Performance Assessment 2.0:
Lessons from Large-Scale Policy & Practice

EXECUTIVE SUMMARY

Ruth Chung Wei
Raymond L. Pecheone
Katherine L. Wilczak

In the last few years, there has been a growing recognition that state accountability systems are limited and often do not assess essential competencies such as higher order thinking skills. This interest corresponds with the establishment of a new policy environment in which the inadequacy of current assessment systems for supporting college and career readiness has been brought into sharper focus. In addition, the widespread adoption of new standards for college and career readiness – the Common Core State Standards – has provided the policy impetus for changing the way students and teachers are assessed.

A significant shift in direction is underway, representing a "swing of the pendulum" away from a decades-long dominance of standardized selected-response testing back towards the use of more diverse and richer forms of assessments.

Performance assessment taps into students' higher order thinking skills – such as evaluating the reliability of sources of information, synthesizing information to draw conclusions, or using deductive/inductive reasoning to solve a problem – to perform, create, or produce something with transferable real-world application. Researchers have found that the use of performance assessments can produce positive instructional changes in classrooms (Koretz et al., 1996; Matthews, 1995); increase student skill development (Spalding and Cummins, 1998); increase student engagement and post-secondary success (Foote, 2005); and strengthen complex conceptual understandings (Chung & Baker, 2003). Fundamentally, performance-based assessments

provide a means to assess higher order thinking skills while helping teachers and principals support students in developing a deeper understanding of content (Vogler, 2002).

During the 1990s, there were a number of large-scale experiments in performance assessment across the country. Despite the benefits of performance assessment documented in the research, many of the states that attempted to integrate performance assessments into their state assessment programs had to abandon the use of performance assessment for a variety of reasons. While some of these experiments were successful, and traces of these initiatives can still be found in existing state assessment programs (e.g., the Connecticut Mastery Tests/Connecticut Academic Performance Test, the New England Common Assessment Program-NECAP), in most cases, large-scale use of performance assessments was discontinued due to a variety of challenges to those systems.

We conducted a retrospective research study of performance assessment initiatives beginning in the 1990s up to today, drawing on available research literature and documentation produced by state assessment programs, as well as interviews with key individuals who participated in developing and administering those assessments, studied the implementation of those assessments, or have expertise in performance assessment. The study addresses three specific questions:

- What were the conditions that helped sustain some of the programs?
- What were the challenges that led to their discontinuation?
- What are some lessons learned that might help inform current assessment initiatives that seek to integrate performance assessment into large-scale student assessment programs?

The performance assessment systems that we examined included the following initiatives:

State	*Initiative Name*	*Years of Administration*
California	California Learning and Assessment System (CLAS)	1993 – 1994
Connecticut	Connecticut Mastery Test (CMT) Connecticut Academic Performance Test (CAPT)	1985 – present 1994 – present
Kentucky	Kentucky Instructional Results Information Systems (KIRIS)	1991 – 1998
Maryland	Maryland State Performance Assessment System (MSPAP)	1991 – 2002
Nebraska	Nebraska School-based Teacher-led Assessment and Reporting System (STARS)	2001 – 2009
Multiple States	New Standards Project (NSP)	1991 – 1999
Rhode Island	Rhode Island Diploma System	2011 – present
Vermont	Vermont Portfolio Assessment Program	1991 – 2004
Wyoming	Wyoming Body of Evidence (BOE)	2001 – present

Overview of Findings

Three kinds of lessons learned have emerged from our synthesis of the research.

1. Lessons about the role of political contexts and the importance of leadership, communication, and public support.
2. Lessons about technical quality and the design of performance assessment systems that support credibility and viability.
3. Lessons about practical issues such as cost and implementation factors that supported or hindered the success of performance assessment systems.

In our analysis, we draw parallels between these retrospective lessons gleaned from the performance assessment initiatives of the 1990s and conditions today (e.g., policy contexts, technical issues, and practical/implementation issues) to inform our understanding of current challenges and areas for opportunity.

What we find is that while many of the technical quality issues for integrating performance assessment into large-scale assessment systems may have been

overcome, there remain political, communication, and implementation challenges that will continue to serve as stumbling blocks to large-scale implementation and scale-up.

Based on our synthesis of the research, we also offer recommendations for the role that performance assessments should play in state assessment systems, for strategies that may support educative use of performance assessments, and for policies that may support the sustainability and viability of large-scale assessment systems that include performance assessments.

1. *Lessons about the role of political contexts and the importance of leadership, communication, and public support*

A crucial factor that either supported or led to the dismantling of large-scale performance assessment programs in the 1990s was the political context in which they were initiated, funded, developed, and implemented. We identified four major factors related to political context and leadership that shaped the outcomes of the programs:

a) **Shifting purposes for educational assessment.** As the policy environment in the U.S. moved toward greater levels of accountability for schools, teachers, and students, the role of educational assessment changed. The design, technical quality, and implementation costs of many of the assessment programs we studied (created during an era in which accountability focused on school-level scores) did not align with the demands of No Child Left Behind (NCLB) to test more grades and more students, and to report student-level scores by the autumn of each year. Only those programs adaptable to the demands of NCLB survived.

b) **Competing priorities and scarce resources.** Designing, implementing, and scoring performance tasks was typically more expensive than administering off-the-shelf basic skills tests. In almost all cases, the new assessments we studied received support initially through special funding streams or the infusion of new legislative appropriations. However, exhaustion of those initial funds, fluctuations in education budgets, and changes in political support led to the defunding of many of the programs.

c) **State politics and educational leadership.** Many of the assessment

programs we studied were initiated by those with significant political clout in the state policy arena, and could not be sustained without strong leadership and on-going legislative support. Unfortunately, with each political cycle and changing leadership, educational programs were vulnerable to shifting political winds. State assessment programs with more consistent political support and leadership were longer lived.

d) **Public acceptance and teacher and parent buy-in.** Due to a lack of understanding about the purposes and benefits of the new standards and assessments, they were often subject to criticism and skepticism by the public, and were often regarded as a burden by teachers despite their initial support. Some of the assessment programs were subjected to damaging media attacks that supported the efforts of vocal oppositional groups to dismantle these programs.

These political factors and contexts continue to be critical to the adoption of performance-based assessment formats in current large-scale assessment systems. While the widespread adoption of the Common Core State Standards initially made the policy environment more hospitable to performance assessment, we are beginning to see significant resistance to the CCSS from both the right and the left. In this highly charged political climate, the importance of leadership and an urgent need for improved communication to rally educator and public support for the CCSS and CCSS-aligned assessments is becoming more evident.

2. *Technical Quality Issues*

In a changing policy context in which school-level accountability was being significantly intensified, the performance-based assessment programs that were dismantled near the end of the 1990s and early 2000s had difficulty producing student-level scores that were defensible on technical grounds. There were four main technical quality issues related to the performance assessments of the 1990s:

a) **Use of matrix sampling and school-level reporting amidst increasing demands for student-level reporting.** Matrix sampling allowed for assessment of a broader range of content standards with greater efficiency and less testing time by administering different performance tasks to students across a school. However, it did not produce comparable student-level

scores. The demand for student-level score reporting across all testing grades could not be met feasibly in some assessment programs, which led those programs to be discontinued.

b) **Lack of standardization and comparability of performance assessments.** One of the problems with some of the performance assessment programs in the 1990s, particularly with portfolio assessments in which teachers designed their own assessments or selected from a task bank (e.g., Kentucky, Vermont), is that the assessments were not always comparable and were completed with the assistance of teachers, parents, or classmates, making it impossible to compare scores of one portfolio to another.

c) **Validity and content issues.** Some of the performance assessments in the 1990s were criticized for lacking clear measurement targets, for inconsistent results when compared with other measures (e.g., National Assessment of Educational Progress, ACT scores), and for including content with bias and sensitivity problems.

d) **Inter-rater reliability and insufficient item reliability.** All performance assessments must be hand-scored by trained scorers using professional judgment. Although sufficient inter-rater reliability was achieved after several years of implementation as scoring protocols were improved, reports of initially poor inter-rater reliability fed into a general skepticism about whether performance tasks can be reliable measures. Local scoring approaches, in particular, were problematic for high-stakes use. Additionally, performance tasks produce a small number of scores on a relatively limited content domain because it is impractical to administer multiple lengthy performance tasks to an individual student.

These technical issues continue to be important considerations in the design of large-scale assessment systems that are expected to be applied to high-stakes purposes. However, the previous limitations of performance assessment in the 1990s that led policymakers and the general public to question their validity, comparability, and reliability have been largely overcome. Today, the field of assessment development has evolved to include more systematic processes, protocols, and safeguards, so that assessment systems that include performance assessment formats can be designed to be comparable,

reliable, and valid measures of targeted learning outcomes. Use of assessment design frameworks, such as Evidence-Centered Design (Robert Mislevy), and task design and content specifications have improved the alignment between assessment design and measurement targets, allowing for greater comparability among performance tasks. Systematic bias/sensitivity review processes for ensuring item quality have also improved the overall quality of test items, and improvements in the design of scoring instruments, training protocols, and moderation processes during scoring have also improved inter-rater reliability and validated the use of hand scoring for large-scale and high-stakes use. The use of performance tasks in combination with other closed response types to measure overlapping measurement targets has also supported greater content validity without sacrificing reliability. These state-of-the-art practices are in use by the testing consortia that are designing and field-testing the Common Core assessments (Partnership for Assessment of Readiness for College and Careers-PARCC and Smarter Balanced Assessment Consortium-SBAC).

3. Practical Issues in Implementing Large-Scale Performance Assessments

A last set of important factors that we found to have an impact on efforts to integrate performance assessments into large-scale assessment systems in the 1990s were the practical issues related to implementing the assessment systems. Included in this set of factors are:

a) **Costs and burdens associated with developing, administering, and scoring performance assessments.** As noted previously, the cost of performance assessment in the 1990s was high relative to other assessment item types, and made it prohibitive to continue using performance assessments under the requirements of No Child Left Behind. NCLB dramatically increased the costs of testing across states due to the requirement to test in more grades, include more students, and report more quickly. State funding was insufficient to sustain the use of performance assessment in most states. Today, states have combined resources through testing consortia (e.g., NECAP, SBAC, PARCC), with the goal of reducing the cost of developing and administering the assessment.

b) **Pressure to quickly scale up and use the assessments for accountability.** It takes time for new assessments to be developed, piloted, field-tested, and refined to bring them to a level of technical quality requisite for high-stakes use. However, state

agencies are often pressured by policymakers to bring assessment programs online more quickly than is warranted due to low tolerance for an accountability vacuum. These pressures often led to sacrifices in quality, both in terms of the assessment items and the manner in which the assessments were implemented.

c) **Need for a coherent system of curriculum, instructional resources, and professional development.** Standards-based reform envisions a coherent system of standards, assessments, curriculum, and instruction. Unfortunately, in many cases, state policies and budgets did not prioritize such comprehensive approaches to instructional change. Instead, the focus was on creating systems of accountability, with little attention to the opportunities to learn needed by teachers and students. A single-minded focus on assessment as a lever for reform did not lead to wide-spread instructional improvement or sustained teacher and parent support.

In the current policy context, in which assessment-based accountability continues to be the main driver of school reform, along with the push to implement the Common Core State Standards, we continue to see the same pressures, resource trade-offs, and potential missteps in implementation. While cross-state collaborations provide a promising strategy for reducing the costs of developing and administering performance assessments, there remain technological and infrastructure roadblocks to smooth implementation. In addition, in rushing to build new assessment systems, policymakers at all levels often neglect a key underlying premise of standards based reform - the need for a coherent system of standards, assessment, curriculum, instructional resources, and professional development. While performance assessments offer the promise of encouraging more varied and deeper learning experiences for students, the performance assessment initiatives of the 1990s show that assessment alone is insufficient to drive large-scale, systematic improvements in instruction and curriculum. An effective CCSS implementation strategy must also make deep investments in supporting instructional change through the provision of curricular and instructional resources and professional learning opportunities for teachers.

Conditions for Sustainability

In our examination of the nine performance assessment initiatives included in this study, we noted that

a few of the initiatives had greater longevity than others. When initiatives did not last more than a few years (e.g., CLAS), this was usually due either to political or leadership changes, or the technical limitations of the assessment (i.e., matrix sampling, lack of comparability across assessments) that could not withstand the increased demands for assessment-based accountability. Those initiatives that lasted for a longer period of time (more than five years), such as the performance-based assessment programs in Kentucky, Maryland, Connecticut, and Wyoming, experienced success due to the continuity of political leadership within the state, the technical quality of the assessment, and the level of buy-in from teacher and other stakeholder groups.

One state in particular, Connecticut, stands out in terms of the longevity of its assessment system. While the Connecticut Mastery Tests and Connecticut Academic Performance Test have evolved over the last 25 years - with some of the on-demand classroom-based performance items being eliminated - the state has been able to sustain a high quality assessment that continues to incorporate performance-based items along with selected-response and short constructed-response items. In fact, it is likely because of the assessment design's balance of multiple item formats, and the program's willingness to adapt to changing policy frameworks toward increasing accountability, that it was able to survive the demands of NCLB. In combination with a technically defensible and balanced assessment approach, Connecticut has experienced a unique continuity of political and educational leadership over the years.

Lessons Learned and Recommendations

Based on our analysis of performance assessment initiatives of the 1990s, we propose the following seven key recommendations for future performance assessment initiatives. These recommendations focus on state and district actions that may support their transition to the Common Core State Standards and implementation of CCSS-aligned assessments.

1. Design assessments that meet intended purposes and meet standards of technical quality

One recurring issue evident in many of the performance assessment initiatives we studied is that the technical quality of performance tasks was not sufficiently robust. Lessons from Connecticut, Maryland, and other large-scale assessment programs that integrate the use of performance components suggest that it is possible to achieve sufficient levels of technical quality if developers design their assessments with the intended uses in mind, and invest in

processes designed to support technical quality.

2. Minimize the costs of hand scoring by involving teachers in scoring performance-based assessments

Hand scoring in the context of large-scale assessments is costly and time-intensive due to the need to recruit and train large cadres of scorers. Yet educator-involved scoring models have been used successfully and have supported the sustainability of performance based assessments (e.g., Nebraska STARS, New York State Regents[1], and Queensland, Australia[2]). Involving educators in scoring can help states minimize the cost of scoring performance assessments. And with robust training protocols and proper controls, educator-involved scoring can be technically sound, and support teachers' professional learning.

3. Minimize the cost of developing and administering performance assessments through economies of scale and cross-state collaboration

The costs of designing and managing assessment programs that included performance tasks led to the demise of many performance assessment initiatives of the 1990s. States that have adopted the CCSS should take advantage of the cost-saving benefits created through economies of scale, specifically those of the Common Core assessment consortia – SBAC and PARCC. In cost-benefit analyses, education agencies should also account for the benefits of using performance-based assessments that promote student use of higher-order cognitive strategies rather than a reliance on selected response items that restrict instruction by focusing on lower-order skills.

4. Build a coherent system of assessments, curricula, and instructional supports

As districts and states transition to the CCSS, they should invest in new kinds of formative assessment practices that include the development of curriculum-embedded performance tasks to evaluate *the full range* of the CCSS, and not just those expected to be measured on summative tests. Developing a comprehensive and coherent system of standards, assessment, and instruction to support rigorous learning should include the development of a) Curricular resources aligned to the desired state/local learning outcomes and assessment; b) Protocols and processes to quickly vet curricula, curriculum-embedded

[1] The New York State Regents has a rich history of local hand scoring that builds into a teacher's workload the resources and time for teachers to be trained and to score performance items on the Regents examinations.

[2] Queensland has a long tradition of implementing a tiered system of social moderation (scoring audit) of student performance assessments that are designed at the local level, peer reviewed and certified across all levels of the system (classroom, school and state level) by independent panels of trained teachers and educators.

assessments, and instructional modules; and c) Data reporting systems of student learning that are structured to include multiple sources of evidence about student learning in relation to the standards.

5. Invest in the development of a crowd-sourced clearinghouse of high quality CCSS-aligned performance tasks to support powerful instruction and assessment practices

Lessons learned from past experiences with performance-based assessment reveal that teachers and schools are oftentimes isolated and unsupported in their efforts to develop and implement richer curricula and assessments that support richer and deeper learning experiences for students. States that have adopted the CCSS should create a cross-state collaborative electronic platform to share resources, information, and best practices that comprehensively address and are indexed to the CCSS. The creation of digital libraries of formative assessments, curriculum resources, and instructional modules has the potential to move away from "one size fits all" approaches to formative assessment toward a system in which instructional leaders and teachers are expected to use their professional judgment and are provided with an array of choices about the design of a formative assessment system that both respects local contexts and better meets the learning needs of their particular students.

6. Actively engage with stakeholders, and develop the capacity of educational leaders and policymakers to deeply understand and champion research-based reforms

One of the enduring themes of successful large-scale use of performance assessment, highlighted in this monograph, is the critical role of communication and engagement with a wide spectrum of key stakeholders in the development and launching of innovative assessment systems. This can be accomplished by maintaining open channels of communication and transparency at all stages of the development process, keeping policymakers informed about the status of the work by actively engaging policymakers at all levels of the system in discussing the design and limitations of the assessment system, as well as highlighting significant areas of progress. Intensive engagement of educators and policymakers early on in the process should produce "champions" and supporters who step forward to advocate for the reform. Because of frequent changes in political leadership, states must also work to develop the organizational capacity of educational leaders at all levels of the system – state, district, and school – to sustain the reform as

new policies and priorities come and go.

7. Actively engage with the public, and provide timely, accessible information about the new assessment systems and the CCSS

Past movements to adopt performance assessment systems failed to build support among teachers, parents, and community members who often lacked any real understanding of why new assessments were adopted; what changes in instruction needed to be made in schools and classrooms to adapt to the assessments; why the new direction was necessary; how the new assessments differed from what already existed; and how the changes were better for students.

To sustain a state's adoption of a new assessment and accountability system, all key stakeholders must have a deep understanding of the standards and assessments as well as the curricular and instructional changes needed to achieve the new standards. Marshaling support for the Common Core State Standards and the assessment consortia (SBAC and PARCC) must move beyond simple claims that the standards are based on research and that high standards lead to more effective teaching and student learning. Instead, the public needs greater transparency about what will actually change with respect to curriculum, instruction, assessment, and student learning.