995). The cost of performance assessment in science: the SCASS perspective. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4 (4), 289-303.
- Educational Testing Service. (May 1987). *Learning by doing: A manual for teaching and assessing higher-order thinking in science and mathematics*. Report No: 17-HOS-80. The Nation's Report Card. Princeton: Author.
- Fenster, M. (1996, April 8-12) *An assessment of "middle" stakes educational accountability: The case of Kentucky*. Paper presented at the Annual Meeting of the Educational Research Association, New York, NY.
- Ferrara, S. (2009, December 10-11). *The Maryland school performance assessment program (MSPAP) 1991-2002: Political considerations*. Presentation at the National Research Council workshop "Best practices in state assessment." Retrieved December 15, 2009, from http://www7.nationalacademies.org/bota/Workshop_1_Presentations.html
- Frederiksen, N. (1984). The real test bias. *American Psychologist*, 39 (3), 193-202.
- Gearhart, M., Herman, J. L., Baker, E. L., & Whittaker, A. (1993). *Whose work is it? A question for the validity of large-scale portfolio assessment*. CSE Tech. Report No. 363. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Gong, B. (2009, December 10-11). *Innovative assessment in Kentucky's KIRIS System: Political considerations*. Presentation at the National Research Council workshop "Best practices in state assessment." Retrieved December 15, 2009, from http://www7.nationalacademies.org/bota/Workshop_1_Presentations.html
- Hambleton, R. K., Jaeger, R. M., Koretz, D. Linn, R. L., Millman, J., & Phillips, S. E. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System, 1991-1994*. Frankfort, KY: Office of Educational Accountability, Kentucky General Assembly.
- Hambleton, R. K., Impara, J., Mehrens, W., & Plake, B. S. (2000). *Psychometric review of the Maryland School Performance Assessment Program (MSPAP)*. Psychometric Review Committee.
- Hamilton, L. S. (1994). *An investigation of students' affective responses to alternative assessment formats*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Hamilton, L. S., & Koretz, D. M. (2002). Tests and their use in test-based accountability systems. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.) (2002). *Making sense of test-based accountability in education*. MR-1554-EDU. Santa Monica: RAND.
- Hardy, R. A. (1995). Examining the cost of performance assessment. *Applied Measurement in Education*, 8 (2), 121-134.
- Haertel, E. H. (1999). Performance assessment and education reform. *Phi Delta Kappan*, 662-666.

- Herl, H. E., O'Neil, H. F., Jr, Chung, G. K. W. K., et al. (1999). *Final report for validation of problem solving measures*. CSE Technical Report 5. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Herman, J. L., & Golan, S. (n.d.). *Effects of standardized testing on teachers and learning—another look*. CSE Technical Report 334. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Hersh, R. (2009). *Teaching to a test worth teaching to: In college and high school*. Retrieved December 18, 2009, from http://www.cae.org/content/pro_collegework.htm
- Hoff, D. (2002, April 3). Md. to phase out innovative program. *Education Week*. Retrieved September 10, 2009, from <http://www.edweek.org/ew/articles/2002/04/03/29mspap.h21.html>
- Hoover, H. D., & Bray G. B. (1995). *The research and development phase: Can a performance assessment be cost-effective?* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Keiper, S., Sandene, B. A., Persky, H. R., & Kuang, M. (2009). *The Nation's Report Card: Arts 2008 Music & Visual Arts* (NCES 2009–488). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.
- Kentucky Department of Education (2008). *Fact sheet: Reconsidering myths surrounding writing instruction and assessment in Kentucky*. Retrieved September 11, 2009, from <http://www.education.ky.gov/kde/instructional+resources/literacy/kentucky+writing+program/fact+sheet+-+reconsidering+myths+surrounding+writing+instruction+and+assessment+in+kentucky.htm>
- Kentucky Department of Education. *On-demand writing released prompts in grades 5, 8, and 12*. Retrieved September 7, 2009, from <http://www.education.ky.gov/kde/administrative+resources/testing+and+reporting+/district+support/link+to+released+items/on-demand+writing+released+prompts.htm>
- Kirst, M & Mazzeo, C. (1996, April 8-12). *The rise, fall and rise of state assessment in California, 1993-1996*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.
- Klein, S. (2008). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. *Institute of Mathematical Statistics Collections, Probability and Statistics: Essays in Honor of David A. Freeman*, 2, 76–89.
- Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). *The collegiate learning assessment: facts and fantasies*. White Paper.
- Klein, S. P., Jovanovic, J., Stecher, B. M., McCaffrey, D., Shavelson, R. J., Haertel, E., Solano-Flores, G., & Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. *Educational evaluation and policy analysis*, 19 (2), 83-97.
- Klein, S. P., McCaffrey, D., Stecher, B., & Koretz, D. (1995). The reliability of mathematics portfolio scores: Lessons from the Vermont experience. *Applied Measurement in Education*, 8 (3), 243-260.

- Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K., & Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11* (2), 121-138.
- Koretz, D., Barron, S., Mitchell, M., & Stecher, B. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)*. RAND Corporation.
- Koretz, D., McCaffrey, D., & Hamilton, L. (2001). *Toward a framework for validating gains under high-stakes conditions*. CSE Technical Report 551. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13* (3), 5-10.
- Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment, 8* (4), 279-315.
- Lawrenz, F., Huffman, D., & Welch, W. (2000). Considerations based on a cost analysis of alternative test formats in large scale science assessments. *Journal of Research in Science Teaching, 37* (6), 615-626.
- Linn, R. L. (2000). Assessment and accountability. *Educational Researcher, 29* (2), 4-16.
- Linn, R. C., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessments: Expectation and validation criteria. *Educational researcher, 20* (8), 15-21.
- Linn, R. L., Burton, E., DeStafano, L., & Hanson, M. (1996). Generalizability of new standards project 1993 pilot study tasks in mathematics. *Applied Measurement in Education, 9* (3), 201-214.
- Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 688-695*.
- McBee, M. M., & Barnes, L. L. (1998). Generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education, 11* (2), 179-194.
- McDonnell, L. M. (2004). *Politics, persuasion and educational testing*. Cambridge: Harvard University Press.
- McDonnell, L. M. (1994). Assessment polity as persuasion and regulation. *American Journal of Education, 102* (4), 394-420.
- McDonnell, L. M. (2009). Repositioning politics in education's circle of knowledge. *Educational Researcher, 38* (6), 417-427.
- Meier, S. L., Rich, B. S., & Cady, J. (2006). Teachers' use of rubrics to score non-traditional tasks: factors related to discrepancies in scoring. *Assessment in Education, 13* (1), 69-95.
- Meisels, S. J., Xue, Y., & Shablott, M. (2008). Assessing language, literacy, and mathematics skills with "Work Sampling for Head Start." *Early Education and Development, 19* (6), 963-981.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23* (2), 12-23.

- Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement*, 24 (4), 367-378.
- Nadeau, J-F R., & Godbour, P. (2008). The validity and reliability of a performance assessment procedure in ice hockey. *Physical Education and Sport Pedagogy*, 13 (1), 65-83.
- National Assessment of Educational Progress. (1987). *Learning by doing: A manual for teaching and assessing higher-order thinking in science and mathematics*. Report No. 17-HOS-80. Princeton, NJ: Educational Testing Service.
- National Assessment of Educational Progress. (2009a). Retrieved September 9, 2009, from <http://nces.ed.gov/nationsreportcard/>
- National Assessment of Educational Progress. (2009b). *Writing Framework for the 2011 National Assessment of Educational Progress*. (Pre-Publication Edition). Retrieved September 9, 2009, from <http://www.nagb.org/publications/frameworks.htm>.
- National Center for Education Statistics. (1995, January). *Windows into the classroom: NAEP's 1992 writing portfolio study*. Washington DC: U.S. Department of Education.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America*. Boston: Author.
- National Occupational Competency Testing Institute. (2009). Retrieved September 9, 2009, from <http://www.nocti.org/>
- National Research Council. (2006). *Systems for state science assessment*. Committee on Test Design for K-12 Science Assessment. M. R. Wilson & M. W. Berthenthal, (Eds.), Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- National Research Council. (1996). *National science education standards*. Washington, D.C.: National Academy Press.
- National Research Council. (2008). *Assessing accomplished teaching: Advanced-level certification programs*. M. D. Hakel, J. A. Koenig, & S. W. Elliott (Eds.). Washington D.C.: National Academy Press.
- Niemi, D., Baker, E. L., & Sylvester, R. M. (2007). Scaling up, scaling down: Seven years of performance assessments development in the nation's second largest school district. *Educational Assessment*, 12 (3&4), 195-214.
- NOCTI (2009). *Site coordinator guide for student assessment*. Retrieved September 14, 2009, from http://www.nocti.org/PDFs/Coordinator_Guide_for_Student_Testing.pdf
- Office of Technology Assessment (1992). *Testing in America's schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.
- Organisation for Economic Co-operation and Development. (2009). *Programme for International Student Assessment*. Retrieved September 8, 2009, from <http://www.pisa.oecd.org/>

- Palm, T. (2008, April). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation*, 13 (4). Retrieved August 31, 2009, from <http://pareonline.net/getvn.asp?v=13&n=4>
- Parke, C. S., & Lane, S. (2008). Examining alignment between state performance assessments and mathematics classroom activities. *Journal of Educational Research*, 101 (3), 132-146.
- Pearson, P., Calfee, R., Walker Webb, P., & Fleischer, S. (2002). *The role of performance-based assessments in large scale accountability systems: Lessons learned from the inside*. Washington DC: Council of Chief State School Officers.
- Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The performance assessment for California teachers (PACT). *Journal of Teacher Education*, 57 (1), 22-36.
- Pecheone, R. L., & Kahl, S. (n.d.). *Lessons from the United States for developing performance assessments*. Unpublished manuscript.
- Peng, S. S., Wright, D., & Hill, S. T. (1995). *Understanding racial-ethnic differences in secondary school science and mathematics achievement* (NCES 95-710). Washington, DC: U.S. Department of Education.
- Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan*, 66 (9), 628-634.
- Queary, P. (2004, March 5). Senate passes WASL changes. *Seattle Times*.
- Raizen, S., Baron, J. B., Champagne, A. B., Haertel, E., Mullis, I. N. V., & Oakes, J. (1989). *Assessment in elementary school science education*. Washington, DC: National Center for Improving Science Education.
- Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M.C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction*. Boston: Kluwer Academic Publishers.
- Rohten, D., Carnoy, M., Chabran, M., & Elmore R. (2003). The Conditions and Characteristics of Assessment and Accountability. In M. Carnoy, R. Elmore, & L.Siskin (Eds.), *The new accountability: High schools and high stakes testing.*, New York: Taylor & Francis Books.
- Romberg, T. A., Sarinia, E. A., & Williams, S. R. (1989). *The influence of mandated testing on mathematics instruction; Grade 8 teachers' perceptions*. Madison, WI: National Center for Research in Mathematical Science Education, University of Wisconsin-Madison.
- Salahu-Din, D., Persky, H., & Miller, J. (2008). *The Nation's Report Card: Writing 2007* (NCES 2008-468). Washington DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.
- Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project, research and development series* (NCES 2005-457). Washington DC: U.S. Department of Education, National Center for Education Statistics.

- Shapley, K. S., & Bush, M. J. (1999). Developing a valid and reliable portfolio assessment in the primary grades: Building on practical experience. *Applied Measurement in Education*, 12 (2), 111-132.
- Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher*, 22-27.
- Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30 (3), 215-232.
- Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (2005). Windows into the mind. *Higher Education*, 49, 413-430.
- Shepard, L. A., & Dougherty, K. C. (1991). *Effects of high-stakes testing on instruction*. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago.
- Shulte, B. (2002, February 4). MSPAP Grading Shocked Teachers. *Washington Post*.
- Solano-Flores, G., Jovanovic, J., Shavelson, R. J., & Bachman, M. (2001). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education*, 21 (3), 293-315
- Spector, J. M. (2006). A methodology for assessing learning in complex and ill-structured task domains. *Innovations in Education and Technology International*, 43 (2), 109-120.
- Stecher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practices. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability*. MR-1554-EDU. Santa Monica: RAND.
- Stecher, B. M., & Klein, S. P. (1997, Spring) The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis*, 19 (1), 1-14.
- Stecher, B. M., & Mitchell, K. J. (1995). *Portfolio driven reform: Vermont teachers' understanding of mathematical problem solving*. CSE Technical Report 400, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Taylor, C. S. (1998). An investigation of scoring methods for mathematics performance-based assessments. *Educational Assessment*, 5 (3), 195-224.
- Topol, B., Olson, J., & Roeber, E. (2010). *The cost of new higher quality assessments: A comprehensive analysis of the potential costs for future state assessments*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.
- United States Department of Education. (1995). Section 2: Reform through Linking Title I to Challenging Academic Standards. In *Mapping out the National Assessment of Title I: The Interim Report*. Retrieved September 9, 2009, from <http://www.ed.gov/pubs/NatAssess/sec2.html>.
- U.S. Department of Education. (n.d.). *Windows into the classroom: NAEP's 1992 writing portfolio study*. Washington DC: Author.

- Washington State Institute for Public Policy (2006). *Tenth-grade WASL strands: Student performance varies considerably over time*. Olympia: Author.
- Webb, N. M., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education*, 13 (3), 277-301.
- Wentworth, N., Erickson, L. D. et al. (2009). A paradigm shift toward evidence-based clinical practice: Developing a performance assessment. *Studies in Educational Evaluation*, 35 (1), 16-20.
- White, K. (1999) Kentucky: To a different drum. Quality Counts '99 Policy Update. *Education Week*. Retrieved September 10, 2009, from <http://rc-archive.edweek.org/sreports/qc99/states/policy/ky-up.htm>
- Zelinsky, A. L., & Sireci S. G. (2002). Technological innovations in large-scale assessments. *Applied Measurement in Education*, 15 (4), 337-362.

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

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

Carol Campbell, Executive Director



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

<http://edpolicy.stanford.edu>