

# What have we learned from pioneers in innovative assessment?

Brian Stecher and Laura Hamilton, RAND

## Introduction

This paper describes recent efforts to develop and implement innovative forms of assessment and summarizes some of the lessons educators can learn from those efforts. We include both pioneering state efforts to innovate large-scale assessment systems that took place in the U.S. in the 1990s and current operational testing programs, both domestic and international, that stretch the form, delivery or content of assessment for individuals. Most of the examples are drawn from K-12 education, but we also look to other levels of education where interesting efforts are underway and other sectors where relevant examples can be found.

Before we answer the question “What have we learned from pioneers in innovative assessment?” we must address the prior question, “What do we *mean* by ‘innovative assessment’?” Multiple choice testing was innovative when first introduced on a large scale in the early 20<sup>th</sup> Century, but it would hardly be considered so today. In fact, we suspect that this meeting was prompted concerns about the validity and impact of large-scale, multiple-choice testing for accountability that predominates in the U.S. today. We interpret innovation in assessment to mean “different from the norm;” in the current setting that means structurally different from pencil-and-paper, multiple choice testing (PPMC).<sup>2</sup>

There are many ways that assessment can be different in form from PPMC testing. For example, assessment can incorporate innovative *prompts* that are more complex than are found on typical printed tests. Such prompts might include hands-on stimulus materials (e.g., science

---

<sup>1</sup> Paper prepared for the National Research Council, Board on Testing and Assessment, Workshop Series on Best Practices in State Assessment, December 10-11, 2009. This paper has not been formally reviewed and does not represent the views of the RAND Corporation.

<sup>2</sup> For our purposes, multiple-choice testing includes true-false tests, matching tests, and other forms of assessment in which the respondent selects answers from a predetermined set of options.

equipment), video stimuli (e.g., records of actual events), or multiple types of materials (e.g., “in basket” tasks). Assessments can have innovative *response options*, such as constructed written responses, collections of respondent-produced materials (portfolios), or interactive responses to computer-administered problems. Innovative responses usually require more complex (and potentially innovative) scoring procedures, including systematic human judgment and “intelligent” computer scoring. Another way in which assessments can be innovative is to use different *delivery systems*, such as computer-administered assessments<sup>3</sup>, computerized adaptive testing, or assessment center exercises. The innovative domain we have outlined here is potentially quite large, including any assessment using one or more non-PPMC structural components: prompts, response options (with accompanying scoring) or delivery systems<sup>4</sup>. Other terms which have been used to describe part or all of this domain include “alternative assessment” and “performance assessment;” the latter emphasizes the use of constructed responses.

The final bit of stage setting before we summarize what has been learned about innovative assessment is to characterize the kinds of issues covered in this review. Broadly speaking, we are interested in five themes: the structure of the assessment, the technical quality and usefulness of the results, the impact of the assessment on practice, the burden imposed by the assessment on respondents and users, and the reactions of stakeholders, particularly in the political realm. The structural features include those noted in the preceding paragraph having to do with the form of the assessment and the nature of its demands on respondents. Technical quality includes, as appropriate, agreement among raters (reliability of the rating process), the

---

<sup>3</sup> We do not include computer-administered multiple-choice tests in our definition of innovative assessments.

<sup>4</sup> Although we do not attend to it, assessments can also be innovative in terms of their reporting options. We would consider real-time scoring/reporting, reporting with guidance for teaching/learning, and the like to be innovations worthy of further attention.

reliability of student scores, the fairness of the assessment for different population groups, and the validity of scores for particular inferences. Tests used in the context of 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. In many cases, innovative assessments have led to changes in practice, and we describe these where they have been documented. We find some evidence about the burdens imposed by innovative assessments relative to PPMC testing. 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). Finally, we discuss the perspective of various stakeholders, particularly politicians, because some of these assessment systems were initiated or eliminated by the actions of politicians.

In the following section, we present evidence on these four themes for a diverse set of non-PPMC assessments that were implemented and then fell out of favor during the 1990s. Subsequently, we review selected assessment innovations in operation since 2000. In most cases, the evidence is incomplete, so not all descriptions cover all the topics mentioned above.

### **What has been learned from innovative K-12 testing efforts in the 1990s?**

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 ten states were exploring the possibility of, or developing plans for, various forms of performance assessment, for a total of 24 states interested in, developing or using performance assessment (Aschbacher, 1991). Nineteen years later, the use of performance assessment has been scaled back significantly. NCLB was a factor in some state

decisions (either because the requirement that all students in grades 3-8 receive individual scores in reading and math each year or because proposed performance assessments were not approved by peer review panels) but as the following review will show some states abandoned their innovating testing for a variety of other reasons. In this section we recap some of this history to explore the reasons for the enthusiasm of the 1990s and the indifference of the 2000s.

### *Vermont Portfolio Assessment Program*

Educators in Vermont began to develop the Vermont Portfolio Assessment Program in 1988. They shared the twin goals of providing high-quality data about student achievement (sufficient to permit comparisons of schools or districts) and inducing instructional improvement. 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 about the reliability of the scoring process and the overall validity of the portfolio system as an indicator of school quality (Koretz et al., 1994).<sup>5</sup> Results suggested that readers had difficulty scoring individual pieces consistently; correlations among raters at the piece level ranged from 0.35 to 0.45 in writing

---

<sup>5</sup> The state recognized that the program was innovative and might require some revision and improvement over time; they cooperated with researchers who wanted to study its quality and impact.

across grades 4 and 8 in 1991-92 and 1992-93, and ranged from 0.34 to 0.50 in mathematics for the same grades and years). Rater reliability improved at the dimension and total score level, but it was still not high enough to draw conclusions about individual student performance. (Rater reliability at the dimension level was 0.05 to 0.10 points higher than at the piece level; and rater reliability at the student total score level ranged from 0.49 to 0.63 in writing and from 0.53 and 0.79 in mathematics.) 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<sup>6</sup>) suggesting that rater training may have played a large role in unreliability in the early years. Another factor affecting score reliability was task variance. Generalizability analyses suggested greater consistency in student performance across writing tasks than across mathematics tasks, where the task by student interaction was one of the larger variance components. One consequence of the relatively large task-by-student interaction is that as many as 25 pieces would have to be included in the student's mathematics portfolio to obtain a student score with reliability of 0.85 (Klein et al, 1995).

While lack of consistency in scoring individual pieces and dimensions is troubling, the Portfolio Assessment was designed to provide comparisons on dimensions at the school level not at the student level. Unfortunately, the scores were not accurate enough to support such school-level reporting, both because of the underlying unreliability of individual pieces and because of the small number of students in many schools (Koretz et al., 1994). Given the low reliability of scores it is not surprising that researchers found limited evidence for the validity of scores as an

---

<sup>6</sup> Daniel Koretz, personal communication, 2009

indicator of school performance. For example, mathematics portfolio scores correlated as highly with scores on the Uniform Test of writing as they did with scores on the uniform test of mathematics. Thus, researchers concluded that portfolio scores were not of sufficient quality to be useful for accountability purposes (Koretz et al., 1994).

On the other hand, researchers did find that the portfolio assessment program had a powerful positive effect on instruction, leading to changes that were consistent with the goals of the developers. For example, 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. The program did involve substantial classroom time, but teachers and principals thought it was a “worthwhile burden.” In fact, in the first years, many schools expanded their use of portfolios to include other subjects. Yet, there were potential shortcomings to some of these changes; researchers noted that teachers in Vermont engaged 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 and Mitchell, 1995). Raters reported that many portfolios contained inappropriate pieces, suggesting that teachers did not really understand the type of work students were supposed to be doing. The program also came with substantial costs, which were never fully assessed because they were distributed throughout the system among teachers, principals, coordinators, etc. Scoring costs alone were estimated to be \$13 per mathematics portfolio. In the end, the Vermont portfolio assessment program was not able to overcome some of the fundamental tensions between quality measurement and instructional improvement.

Largely due to concerns about the quality of the scores, the portfolio assessment program was replaced in the late 1990's with the New Standards Reference Exam (NSRE) (Rohten et al.,

2003). 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. 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 spring 1992<sup>7</sup>. KIRIS tested students in grades 4, 8 and 12 in a three-part assessment that included multiple-choice and short-essay questions, performance "events" requiring students to solve practical and applied problems (working first in groups and then individually), 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).

---

<sup>7</sup> 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.

KIRIS was designed as a school-level accountability system, and schools received rewards or sanctions based on the aggregate performance of all their students.<sup>8</sup> School ratings were based on a combination of cognitive and non-cognitive indicators (including drop out rates, retention rates, and attendance rates), and the school accountability index that was used to combine cognitive and non-cognitive indicators 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 school improvement at the discretion of the faculty. 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.

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 and Stecher, 1996). Teachers were more likely to report that open-response items and portfolios had an effect on practice than multiple choice items or performance events. Teachers devoted effort to aligning classroom instruction with KIRIS assessments, including using KIRIS-like tasks and “direct test preparation.” Principals, too, supported the reform; although they found it to be burdensome, they reported that the benefits outweighed the burdens. Most encouragingly, scores were improving.

However, researchers raised questions about the quality of the portfolio scores, the nature of the changes that were occurring in classroom practices, and the validity of the score gains

---

<sup>8</sup> Including students with disabilities.

(Koretz, Barron, Mitchell and Stecher, 1996). Unlike Vermont, where raters assigned a score to each piece in a student's portfolio, Kentucky portfolios were only assigned a single score for their portfolio as a whole. 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.<sup>9</sup> 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 addition, 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. Finally, researchers found that fourth grade teachers were spending significantly more time on writing and science (subject tested in fourth grade) than teachers in other grades while fifth grade teachers were spending significantly more time on mathematics and social studies (subjects tested in fifth grade) than teachers in other grades. This behavior might lead to inflated scores on a grade-to-grade basis, undermining the inferences one can draw about overall student performance. One telling comparison was that scores on KIRIS were improving while comparable scores on NAEP and on the American College Testing program were not (Koretz and Barron, 1998). Yet, the state standards were generally consistent with the NAEP standards. 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 to demonstrate accountability (Fenster, 1996).

The Kentucky State 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

---

<sup>9</sup> Mathematics portfolio scores were not included in the accountability index in the early years.

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 NCLB reporting, the Kentucky Core Content Test (KCCT) for math (grades 3-8, 11), ELA (grades 3-8, 10) and sciences (grades 4, 7, 11), which includes some constructed response items.

However, the KCCT continues to assess student achievement in writing 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 (grade 12 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).

Interestingly, after having assessed writing using portfolios and writing prompts for over fifteen years, there remains sufficient concern among parents and other groups that the Kentucky Department of Education (KDE) in 2008 published a fact sheet on "*Considering Myths Surrounding Writing Instruction and Assessment in Kentucky*" (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 Program (MSPAP)*

The Maryland School Performance Assessment Program (MSPAP) was created in the late 1980s and early 1990s to assess progress towards the state's educational reform efforts. 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, all responses were scored by human raters. 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 which were not scored. MSPAP items were matrix sampled, i.e., every student took a portion of the exam in each subject but not all items. 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. In addition, there were rewards and sanctions at the school level associated with performance on the MSPAP (Pearson et al., 2002).

Reports from the state testing contractors showed that the test scores met reasonable expectations for reliability and validity. For example, in 1997 the alpha internal consistency coefficients in the calibration samples in each content area for grades 3, 5 and 8 were all about 0.85, except for writing, which was generally around 0.70 (Maryland State Department of Education, 1998). 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.

MSPAP was also designed to influence instruction, both through incentives to achieve higher scores and through modeling of more complex and authentic types of tasks. Teachers and principals reported that MSPAP had a positive effect on instruction (Lane et al., 2000). For example, researchers reported that most mathematics teaching activities were aligned with the state standards and performance assessments (although classroom assessments were less consistent with state assessments) (Parke and Lane, 2008). Teachers in Maryland also 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, and Stone, 2002).

Nevertheless, many of the features of MSPAP were unusual for state testing programs, and some stakeholders raised concerns about the quality of MSPAP school results. 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. The external evaluation commissioned by the Abell Foundation criticized the content of the tests (Hambleton et al., 2000), and there were objections from some to the Maryland Learning Outcomes on which the test was based (Ferrara, 2009). Some prominent supporters of the program turned against it as well. Partially due to these concerns and partially due to a desire (and an NCLB requirement) to have individual student scores, MSPAP was replaced in 2002 by the Maryland School Assessment (MSA) (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 3-8 and 10), writing (grades 4, 7, and 10), mathematics (grades 3-8 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 were reported for schools and districts.

Researchers examined the technical quality of WASL scores, the impact of WASL on student performance, and the links between teacher practices and WASL scores. Test scores on multiple choice components met common standards for reliability; for example, in 1998 the reliability and/or generalizability coefficients for the fourth grade Listening, Reading, Writing and Mathematics tests were 0.60, 0.87, 0.70 and 0.88, respectively (Taylor, 1999).<sup>10</sup> Similarly, inter-rater agreement on the open-ended items was also reasonably high; for example, exact agreement between pairs of judges on the Reading and Listening open-ended items ranges from 76 to 97 percent.

Much like in Vermont and Kentucky, researchers found that teachers adapted classroom practice to align with the emphasis they perceived to be included in WASL, and there was some weak evidence that changes in practice were related to improvement in WASL scores at the

---

<sup>10</sup> The writing value is a generalizability coefficient, the others are alphas coefficients (measuring internal consistency).

school level (Stecher, Chun, Barron and Ross, 2000; Stecher and Chun, 2002). However, negative findings surfaced as well, and there was much disagreement and controversy about the test, which led to delays in full-scale implementation and changes in plans.

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 in light of the low pass rates of 10th graders (Queary, 2004). Other concerns arose during this period; for example, there was 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 3-8 and the High School Proficiency Exam (HSPE) in grades 10-12. 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 EARL 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 and Mazzeo, 1996). First administered in 1993, CLAS assessed students' achievement in reading, writing and mathematics in grades 4, 8 and 10. In reading and writing, CLAS used group activities, essays, and reading short stories to measure students' critical thinking. In math, students were asked to show how they had 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 and Mazzeo, 1996). In addition, the debate in California highlighted a 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, CLAS was reauthorized by the legislature 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).

*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 assess 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*

This brief history of states’ use of innovative assessment in the 1990s suggests that renewed enthusiasm for assessment innovation should be tempered by the realities of implementation. The boldest state assessment innovations from the 1990s, including the use of unstructured portfolios, group assessments, and problem solving tasks, did not survive their use in a large-scale accountability context. In hindsight it appears that these changes foundered on a few practical realities. First, states may have moved too quickly on too large a scale. Although there were nominally “pilot” phases in many cases, policymakers rushed to implement without the benefit of adequate developmental cycles. Problems with scoring, score reporting, content, etc., could have been reduced or eliminated with a more gradual and nuanced roll out. Second, states may have pushed the envelope too far in terms of innovation. They attempted bold changes, in some cases rushing ahead faster than the underlying science warranted. Third, it was difficult to sustain support in the face of the costs and burdens associated with the assessments. It was not clear that the resources needed for development, scoring, etc., and the classroom time devoted to assessment related instruction yielded adequate benefits. Fourth, states could not always maintain the human and financial resources necessary to continue to operate the assessment systems over time. Fifth, states did not give adequate attention to the potential concerns of stakeholders and the demands of politicians. It is one thing to offer a new “cutting edge” product for sale in the marketplace, where buyers have the option of being in the vanguard or waiting for the next version. It is something else to withdraw an existing product in favor of a new one without any buyer choice. At a minimum, a better public education effort was needed

to help stakeholders understand the innovation. To insure acceptance, it might have been better to engage stakeholders in development and shaping of the reform. Sixth, states were not always able to reconcile the views of politicians and assessment developers about the roles of assessment (e.g., as a tool for measurement, instructional improvement, persuasion and regulation) and the implications of the innovation. When conflicts could not be resolved, support waned. Finally, the NCLB requirement for annual test scores on every student in grades 3 to 8 (and one high school grade) could not be met by many of the innovative programs because they tested in only some grades or did not provide scores for each student; modifying these programs to meet the NCLB guidelines would have been both costly and burdensome.

### **Innovative Assessments in Use Today**

In this section we describe several examples of assessments that are currently in operation and that incorporate innovative features. These examples are drawn from K-12 education as well as from other realms. We explore three broad approaches: (1) on-demand performance assessments; (2) portfolios (including video); and (3) computerized testing, which includes computerized adaptive testing, computer-administered simulations or other complex prompts, and automated scoring of open-ended items. We identify one or more examples to illustrate each approach. The set of approaches and specific assessments discussed here is far from exhaustive, but is intended to demonstrate some of the ways in which large-scale assessment programs have successfully incorporated innovative formats, delivery systems, and scoring procedures.

#### *On-Demand Performance Assessment*

Despite the lack of enthusiasm for the statewide K-12 performance assessments described earlier, performance assessments continue to be used in various contexts both within and outside

the K-12 education sector. Written essays are particularly common and are used by many states to assess writing skills and to supplement multiple-choice items in mathematics, reading, and other subjects. They are also widely used in other secondary school assessment programs such as the SAT and Advanced Placement (AP) exams, in admissions tests for graduate schools as well as licensure and certification exams, and in NAEP. They may be administered using paper and pencil or computers. Although the essay format is certainly not “innovative,” its use in large-scale accountability testing has been uneven and has posed technical and cost challenges, so we include it in our discussion. We also discuss other forms of performance assessment. We refer to this class of assessments as “on-demand” because they require students to take them at a particular time and in a particular place, as opposed to portfolio tasks that students complete on their own time outside a formal testing context.

One current example of the essay format is the Multistate Essay Examination (MEE) component of the Bar Examination. MEE includes nine 30-minute essay questions, and officials in each state may select a subset of these for inclusion in their state examination system (see <http://www.ncbex.org/multistate-tests/mee/mee-faqs/description-of-the-mee>). States also have responsibility for scoring the essays and for determining how they are used (e.g., how much weight the essay scores receive relative to scores on the multiple-choice portion of the exam), so this system provides more opportunity for state-level customization than many other assessment programs. Because scores on the essays tend to be highly correlated with those on the multiple-choice portion and the two formats are generally intended to measure similar constructs (Kane & Mroch, in press), it is feasible to combine the scores from the two components into a single score that is used to determine whether an examinee passes. This example illustrates one approach to addressing the often inadequate levels of score reliability for open-ended tasks—combining them

with other test-score data produces a composite score that has higher reliability. However, that composite also lacks information about any unique constructs that the open-ended portion was intended to measure. This example also shows how assessments can be built to address common objectives while also being customized to meet the needs of different jurisdictions.

National assessments used in other countries provide several additional examples of essays and other open-ended response formats. Compared with large-scale assessment programs in the U.S., many other nations' systems rely much less on multiple-choice items and much more on open-ended prompts that involve extensive writing, with an emphasis on higher-order skills such as analysis (Darling-Hammond & McCloskey, 2008). The assessment system used in Queensland, Australia illustrates some of the innovative features that have been incorporated into assessments used elsewhere. The Queensland Studies Authority's (QSA) assessment comprises a variety of assessments linked to common course syllabi, with each set of assessments serving a specific purpose. The system includes diagnostic assessments for the youngest students with locally and centrally developed assessments for students in grade levels 3 through 9. The national assessment system that is intended to support comparisons across states and territories is administered in a subset of grade levels and relies heavily (but not exclusively) on multiple-choice items, but it is supplemented with a set of performance-based tasks (Queensland Comparable Assessment Tasks, or QCAT) that are administered in the years when the national assessment is not given, and with a bank of assessment tasks that can be used at local educators' discretion. Both the QCAT and the additional assessment tasks rely heavily on essays that are scored locally by teachers, a practice that is intended both to support teachers' professional learning and to maximize the utility of the information for instructional decision making (see <http://www.qsa.qld.edu.au/assessment/3111.htm>).

The variety of assessment types at the high school level in Queensland is even larger and includes self-directed projects as well as centrally developed exams consisting of essays, oral recitations, and performances. Several components of the system rely heavily on written constructed responses, but because the scoring of many of these is done locally rather than centrally, the scoring burden is distributed across a larger number of people and jurisdictions than is the case with programs that rely on centralized scoring. Of course local scoring raises concerns about comparability, but Queensland's system is explicitly not intended to support comparisons of schools with one another, so comparability across schools is not as critical as in systems such as those adopted in the U.S. in response to NCLB. In fact, many of the features of the QSA system appear to have been adopted in response to a desire to avoid perceived pitfalls associated with U.S.-style testing, such as score inflation and narrowed instruction (Queensland Studies Authority, 2009).

An additional example of performance assessments comes from the higher education sector in the United States. The Collegiate Learning Assessment (CLA) is one of several measures that have been developed to assess student learning in colleges and universities. CLA has been administered across a wide range of postsecondary institutions, and its reach continues to expand. It is administered on-line and consists of two task types—Performance Tasks and Analytic Writing Tasks—that require students to produce written responses in which they apply critical thinking, problem solving, and communication skills (see [http://www.collegiatelearningassessment.org/files/Architecture\\_of\\_the\\_CLA\\_Tasks.pdf](http://www.collegiatelearningassessment.org/files/Architecture_of_the_CLA_Tasks.pdf) Like the MBE and some of the QSA measures discussed above, the Performance Task component of CLA asks students to respond to a variety of stimulus materials such as letters, maps, photographs, research reports, and diagrams. Until recently, responses have been centrally

scored by trained raters. CLA is exploring the use of automated essay grading, an application of technology that we discuss later in this paper. Because the CLA is intended to support inferences about the performance of institutions rather than individual students, it uses matrix sampling to allow the administration of multiple tasks with each institution, even though each student completes only one Performance Task or two Analytic Writing Tasks (Klein et al., 2008).

Although most performance assessments primarily involve written materials, some fields have adopted assessments that also include equipment, often called “hands-on” assessment. This format is most commonly used in science, in an effort to engage students in the scientific inquiry process, as discussed earlier in this paper. An example from outside K-12 education is the United States Medical Licensing Examination (USMLE) Step II Clinical Skills Examination which requires prospective physicians to interact with standardized patients. Designed to “test medical students and graduates on their ability to gather information from patients, perform physical examinations, and communicate their findings to patients and colleagues” (USMLE, 2009, p.3), this assessment involves twelve different cases in which the examinee interacts with a trained participant who presents with a specific set of complaints or symptoms. The USMLE Step II Clinical Skills Examination probably comes closer than almost any other assessment to achieving one of the primary goals of performance assessment by measuring an examinee’s performance in a context that is very similar to the real-life behavior that the test is designed to predict. This assessment has encountered many of the technical challenges described earlier, including variability stemming from both raters and tasks (with the standardized patients essentially functioning as tasks; see Clauser et al., 2008). The inclusion of twelve cases (a few of which are used for pilot purposes and not counted in the scores) helps to ensure broad coverage

across a range of areas of medicine and therefore addresses the task sampling problem that other performance assessments have encountered. The drawback of including this number of patients is its effect on test length; the exam requires eight hours to complete. Double scoring is used to address rater variability, but because the exam is scored as pass/fail, double scoring is done only for examinees whose scores from a single rater are close to the cut score. This ensures greater accuracy where it is needed, near the cut score, but keeps rating costs to a minimum by not requiring every set of responses to be double-scored.

As this discussion illustrates, the use of performance-based assessment tasks raises several challenges related to cost, feasibility, and technical quality. The nature and extent of these challenges depends in large part on the features of the tasks. The discussion in this section illustrates the wide range of task types, from relatively brief written tasks that rely on paper and pencil to extended activities that involve equipment and other people and that closely simulate a real-world experience. In the K-12 context, where large-scale assessments are typically designed to support inferences about a broad range of skills and knowledge, hands-on tasks or written essays that are relatively brief, and therefore that can support the administration of multiple tasks, are probably most appropriate, though for some specialized domains it may be desirable to use an approach that more closely resembles the USMLE Step II exam. As discussed in the context of the MEE, it may be beneficial to combine scores on hands-on tasks with scores on assessments in other formats to produce scores with adequate reliability.

In all of the examples discussed here, developers and users have developed strategies to address the technical and logistical challenges that contributed to the demise of many states' K-12 performance assessment programs, but some of the solutions would be difficult to enact in the context of today's state accountability systems. Task sampling is one of the most pervasive

problems in performance assessment—multiple tasks are generally needed to support inferences about the construct of interest. CLA addresses the task sampling problem through matrix sampling, which works well for the kinds of school-level inferences CLA is designed to support but does not provide the individual student scores that are needed under NCLB. The USMLE, by contrast, is intended to support inferences about individual examinees and must administer a large enough set of tasks to each prospective physician to provide an accurate estimate of proficiency. As a result of the need for multiple tasks and the complexity of each task, the time required for testing is much longer than would generally be feasible for a statewide K-12 assessment program.

Another challenge associated with performance assessment is the tension between producing scores that can be compared across institutions or jurisdictions and the desire to make the assessment process instructionally useful for teachers and students. The Queensland program and the essay portion of the Bar Exam rely heavily on local input into task selection and scoring, which may enhance the utility of results for local decision making but which also limits the comparability of scores. Centralized task selection and scoring are typically necessary for high-stakes programs that require comparability, but these features may reduce local buy-in and support for the program and diminish the value of the information for some kinds of decisions.

Another characteristic that is common to the assessments discussed here is that none of them involve the kind of consecutive-grade administration that NCLB has required. The Bar Exam and USMLE exams are used to support inferences about performance at a single point in time, and CLA is often used to measure change from freshman to senior year, requiring two administrations. The Queensland system comes closest to NCLB in its administration schedule but the performance tasks are not administered each year, nor are they intended to support

inferences about change over time, something that is likely to be required in a reauthorized set of federal accountability provisions in the U.S. In short, each of the examples presented in this section has features that make it work well for its intended purpose, but it would be difficult to adopt any of these approaches in today's K-12 testing system without making significant changes to state policy surrounding accountability and other uses of test-score data.

### *Portfolio-Based Assessment*

As illustrated by the Vermont experience, portfolios typically involve using products drawn from the regular instructional program as inputs into the assessment system, so that scores do not depend on how students respond to an on-demand set of tasks but are based on work they have done outside the formal assessment setting. The inclusion of these types of tasks is common in the assessment systems used in other high-performing nations, many of which incorporate performance on reports, science investigations or projects into their national assessment systems (Darling-Hammond & McCloskey, 2008). Integrating this type of student work into the large-scale assessment has the potential to broaden the range of topics and skills that are measured while promoting better coherence between instruction and assessment than is typically obtained with on-demand testing. However, as discussed earlier, it raises technical concerns related to scoring and standardization. Some other nations have attempted to enhance consistency by providing centrally developed tasks and scoring systems but allowing local educators some discretion over what is administered and when (Darling-Hammond & McCloskey, 2008).

A prominent example of a portfolio-based assessment system that is in use today is the National Board for Professional Teaching Standards (NBPTS). This assessment is designed for K-12 teachers rather than students, but many of the experiences of the NBPTS are relevant to

student assessment. The NBPTS assessment includes a computer-based portion that measures teachers' knowledge using six constructed-response items, and a portfolio component that includes four exercises designed to allow teachers to demonstrate their teaching skills. The assessment is designed to align with content standards that the National Board developed to describe attributes of accomplished teachers. The board developed five "core propositions" and used these to create specific standards for each of the 25 teaching fields in which teachers can apply for National Board certification. The core propositions and the corresponding standards address teachers' content knowledge but also include many so-called "soft" skills, practices, and attitudes. The core propositions include, for example, "Teachers are committed to students and their learning" and "Teachers think systematically about their practice and learn from experience" (National Research Council, 2008). These kinds of attributes are reminiscent of what many K-12 education reformers have tried to promote through new standards efforts such as the promotion of "21<sup>st</sup> Century Skills" (Partnership for 21<sup>st</sup> Century Skills, 2008). They reflect a recognition that successful workplace performance requires more than knowledge of a particular field but also depends on the attitudes and other non-cognitive attributes that workers bring to their jobs. Therefore efforts to adopt these kinds of standards in the K-12 context, and to measure student progress toward meeting them, may be informed by the NBPTS experience.

The portfolio portion of the NBPTS assessment requires examinees to complete four tasks: three "classroom-based entries" that consist of direct evidence of practice through videos and/or student work, and one "documented accomplishments entry" that involves evidence of work the teacher pursued outside the classroom and how this work affects student learning (National Research Council, 2008). The examinee submits written commentary to accompany videos or artifacts, and this commentary provides evidence related to the teacher's ability to

analyze and reflect on his or her practice. Trained raters score each submission using a four-point rubric that is provided in advance to all examinees as a way of conveying information about what is expected of them.

This assessment shares many of the drawbacks that researchers on the Vermont and Kentucky portfolios identified, as discussed earlier in this paper. The National Research Council's 2008 report documents these challenges. First, the assessment process is time-consuming: The entire assessment is estimated to require up to 400 hours of the examinee's time and typically takes 12-18 months. The board's efforts to include authentic samples of teaching, which require lengthy videos and extensive documentation, result in a relatively small number of tasks that can be included in the entire assessment—four portfolio activities and six computer-based assessment center tasks. As a result, the scores suffer from fairly low reliability stemming from task sampling variability. Rater variability is also a concern, in part because the process of translating the content standards into assessment tasks and scoring rules lacks structure. On the other hand, a validity investigation suggests that the assessment is capturing elements of accomplished teaching. The NRC panel makes some suggestions for improving reliability, such as adding some additional, shorter constructed-response or closed-ended items to the assessment center, but it acknowledges that efforts to improve reliability could compromise validity. Because of these limitations it is unlikely that portfolio assessments alone could meet the needs of today's K-12 state accountability testing programs, though they could become a part of a broader assessment strategy that combines formats that can support cross-institution inferences about performance with formats that are more clearly targeted toward supporting decision making at the school and classroom levels.

### *Technology-Supported Assessment*

Computers have played a role in large-scale testing for many decades, but this role has traditionally been limited to a set of fairly narrow activities including scoring of selected-response items and statistical modeling of scores (e.g., using item response theory (IRT) approaches). More recently, several large-scale testing programs have incorporated advances in information technology to improve the technical quality of assessments and/or to reduce their costs. Innovations that have been adopted for use in operational testing programs include computerized adaptive testing (CAT), automatic item creation through the use of item shells, automatic scoring of essays, and web-based administration of assessments. Although these advances promise to address costs and other burdens associated with the development, administration, and scoring of large-scale assessments, for the most part they have served to facilitate continued administration of the traditional multiple-choice and essay tests that have been the staple of large-scale testing. However, researchers and technology developers have recently predicted that technology is likely to usher in some fundamental changes to the assessment of student learning (Bennett, 1998; Pellegrino, Chudowsky, & Glaser, 2001; Quellmalz & Pellegrino, 2009; Tucker, 2009). In this section we describe three applications of technology that have been applied to assessment, with a recent example of each.

*Computerized Adaptive Testing (CAT).* CAT has been used in a variety of assessment contexts, including exams used to make decisions about graduate admissions, professional licensure and certification. As with some of the other approaches discussed in this paper, CAT is not a novel approach, but it deserves attention because the increasing availability of the required technology in schools offers opportunities to increase the use of CAT in large-scale assessment. Another factor that may contribute to greater interest in CAT is the pending reauthorization of NCLB. Provisions of NCLB that required grade-level testing were interpreted by most states as

prohibiting the use of CAT (Oregon is the only state that uses a CAT-based system for its NCLB testing). However, changes to the law may eliminate that restriction, particularly if the reauthorized version requires growth modeling, a technique that could benefit from the greater accuracy of individual student measurement that CAT affords for students who are performing well above or below grade level.

Probably the most well-known application of CAT in K-12 testing is the Measures of Academic Progress (MAP) developed by the Northwest Evaluation Association (NWEA). Developed to measure student achievement in mathematics, reading, language usage, and science, the MAP assessments are customized to each state's standards and can be administered multiple times throughout the year. A Rasch-based scale is used to create scale scores that are intended to measure student growth over time, providing a different type of information to educators than what is available through the state accountability testing system. These assessments also allow scores for students in different states to be placed on a common scale, thereby providing a means to compare performance across states and to evaluate the difficulty of different states' "proficiency" cut scores (Cronin et al., 2007). However, these tests are designed to serve as broad measures of achievement, and they rely on traditional item formats so they may be less ideal than other assessments for providing detailed diagnostic information about complex skills and problem solving processes.

*Computer-Based Simulations.* A second area in which technology has shown considerable promise for improving the scope and quality of large-scale assessment is in the use of simulated environments that allow students to interact with people or objects in ways that mirror real-world activity. These assessments share some features of the hands-on tasks discussed earlier, but technology can broaden the range of possible activities—for example,

students can engage in experiments that involve observing plant growth or changing the architectural features of a bridge. Computers also address some of the logistical problems that plague assessments that involve a lot of equipment and materials, and offer opportunities for rapid feedback through automated scoring. Computer-based simulations also have the advantage of being able to track students' problem-solving steps and errors in ways that are impossible when students record their responses using paper and pencil.

One of the most prominent operational examples is the USMLE Step 3 Exam used for licensing physicians (see <http://www.usmle.org/Examinations/step3/step3.htm>). It combines multiple-choice items with a simulation component in which the examinee is given a scenario describing a patient's conditions, and is asked to respond as if he or she were in an actual clinical setting—e.g., by ordering tests or treatments. The examinee enters responses into the computer in text format, and the condition of the patient changes as a result of these responses in a way that reflects what would be likely to happen in the real world. These tasks are scored automatically using rules that are consistent with expert clinical judgment.

The Minnesota Comprehensive Assessment Series II (MCA-II) science assessment illustrates the promise of technology in a statewide K-12 testing context. The assessments are described as **“scenario-based, computer-delivered tests that present students with realistic representations of classroom experiments and real-world phenomena”** (Minnesota Department of Education, 2008, p.6). They require students to respond to information presented in text, graphic, audio, and/or video form, and include multiple-choice, short constructed-response, and extended constructed-response items, along with a format called **“figural response”** that enables students to interact with graphics by manipulating elements, selecting points on a figure or graph, or dragging and dropping elements. These figural response items illustrate the power of technology to enable

**students to produce responses that are not limited to selection (as in multiple-choice items) or to written text (as in traditional essay items) and that can be scored automatically.**

*Electronic Scoring of Open-Ended Responses.* The final technology-based innovation we examine is the application of software to score open-ended responses to assessment items, often called automated essay scoring (AES; see Ben-Simon & Bennett, 2007). One reason for the popularity of multiple-choice items has been the availability of technology to score those assessments and report results quickly and inexpensively. This technology not only saves money but increases the utility of results for decision making by providing rapid feedback, so the application of automated scoring to essay formats could have significant benefits. Although the idea is sometimes greeted with skepticism from members of the public, studies suggest that in most cases the correlation between results from automatic scoring correlated as highly and scores obtained by a human rater were similar to the correlation between scores generated by two human raters (Klein, 2008). High human-computer correlations are not sufficient to assume adequate technical quality of AES (Ben-Simon & Bennett, 2007) but they do provide some useful information about how these systems are likely to work in practice. To the best of our knowledge, the NAEP Writing assessment is scheduled to be administered and scored using computers in 2011, and other testing programs already make use of automated scoring. One of these is the USMLE Step 3 exam discussed above.

Automated scoring systems vary in their methodology and algorithms but generally involve having a set of essays scored by trained human raters and giving this information to the scoring software so that it can identify a set of criteria and weights that predicts the human-assigned scores. These empirically derived scoring rules do not always align closely with the criteria that human experts would argue should be given the greatest weight (Ben-Simon & Bennett, 2007). To take a simple example, most writing experts would not use essay length as a

criterion for evaluating high-quality writing, but if length is correlated with other aspects of high-quality writing, an AES system might produce a scoring algorithm that assigns a high weight to length. This mismatch, if known to examinees or test users, could damage the credibility of the system and could encourage undesirable test preparation efforts and other “gaming” behaviors that lead to high scores in the absence of proficiency in the construct of interest. Some AES systems have addressed this problem by incorporating information about the criteria expert human raters apply into the empirical derivation of the scoring algorithm so that the resulting scoring criteria more closely reflect those that are consistent with experts’ views of proficiency (see Ben-Simon and Bennett, 2007, for a discussion of scoring approaches used by popular AES systems). It is likely that the extent of the mismatch varies across subjects, age groups, and other aspects of the testing context, but it is clearly something that users of AES systems should be aware of, particularly with respect to the likelihood of undesirable responses on the part of examinees or those responsible for preparing them for the test. There is also a risk that AES will limit the kinds of tasks that can be included in the assessment.

*Summary: Technology-Supported Assessment*

The various applications of technology discussed here have the potential to improve large-scale assessment in a relatively cost-effective way, but they do present some challenges that need to be overcome. There is some evidence that performance can be affected by examinees’ computer skills, which would lead to inaccurate inferences about performance in the domain of interest, at least in K-12 settings (Bennett et al., 2008). At the same time, however, as students become more accustomed to working with technology it is possible that they will perform better on technology-based assessments. Students who are accustomed to producing written essays on computers do benefit from computer-based essay exams (Russell & Haney,

1997), for example. Other potential barriers are the uneven availability of the required infrastructure in schools (though this is improving rapidly, and web-based administration may address this problem to some degree) and the need to provide professional development to educators so that they can prepare their students appropriately and can make use of the vast amount of data that some technology-based assessments produce. In addition, while some of the most promising advances in technology-based assessment reflect a growing understanding of how students learn, other aspects of the education system including curriculum, instruction, standards, and professional development must be aligned with this more sophisticated understanding of learning. As noted above, **some kinds of computer-based assessments can produce detailed data on students' response processes, but the potential for this kind of information to improve student learning depends on training teachers and other test users to interpret and use it appropriately.** Finally, despite the appeal and the relative success of some approaches to automated essay scoring, there are many unanswered questions about its contribution to the validity and reliability of scores. Despite these challenges, however, as we discuss in the next section, technology is likely to play a growing role in large-scale achievement testing in the K-12 sector.

### **What the Future Might Hold**

All of the examples of innovative assessments past and present suggest that developing high-quality, cost-effective assessments that measure important knowledge and skills poses significant challenges that have not yet been fully resolved. At the same time, the number and diversity of innovations that have influenced large-scale testing over the past several years is impressive and is likely to shape future test-development efforts in significant ways.

Much of the current work in assessment development involves new applications of technology to measure constructs that have been difficult to measure using existing assessment

approaches. As with the hands-on format, the subject that has been at the forefront of innovation in this area is science, in large part because of its emphasis on problem solving and inquiry. The Problem Solving in Technology-Rich Environment (TRE) project was designed by ETS researchers to explore the feasibility of including technology-based simulations in the NAEP Science Assessment. It includes several tasks that require students to apply scientific inquiry and reasoning skills. Although these tasks have not been used in the operational NAEP program, preliminary validation work suggests that scores produced from these tasks have reasonable technical quality and provide information about the kinds of skills the developers had intended to assess (Bennett et al., 2007). Other large-scale assessments for which complex, simulation-based science items have been explored are the Programme for International Student Assessment (PISA) and the statewide science test in Minnesota discussed earlier (Quellmalz & Pellegrino, 2009). These efforts not only provide opportunities to measure skills and knowledge in new ways, but they also can be designed to produce much richer student performance data than what is typically obtained through traditional testing. They can, for example, provide records of students' problem-solving strategies and errors, which can be reported and analyzed in ways that might be useful for subsequent instructional decision making. Technology could also facilitate efforts to meet the needs of students with disabilities and English language learners who often require modifications or accommodations that could be built into a computer-delivered system.

Another innovative assessment system developed at ETS is intended to improve the utility of assessment information for both accountability and instructional purposes and address the need for a coherent *system* of assessments (Shepard, Hannaway, & Baker, 2009). The CBAL (Cognitively Based Assessment *of, for, and as* Learning) involves the development of assessment tasks aligned to cognitive models of student performance in reading, writing, and mathematics.

It includes three components---accountability assessments, formative assessments, and professional support for teachers and other educators—and is designed to improve instruction in addition to serving as an accountability tool. Its features include accountability measures that are administered in small doses throughout the year rather than at a single session, as well as the use of technology that facilitates administration of novel tasks such as having students read passages aloud (O'Reilly & Sheehan, 2009).

As demonstrated by the CBAL example, innovations in technology may facilitate a more flexible approach to large-scale testing. Several scholars have predicted that the nature of K-12 education is likely to experience a fundamental shift as a result of technology that allows schools to break away from the traditional model of a teacher instructing a group of a few dozen students in a classroom, and that the potential for truly individualized instruction is significant (Christenson, Horn, & Johnson, 2008; Moe & Chubb, 2009). Whether these visions of a radically reformed education system actually materialize in the next few decades remains unclear but if they do, assessment is likely to follow suit. It is possible to imagine technology facilitating test administration at any time and place, in a way that is consistent with visions of individualized instruction. For example, an assessment would not have to be given during a two-week window in April but instead could be administered when other information suggests the student is ready to take the assessment, and could be customized to a student's profile of skills, experiences, and interests. Even though this model may be far from a reality now, it is important for policy makers and assessment developers to consider its implications when designing K-12 assessment systems today.

### **Implications for the Development of Assessment Systems and Policies**

This review of innovative assessment suggests that we can measure important achievement constructs in ways that are less constrained and more authentic than PPMC tests. There is clearly a desire among policymakers to create assessments that do a better job of measuring complex skills and processes than current tests, as evidenced recently in speeches by the President and officials from the U.S. Department of Education and by the emphasis on innovative assessments in federal funding programs. Increasing the aspects of achievement we measure and broadening the way we measure them may be beneficial because the results will reflect more of the domains that matter and will generalize to more settings in which the skills are likely to be used.<sup>11</sup>

For example, new technologies permit states to expand the scope and coverage of their existing accountability assessments, and also may facilitate the development of assessments in subjects that have not traditionally been included in state testing programs. Software that records and analyzes spoken language may allow states to measure foreign language production, an important outcome that is not measured well by PPMC. These technologies could also be used for assessments of musical skills. Similarly, technologies that allow students to manipulate and produce objects such as diagrams and graphs are useful not only for science and mathematics assessment but could facilitate measurement of skills in engineering, design, and visual arts. Of course, incorporating all of these measures into an accountability system is not necessarily feasible or desirable; these examples are simply illustrations of ways that states might broaden the scope of their assessment programs to meet the needs of students, educators, and policymakers.

---

<sup>11</sup> One issue that is critical to assessment design that we did not explore is the nature of desired performances as described in academic content standards. In a standards-based system, assessment should be developed to reflect knowledge as portrayed in the standards. PPMC tests may be an appropriate way to assess mastery of some kinds of academic standards.

Whether states choose to move to more innovative forms of assessment depends on a number of technical, practical, political, and theoretical factors, some of which were described in this paper. To recap briefly, recent history suggests that the less you constrain prompts and responses the more technically and logistically difficult it can be to obtain high-quality results. In addition to concerns about technical quality, statewide efforts to adopt innovative assessments in the 1990s were also plagued by other problems, including inadequate pilot-testing, adopting “cutting edge” approaches that were not fully operational, failure to adequately plan for or justify added costs and burdens, insufficient education and engagement of stakeholders, and lack of agreement among assessment developers and politicians about the role and purposes of assessment. These factors are still relevant to large-scale assessment innovation today.

It is also important to point out that an accountability context places important constraints on the use of assessments. History is clear that when used for accountability purposes assessment has a strong influence on classroom practice. Many of the early developments in K-12 performance assessment were motivated by the idea that it was possible to create “tests worth teaching to”—that is, tests that would not lead to undesirable effects on instruction because engaging in test preparation would be consistent with high-quality teaching (see Resnick & Resnick, 1992, for a discussion of the rationale for performance assessments). However, assessment is a two edged sword, i.e., teachers mimic in their classes the form and content of assessment, which may transform curriculum in desired ways at the same time it is narrowing curriculum and leading to score inflation. We do not know what the balance is between these positive and negative effects, but there is no approach to assessment that has been definitively shown to be resistant to the negative effects of traditional testing when used for accountability purposes. The presumed benefits of performance assessment have only partially been supported

by evidence, and the cost and logistical challenges associated with innovative assessments have created incentives for states to rely on the tried-and-true approach of multiple-choice testing.

In addition to calling for new assessments, the current policy debate has emphasized the role that data should play in decision making at all levels of the education system, from determining teacher and principal pay to informing day-to-day instructional decisions. Test-based accountability systems and improvements in technology have led to widespread availability of assessment data, and to a desire to use those data to inform various kinds of decisions. Although there is a tendency for policymakers to focus on data from large-scale assessments, data use is most likely to lead to good decisions if it incorporates information from multiple sources, and if data users understand that the quality of the underlying assessments influences the utility of the data for decision making (see Hamilton et al., 2009). Therefore policies to promote better data use should include attention to the variety of assessments that are likely to be used, and policymakers should think of assessment reform broadly to incorporate classroom-based assessments as well as large-scale assessments in a coherent system that includes professional development and other supports to help educators use information effectively and appropriately.

It is also important to consider the trade-offs inherent in any proposed use of large-scale assessment data. One of the most widely discussed uses of test-score data in today's policy debate is the development of statistical models that can be used to support inferences about educator performance and, in some cases, to make decisions about educator compensation. Most of these discussions have focused on the use of growth or value-added models that aim to isolate the effects of educators or schools from the effects of other factors (McCaffrey et al., 2003). Proponents of these models have provided strong rationales for why they are superior to single-

point-in-time testing, but the use of growth models requires testing in consecutive grade levels using assessments that adequately sample the constructs of interest and that support comparisons across institutions. These requirements may preclude the use of some types of innovative assessments, and place significant constraints on any innovations that are adopted. At the same time, experts who provided input for consideration in the Race to the Top competition argued that assessments should be designed to inform instructional decision making and that assessment policies should provide opportunities for teacher involvement in assessment development, use, and scoring (Sawchuk, 2009). Accomplishing these goals while simultaneously providing the type of data needed for large-scale longitudinal growth modeling is likely to be difficult.

Combining innovative formats with more traditional formats might be a reasonable approach for addressing some concerns related to costs, feasibility, and technical quality. Well-designed multiple-choice tests can measure many important constructs and there is no reason to abandon them as long as they are working well for some purposes. Decisions to combine formats should carefully consider the constructs measured by each portion of the assessment and the value added by the more expensive format. Although adding innovative formats may not contribute to the technical quality of the measure and may in fact hinder efforts to construct scores with high reliability (Wainer & Thissen, 1993), other considerations such as the likely effects of the assessment on instruction should also be considered. Scoring systems, regardless of whether they are automated, must be based on a careful analysis of the kinds of responses the assessment is intended to elicit (Bennett & Bejar, 1998).

In addition, assessment policies should strive to improve the integration between large-scale and classroom assessments. Despite the common practice of using large-scale achievement test data for school and classroom-level decisions (Stecher et al., 2008; Hamilton et al., 2008),

most of those assessments do not have the features that have been described as essential for formative use of assessment data. In particular, large scale assessments, which are not embedded in specific learning activities nor linked to the current instructional unit, do not provide optimum information for “diagnosing where students are in their learning, where gaps in knowledge and understanding exist, and how to help teachers and students improve student learning” (Perie, Marion, and Gong, 2007, p. 3). No single assessment is likely to serve the needs of a large-scale program and classroom-level decision making equally well. Instead, it would be preferable to have a coordinated system that includes a variety of assessment types that addresses the distinctive needs of policymakers (for accountability purposes), teachers (for instructional purposes) and other user groups. Such integrated systems could use innovative assessments as described in this paper where those assessments provide better information about relevant aspects of student learning.

## References

- Aschbacher, P. (1991). Performance Assessment: State Activity, Interest and Concerns. *Applied Measurement in Education*, 4(1), 275 – 288.
- Ben-Simon, A. & Bennett, R.E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment*, 6(1). Retrieved 11/30/09 from [www.jtla.org](http://www.jtla.org).
- Bennett, R.E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Princeton, NJ: Educational Testing Service.
- Bennett, R.E., & Bejar, I.I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, 17, 9-17.
- Bennett, R.E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, 6(9). Retrieved 11/30/09 from <http://www.jtla.org>.
- Bennett, R.E., Persky, H., Weiss, A.R., & Jenkins, F. (2007). *Problem Solving in Technology-Rich Environments: A report from the NAEP Technology-Based Assessment Project* (NCES 2007-466). Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- Blumberg, F., Epstein, M., MacDonald, W., and 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.
- Catterall, J., Mehrens, W., Flores, R. G., and Rubin, P. (1998, January). *The Kentucky Instructional Results Information System: A technical review*. Frankfort, KY: KY Legislative Research Commission.
- Christensen, C.M., Horn, M.B., & Johnson, C.W. (2008). *Disrupting class: How disruptive innovation will change the way the world learns*. New York City: McGraw-Hill.
- Clauser, Brian E.; Harik, Polina; Margolis, Melissa J.; Mee, Janet; Swygert, Kimberly; Rebbecchi, Thomas (2008). The Generalizability of Documentation Scores from the USMLE Step 2 Clinical Skills Examination. *Academic Medicine* 83(10), s41-s44. Retrieved 11/30/09 from [http://journals.lww.com/academicmedicine/Fulltext/2008/10001/The\\_Generalizability\\_of\\_Documentation\\_Scores\\_from.11.aspx](http://journals.lww.com/academicmedicine/Fulltext/2008/10001/The_Generalizability_of_Documentation_Scores_from.11.aspx)
- Cronin, J., Dahlin, M., Adkins, D., & Kingsbury, G. (2007). *The Proficiency Illusion*. Washington, DC: Thomas B. Fordham Institute.

- Darling-Hammond, L., & McCloskey, L. (2008). Assessment for learning around the world: What would it mean to be internationally competitive? *Phi Delta Kappan*, 90(4), 263-272.
- Fenster, M. (April, 1996) *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, April 8-12, 1996).
- 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 on December 15, 2009 from [http://www7.nationalacademies.org/bota/Workshop\\_1\\_Presentations.html](http://www7.nationalacademies.org/bota/Workshop_1_Presentations.html)
- 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 on December 15, 2009 from [http://www7.nationalacademies.org/bota/Workshop\\_1\\_Presentations.html](http://www7.nationalacademies.org/bota/Workshop_1_Presentations.html)
- Hambleton, R.K., Jaeger, R.M., Koretz, D. Linn, R.L., Millman, J., and 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., and Plake, B.S. (2000). Psychometric review of the Maryland School Performance Assessment Program (MSPAP). Psychometric Review Committee.
- Hamilton, L.S., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Hamilton, L.S., Stecher, B.M., Russell, J.L., Marsh, J.A., & Miles, J. (2008). Accountability and teaching practices: School-level actions and teacher responses. In B. Fuller, M.K. Henne, & E. Hannum (Eds.), *Strong state, weak schools: The benefits and dilemmas of centralized accountability (Research in the Sociology of Education, Vol. 16*, pp.31-66). St. Louis, MO: Emerald Group Publishing.
- Hoff, D. (2002) *Md. to Phase Out Innovative Program*. Education Week, April 3. Retrieved September 10, 2009 at: <http://www.edweek.org/ew/articles/2002/04/03/29mspap.h21.html>
- Kane, M.T., & Mroch, A.A. (in press). Modeling group differences in OLS and orthogonal regression: Implications for differential validity studies. *Applied Measurement in Education*.

- Kentucky Department of Education (2008). *Fact Sheet: Reconsidering Myths Surrounding Writing Instruction and Assessment in Kentucky*. Retrieved on Sept. 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. Accessed on Sept. 7, 2009. <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). *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, April 8-12, 1996.
- Klein, S.P. (2008). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. In D. Nolan & T. Speed (Eds.), *Probability and statistics: Essays in honor of David A. Freedman* (Vol. 2, pp.76-89). Beachwood, OH: Institute of Mathematical Statistics.
- Klein, S., Freedman, D., Shavelson, R., & Bolus, R. (2008). *Assessing school effectiveness*. Paper retrieved 11/30/09 from <http://www.stat.berkeley.edu/~census/finalER.pdf>
- Klein, S.P., McCaffrey, D., Stecher, B., and Koretz, D. (1995). The reliability of mathematics portfolio scores: Lessons from the Vermont experience. *Applied Measurement in Education*, 8(3),243-260.
- Koretz, D.M., and Barron, S.I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica: RAND.
- Koretz, D., Barron, S., Mitchell, M. and Stecher, B. (1996) *Perceived Effects of the Kentucky Instructional Results Information System (KIRIS)*. RAND Corporation.
- Koretz, D. Stecher, B., Klein, S. and 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., Cerrillo, T.L. and Hansen, M.A. (2000). *MSPAP Impact Study: Science*. U.S. Department of Education, Assessment Development and Evaluation Grants Program (CFDA 84.271) for the Maryland Assessment System Project.
- Lane, S., Parke, C.S., and 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.
- Maryland State Department of Education (1998). *Maryland School Performance Assessment Program (MSPAP) 1997. Technical Report* (ERIC ED 461 666)
- 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.
- Minnesota Department of Education (2008). *Minnesota Comprehensive Assessments Series II (MCA-II): Test specifications for science*. Roseville, MN: Author. Retrieved 11/30/09 from <http://education.state.mn.us/mdeprod/groups/Assessment/documents/Report/006366.pdf>
- Moe, T.M., & Chubb, J.E. (2009). *Liberating learning: Technology, politics, and the future of American education*. San Francisco: Jossey-Bass.
- 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 Research Council (2008). *Assessing accomplished teaching: Advanced-level certification programs*. Committee on Evaluation of Teacher Certification by the National Board for Professional Teaching Standards (M.D. Hakel, J.A. Koenig, and S.W. Elliott, Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.
- O'Reilly, T., & Sheehan, K.M. (2009). *Cognitively based assessment of, for, and as learning: A framework for assessing reading competency* (ETS report RR-09-26). Princeton, NJ: Educational Testing Service.
- Parke, C.S. and Lane, S. (2008). Examining alignment between state performance assessments and mathematics classroom activities. *Journal of Educational Research*, 101(3), 132-146.
- Partnership for 21<sup>st</sup> Century Skills (2008). *21<sup>st</sup> century skills, education, and competitiveness: A resource and policy guide*. Retrieved 11/30/09 from [http://www.21stcenturyskills.org/documents/21st\\_century\\_skills\\_education\\_and\\_competitiveness\\_guide.pdf](http://www.21stcenturyskills.org/documents/21st_century_skills_education_and_competitiveness_guide.pdf)

- 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.
- Pellegrino, J.W., Chudowsky, N., & Glaser, R., eds. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: Committee on the Foundations of Assessment, Board on Testing and Assessment, Center for Education, National Research Council, National Academies Press.
- Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments*. Dover, NH: National Center for the Improvement of Educational Assessment.
- Queary, P. *Senate passes WASL changes*. Seattle Times, March 5, 2004.
- Queensland Studies Authority (2009). *Student assessment regimes: Getting the balance right for Australia*. Draft paper retrieved 11/30/09 from [http://www.qsa.qld.edu.au/downloads/publications/ksa\\_paper\\_assess\\_balance\\_aust.pdf](http://www.qsa.qld.edu.au/downloads/publications/ksa_paper_assess_balance_aust.pdf)
- Quellmalz, E.S., & Pellegrino, J.W. (2009). Technology and testing. *Science*, 323, 75-79.
- 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 assessment: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer.
- Rohten, D., Carnoy, M., Chabran, M. and Elmore R. (2003). The Conditions and Characteristics of Assessment and Accountability. In *The New Accountability: High Schools and High Stakes Testing*. Carnoy, M, Elomore, R. and Siskin, L (Eds), Taylor & Francis Books, New York, New York.
- Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, 5(3), available at <http://epaa.asu.edu/epaa/v5n3.html> (retrieved 10/8/09).
- Sawchuk, S. (2009, November 13). Funding for common assessments poses challenge. *Education Week*, 29(12). Retrieved 11/30/09 from <http://www.edweek.org>.
- Schmidt, W.H., Wang, H.C, & McKnight, C. (2005). Curriculum coherence: An examination of U.S. mathematics and science content standards from an international perspective. *Journal of Curriculum Studies*, 37, 525-559.
- Shulte, B. *MSPAP Grading Shocked Teachers* Washington Post, February 4, 2002.
- Shepard, L., Hannaway, J., & Baker, E. (Eds., 2009). *Standards, assessments, and accountability: Education Policy White Paper*. Washington, DC: National Academy of Education.

- Stecher, B. M. and Chun, T. (2002). *School and Classroom Practices During Two Years of Education Reform in Washington States*. CSE Technical Report No. 550. Los Angeles: UCLA National Center for Research on Evaluation, Standards and Student Testing.
- Stecher, B. M., Chun, T., Barron, S. and Ross, K. (2000). *The effects of the Washington education reform on schools and classrooms: initial findings*. RAND, DB-309-EDU.
- Stecher, B.M., Epstein, S., Hamilton, L.S., Marsh, J.A., Robyn, A., McCombs, J.S., Russell, J.L., & Naftel, S. (2008). *Pain and gain: Implementing No Child Left Behind in California, Georgia, and Pennsylvania, 2004 to 2006*. Santa Monica, CA: RAND.
- Stecher, B. M. and 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. (1999). *Washington Assessment of Student Learning, Grade 4: 1998 Technical Report*. Olympia: Washington Office of the State Superintendent of Public Instruction. (ERIC ED 445 018)
- Tucker, B. (2009). *Beyond the bubble: Technology and the future of student assessment*. Washington, DC: Education Sector.
- United States Department of Education. (1995). *Mapping out the National Assessment of Title I: The Interim Report*. Section 2: Reform through Linking Title I to Challenging Academic Standards. Retrieved September 9, 2009 from <http://www.ed.gov/pubs/NatAssess/sec2.html>.
- United States Medical Licensing Examination (2009). *2010 Step 2 Clinical Skills (CS) content description and general information*. Retrieved 11/30/09 from <http://download.usmle.org/2010/2010CSinformationmanual.pdf>
- Wainer H., & Thissen D. (1993). Combining multiple-choice and constructed-response test scores: toward a Marxist theory of test construction. *Applied Measurement in Education*, 6,103-118.
- Washington State Institute for Public Policy (2006). *Tenth-Grade WASL Strands: Student Performance Varies Considerably Over Time* Olympia: Author.
- White, K. (1999) Kentucky: To A Different Drum. Quality Counts '99 Policy Update: Education Weekly on the Web. <http://rc-archive.edweek.org/sreports/qc99/states/policy/ky-up.htm>, Accessed Sept. 10, 2009.