PERFORMANCE ASSESSMENT 2.0

Lessons from Large-Scale Policy & Practice

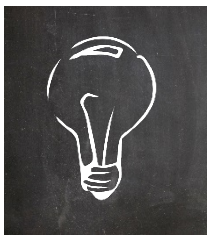
Ruth Chung Wei, rchung@stanford.edu
Raymond L. Pecheone, pecheone@stanford.edu
Katherine L. Wilczak, kwilczak@stanford.edu

Stanford | GRADUATE SCHOOL OF EDUCATION

SCALE

Stanford Center for Assessment, Learning, and Equity
Stanford University | December 2014

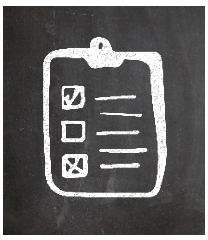
Table of Contents


CH 1


CH 2


CH 3


CH 4


CH 5

ACKNOWLEDGEMENTS

The authors would like to thank the following people for their contributions to this project: Janet Bagsby, Douglas Christensen, Scott Cook, Linda Darling-Hammond, Steve Ferrara, Pascal Forgione, Brian Gong, Milica Gurney, Edward Haertel, Joan Herman, Bill Honig, Kate Jamentz, Stuart Kahl, Michael Kirst, Daniel Koretz, Ann Lieberman, Claudia Long, Enrique Lopez, Robert Marzano, Lorraine McDonnell, Jay McTighe, Scott Marion, John Mergendoller, Marge Petit, Lauren Resnick, Douglas Rindone, Bob Rothman, Richard Shavelson, Lorrie Shepard, Elizabeth Stage, Matthew Tungate, and Grant Wiggins.

This report is made possible through a generous grant from the William and Flora Hewlett Foundation.

ABOUT SCALE

The Stanford Center for Assessment, Learning, and Equity (SCALE) is a research and development center based at Stanford University in California. SCALE provides technical consulting and support to schools, districts, and states that have committed to adopting performance-based assessment as part of a multiple-measures system to evaluate student learning and measure school and teacher effectiveness. SCALE works with education agencies and practitioners to develop customized assessment materials, establish and oversee scoring procedures, provide professional development to support teachers and schools engaged in the work, and conduct research to support the validity and reliability of the assessment system. At the core of our work is the belief that a performance assessment system should be educative for students, teachers, and schools.

EXECUTIVE SUMMARY

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

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.

INTRODUCTION

Purpose

A growing consensus among policymakers, educators, and the public is beginning to be forged that the current accountability systems used to assess student learning are limited, narrowly constructed, and significantly flawed. This growing dissatisfaction with standardized testing 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 and where policymakers are searching for richer and better alternatives. The 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. New standards require a paradigm shift in curriculum, instruction, and assessment.

This change in accountability is much more than just better "tests". 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, and 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, including a greater emphasis on cooperative work; a stronger focus on writing, problem solving, and real-world, hands-on activities; and a reduced emphasis on rote learning and teaching (Koretz et al., 1996). 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. 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 in the 1990s, tapping into the expertise of those with a deep reservoir of experience from these earlier efforts to integrate the use of performance assessments into large-scale assessment programs in the United States. The study documents and synthesizes the common political, technical, and practical issues related to these earlier efforts, with the goal of answering 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 goal of this study is to inform policymakers and educational leaders about the conditions necessary for successful integration of performance assessments into large-scale assessment programs. Lessons learned from past large-scale performance assessment initiatives can help states synchronize their policy and practices to align standards, assessment, curriculum, and instruction in ways that more effectively support student preparation for college and careers. This study also seeks to contribute to the knowledge base around effective and ineffective designs for performance assessments that have been used in large-scale assessment programs in terms of the actual scope and design of the assessments themselves, the ways in which the assessments were developed and implemented, and the policy frameworks under which the assessments operated.

Rationale

Since the adoption of the Common Core State Standards (CCSS) by a large majority of states[3] and the formation of two consortia of states – the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) – to design common assessments that measure the Common Core State Standards, the policy environment in the United States appears to be changing with regard to the possibility of integrating performance assessments into state assessment and accountability systems. The response from states and local districts to the Common Core State Standards has been mostly positive.[4] This is largely because the CCSS depart from traditional standards by focusing on college and career readiness expectations (that are research- and evidence-based) (Common Core State Standards Initiative, 2014), as well as the application of deep content understanding and higher order thinking skills. They have also been well-received because they have been viewed as opening up new opportunities and possibilities for the design of assessment systems that move beyond basic skills testing toward more complex and authentic forms of assessment.

This level of acceptance by state education agencies across the country of a common set of standards is unprecedented. Overwhelming endorsement of the intent of the standards by educators and state leaders, as well as the business and higher education communities, signals a growing awareness of the inadequacies of the current levels of student preparation for college and work, and a general dissatisfaction with current testing and accountability systems.

There is a general acknowledgement among educators and policymakers that existing testing programs are limited in their ability to assess the higher order thinking skills embedded in the CCSS, and are even more limited in assessing the application of these skills. Assessments of the full range of the new standards would require students to engage in more authentic investigations and tasks that address both the ability to apply knowledge and to communicate more effectively, both orally and in writing. A renewed focus on assessments that support inquiry and deeper learning has become a catalyst for the development of performance assessments to be included in the

[3] At this time, 45 states, the District of Columbia, and four U.S territories have adopted the CCSS.
[4] Recently, with a change of leadership in many states, states have become embroiled in local political battles to reject the Common Core State Standards. Indiana recently paused implementation of the Common Core, and Michigan legislators fought (unsuccessfully) to do the same. Other states, including Florida and Wisconsin, anticipate bringing the issue to a vote in their legislatures in the coming months.

design of new state and national assessments. Some performance assessments under consideration are short research tasks, mathematical problem-solving tasks, and document-based prompts that support literacy and text-based argumentation within and across disciplines. This interest in performance assessments that focus on the measurement of higher order thinking skills and application of knowledge to solve problems is not a new phenomenon; there are multiple examples of schools, districts, and states using performance assessments as a key driver to reform curriculum and instructional practices, particularly in the 1990s.

In the beginning of the 1990s, federal education policy was much more "loosely coupled" (Weick, 1976) than it is today, and states and local districts were given freer rein over their standards and testing programs. During that era, states and networks of states experimented with new and alternative forms of assessment. Over time, nearly all of these assessment programs and systems were dismantled or dramatically scaled back due to a variety of challenges to those systems, including changes in political leadership, concerns about the technical quality of performance assessments, and the costs of development and administration. Now in 2014, we are in an era of unprecedented federal control over assessment and accountability that few could have imagined during the previous era of "local control" and state designed and controlled accountability systems.

Interestingly, the regulatory framework set by the federal government over assessment and accountability (No Child Left Behind-NCLB) and the widespread adoption of the CCSS have converged to create a policy environment in which performance assessment has become an essential component in the design of a new accountability system.

The move toward performance assessment and the development of new policy frameworks supportive of the integration of the CCSS should be informed by lessons learned from the history of performance assessment initiatives in the 1990s. Some of the challenges that led to the demise of past performance assessment initiatives may have been overcome - e.g., the technical quality of the assessments (reliability and validity) in relation to their proposed use - while other issues continue to be challenges, such as the costs and burdens of administering and scaling up those assessments, the professional learning needs of teachers who must adopt new curricula and instructional strategies, and learning opportunities for students.

Methods and Data Sources

The research team gathered information about the design, conduct, and outcomes of performance assessment initiatives in the U.S. in the 1990s and in the following twenty years, including both those that have had some longevity as well as those that were quickly dismantled. The study includes a synthesis of research studies and other documentation on those initiatives, as well as the results of interviews with major players in these past initiatives.

Literature and Document Review

Through an extended literature search of available research papers and a solicitation of internal reports from the administrators of the performance assessment systems, we gathered key information about 1) the design of the assessment programs, including evidence about the technical quality of those assessments (reliability and validity); 2) the goals and policy framework of the performance assessments; and 3) the implementation of the assessment programs, including the cost of implementation.

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
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	2001 – present
Vermont	Vermont Portfolio Assessment Program	1991 – 2004
Wyoming	Wyoming Body of Evidence (BOE)	2001 - present

Interviews

For each selected performance assessment program, the research team conducted telephone or in-person interviews with multiple key personnel involved in designing, implementing, and/or building the policy framework of the program. These included past state education officers or agency leads, key designers of the assessment program, researchers involved in studying the technical quality and/or validity of the assessments, and assessment contractors. Other educational assessment experts and policy researchers who have studied these initiatives were also interviewed. In all, we conducted interviews with 30 individuals representing the selected performance assessment programs as well as other national assessment experts who studied or were familiar with those programs. (See Appendix A for a full list of interviewees.) These individuals were asked to fill in gaps in information about the assessment programs that were not accessible through literature searches and document reviews. In addition, these individuals were asked to describe their understanding of a) the most important challenges and impediments to integrating performance assessment into their assessment systems; b) the key conditions that supported effective implementation of performance assessments; and c) the most important lessons learned that may inform the current efforts of states with regard to the Common Core State Standards and the consortia assessments aligned to those standards.

Overview of Findings

Three kinds of lessons learned have emerged from our synthesis of the research. These will be described in much further detail and depth in each respective chapter.

The three kinds of lessons learned are:

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 1990s performance assessment initiatives 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.

CHAPTER 1

Political Context and Leadership Issues

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. Spending on public education is one of the largest expenditures for many states, and the power of the public education system to shape future generations of Americans is not one that is taken lightly. Consequently, efforts to control the content and form of public education, as well as how public funds are used, are understandably steeped in ideological and value conflicts. In this context, the use of assessment as a tool for accountability has become increasingly contentious over the last twenty years. The large-scale assessment programs that attempted to include performance assessment emerged because of a growing reliance on assessment as a policy tool to reshape American education and to hold education agencies at all levels accountable for student performance. At the same time, the assessment programs often became the casualties of the same political processes that helped bring them into being when demands for accountability shifted.

In our study of the performance assessment initiatives of the 1990s, we identified four major factors related to political context and leadership that shaped the outcomes of the programs:

1. Shifting purposes for educational assessment
2. Competing priorities and scarce resources
3. State politics and educational leadership
4. Public acceptance and teacher and parent buy-in

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 (CCSS) 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.

Shifting Purposes for Educational Assessment: The Move toward Greater Accountability

When performance assessment began to emerge in state assessment programs in the late 1980s and early 1990s, there was an increasing sense of urgency among national and state policymakers to improve the rigor of public education, as evidenced in reports like *A Nation at Risk* (1983). Prior to this time, there had been few attempts to use educational assessments as a metric for holding school organizations or personnel accountable. In the 1990s, the administrations of George H.W. Bush and Bill Clinton embraced the idea of comprehensive standards-based reform. As part of this reform effort, policymakers looked to standards as a means of bringing consistency and a common level of rigor to school curricula.

In 1994, the Clinton administration authorized the Goals 2000: Educate America Act, which asked states to voluntarily establish high state-level learning standards along with assessments that were aligned to those standards. The act also asked states to bring coherence to their curriculum, teacher preparation programs, textbooks, and in-service professional development. Congress funded the act with a $105 million appropriation for 1994, which provided incentives for states to develop a plan toward meeting the goals of the legislation. Goals 2000 also established a National Education Standards and Improvement Council to review and certify the voluntary state standards and assessment systems (Resnick, 1995). It was in this context of setting high "world class" academic standards and the establishment of the standards-based reform movement that assessment first emerged as a high-leverage strategy to hold schools and districts accountable for attending to a set of common standards for student performance.

States that introduced performance assessment into their assessment programs did so with the theory of action that including richer, more authentic types of work would lead to improvements in curriculum,

instruction, and student learning (Cohen and Hill, 1998; Ferrara, 2010; McDonnell, 2004; Simmons and Resnick, 1993; Stecher, 1998). Lorraine McDonnell, commenting on the California CLAS initiative, laid out the policy framework that drove CLAS: "The assessment would be linked to well-defined standards and curricula; the underlying curricular values – combined with the public notification and consequences associated with an accountability system – would prompt changes in teaching; and as a result, students would not only learn more effectively, but would also acquire knowledge of greater worth" (2004, p. 50). Along these lines, in Maryland, the new Maryland Learning Outcomes and MSPAP performance tasks were intended to move educators toward the teaching of higher order thinking skills and counteract the narrowing and "dumbing down" of curriculum and instruction (Michaels and Ferrara, 1999, p. 105). This may sound familiar because it is the same theory of action that undergirds the latest Common Core State Standards movement and consortia assessments.

In the 1990s, the local control culture and a strong belief in state primacy with regard to public school policy remained strong. Prior to 1980s, the majority of U.S. states did not have state education standards, and the development of national standards was unheard of. In many states that enacted state standards, especially those with a long history of local control, state demands for local school districts to adopt and implement state standards resulted in uneven implementation at best. However, in the last decade, federal mandates such as Title I and the Individuals with Disabilities Education Act (IDEA) have effectively been used as economic incentives to enact federal education policy in ways that effectively pushed forward the standards-based reform and accountability agenda for many districts that have come to rely on federal funding to supplement local education budgets.

With a growing demand for state, district, and school-level accountability (balanced by claims of local control and autonomy), politicians looked to assessments as one way of holding districts and schools responsible for meeting the educational needs of their students. At this time, there was no demand for student-level accountability (e.g., exit exams, retention/ remediation policies, or other promotion or graduation requirements) or for teacher-level accountability.

Because school-level accountability did not depend on every student receiving an individual score, testing could be conducted using a matrix sampling strategy. Matrix sampling is the practice of generating multiple forms of an assessment, each with a different set of items

(though some items overlap across forms), and administering these different forms to different students within classrooms and schools. This means that not all students take the same exam, and different items and performance tasks can be administered to different students. In the 1990s, matrix sampling allowed for a greater range of performance targets to be measured within a school, less testing time for individual students, less administrative burden for schools, less scoring at the state level, and greater efficiency in terms of cost. Along with its advantages, however, matrix sampling had some disadvantages: differences across demographic groups within a school could not be measured (all students need to take the same test to produce comparable scores); the school-level score results produced by matrix sampling were less transparent and more difficult to explain to parents and the public; and no comparable student-level scores could be generated because students were, in effect, taking different tests. Subsequently, performance data was valid only for a school or district as the unit of analysis.

During the late 1990s, there was an increasing demand for student-level scores among parents and policymakers (Koretz, Mitchell, Barron, and Keith, 1996; NRC, 2010), and by the time No Child Left Behind (NCLB) was enacted in 2001 (during the George W. Bush administration), state- district-, school-, *and* student-level accountability had become the focus of federal education policy[5], which re-shaped state-level accountability systems. In this environment of increased demand for individual student-level scores, it became impossible for states to continue using a matrix sampling approach in which tests administered to different students were not equivalent. In addition, the need to test every student, every year, in grades 3-8 and once in high school significantly increased the amount of testing administered by the state. Total testing costs and testing time increased markedly as states began to test every student across multiple years, as required by NCLB.

In this shifting policy context, the purpose of educational assessment also changed. Whereas test scores under the school-level accountability framework were used primarily to inform school leaders about the efficacy of local curricula and instruction and how to target

[5] NCLB Sec 1111(b)(3)(C) says "REQUIREMENTS- Such assessments shall-(xii) produce individual student interpretive, descriptive, and diagnostic reports, consistent with clause (iii) that allow parents, teachers, and principals to understand and address the specific academic needs of students, and include information regarding achievement on academic assessments aligned with State academic achievement standards, and that are provided to parents, teachers, and principals, as soon as is practicably possible after the assessment is given, in an understandable and uniform format, and to the extent practicable, in a language that parents can understand".

resources for the improvement of curricula, instruction, and student learning, in the new era of accountability, student- and teacher-level scores were used to hold schools and individuals, including students and teachers, accountable. In the last 10 years, there has been a growing push to hold teachers accountable (to reward "effective teaching" and penalize "less effective" teaching), with "effective teaching" being measured by teachers' contributions to student achievement through the use of standardized test scores and value-added modeling, a statistical method that estimates student learning gains in the areas tested during a particular year using prior achievement and sometimes other student characteristics as controls.

In addition, NCLB required schools, school districts, and states to report test results for specific student subgroups, including students from low-income families, students with disabilities, English language learners, and major racial and ethnic groups, with the aim to improve educational opportunities for these student groups.

In this context of rising demand for teacher and student accountability, performance assessment was edged out of most state assessment programs because of the cost and time associated with administration and scoring, as well as a general uneasiness about an assessment format deemed to be less "scientifically validated" than selected-response item formats.

Competing Priorities and Scarce Resources

Another factor that impacted the sustainability of performance-based assessments in state assessment programs in the 1990s was the availability of public and private funding for their development and implementation. In some cases, there were special state appropriations associated with legislation calling for the development of more rigorous state assessments. For example, Maryland increased education funding to 20% of the state budget to fund its comprehensive Maryland School Performance Plan (MSPP), of which performance assessment was a central component (Ferrara, 2010). Kentucky's performance assessment system, KIRIS, was part of a larger education reform plan that allocated nearly $700 million to public education over two years (McDonnell, 2004). Vermont's portfolio program was established in the midst of a $600 million state investment in education (Mills, 1996).

In the case of the New Standards Project, philanthropic funds from the Pew Charitable Trusts and the John D. and Catherine T. MacArthur Foundation were used in combination with the project's state membership dues to support development and piloting of New

Standards exams in numerous states (Simmons, 1993). Eventually, the foundation funds dried up, and when it began to appear that implementing the New Standards exams was not going to be profitable, the testing company that bought the rights to administer the exams shelved the performance tasks (L. Resnick, interview, June 14, 2012).

Both philanthropic resources and public funding for education are "soft money" – meaning that the funding fluctuates and almost always disappears. Both are subject to the booms and busts of the economy, especially at the state level, as well as changing political priorities. When initial seed money is expended, it is often impossible to sustain an expensive education program, especially when the program has insufficient political or public support or experiences issues of credibility.

During the 2000s, as the demands for accountability increased, the total costs of state testing rose substantially. As a result of No Child Left Behind, state testing costs went from an average of $8.4 million in 2001 to an average of $22 million in 2007-2008 (Vu, 2008).[6]

The federal government funded the increased costs of assessment with an initial 2002 investment of just $378 million[7] (USDOE, 2013). In both 2007 and 2008, the federal government appropriated $408 million for state assessments (USDOE, 2013), which works out to slightly more than $8 million per state, far below the $22 million average total testing cost per state. This meant that states had to reallocate funds from other state education priorities to meet new annual testing demands, and legislators felt pressured to eliminate higher-cost testing programs like those that incorporated performance-based items.

In comparison to state testing programs that exclusively use machine-scored selected-response items, programs that include extended constructed-response items and performance-based items are simply more expensive due to the cost of developing, administering, and hand scoring those types of items. "Typical" assessments (i.e., those with selected-response items only) had an average cost of $19.93 per student in 2010[8], while "high quality

[6] Total U.S. spending on standardized tests was almost $423 million in 2001; for the 2007-2008 school year it was almost $1.1 billion (Vu, 2008).
[7] Section 6113 of the No Child Left Behind Act of 2001 authorized $490 million to be appropriated for state assessments for fiscal year 2002 (NCLB, 2002), however the final 2002 federal budget included just $387 million in appropriations for state assessments (USDOE, 2013).
[8] Darling-Hammond and Adamson (2013) argue that the cost of "typical" assessments is actually much higher when the costs of test-prep, benchmark assessments, misdirected classroom instructional time, and other factors are included in estimates. They argue that the financial cost of implementing systems of performance assessment may actually be *lower* than the financial cost of traditional standardized

assessments" averaged $55.67 that same year (Topol, Olson, and Roeber, 2013). Studies show that, in past initiatives, the cost of scoring performance tasks and on-demand essays ranged from $1.50 to $15 per student (Stecher, 2010). Faced with increased requirements for testing, states made the difficult decision to scale back the proportion of performance-based items in their state assessment programs. For example, Connecticut, which legally challenged NCLB's requirements but was unsuccessful[9], reluctantly eliminated some of its expensive hands-on performance tasks in favor of more constructed-response items. In other states, like Maryland, performance-based item formats were eliminated altogether.

Once again, to fuel the redesign of state accountability systems, the federal government has harnessed the popularity of the CCSS and has established new funding streams to incentivize changes in state assessment systems. Currently, the two Common Core assessment consortia - the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC) - which both include performance-based items, operate on an infusion of federal dollars from the Obama administration's Race to the Top (RTTT) competition, which awarded the consortia $160 and $170 million, respectively, in 2010. The RTTT competition receives funding through the American Recovery and Reinvestment Act (2009). This federal stimulus package came at a time when the U.S. had descended into the worst economic recession since the Great Depression of the 1930s. With many states experiencing severe budget crises and the immobilization of government functions, these federal funds came as welcome relief. The RTTT competition also enticed states to join the two assessment consortia with the underlying theory that federal funding and the aggregation of state resources for a testing program would reduce the costs of developing and administering the assessment programs sponsored by the federal government.[10]

Like the assessment programs of the 1990s, the Common Core assessment consortia are funded through "soft money," and it is unclear whether the consortia can be sustained once current federal funds are used up or if the political

exams. Moreover, performance assessments support a system of deeper learning while traditional exams divert financial resources towards ineffective teaching and learning practices.

[9] In 2005, Connecticut sued the U.S. Department of Education, claiming that NCLB illegally required the state to spend millions of extra dollars on unnecessary tests. The case went all the way to the U.S. Supreme Court, where it was dismissed in 2011.

[10] Current average state spending for math and ELA summative assessments is approximately $20-25 per student; states range from $4.12 million to $114.46 million in annual summative NCLB testing (Topol, Olson, & Roeber, 2013, pp. i, 7).

winds change once again. State members of each consortium may pick up where federal funding leaves off to continue with development and implementation efforts, but the costs will be significant.[11]

Performance-based assessments are not only more costly to develop, administer, and score than traditional selected-response tests, they are also more expensive to support and implement. Teachers need professional development both to help them understand the nature and content of performance assessments and to effectively implement classroom-embedded performance tasks (e.g., the MSPAP and CAPT hands-on science labs). Yet state budgets rarely have the capacity to invest in the education and training of teachers. The performance assessment initiatives of the 1990s depended on the knowledge and skill of educators who used, and in some cases helped to develop, the performance-based assessments, however, many of the initiatives we studied failed to set aside resources to expand or deepen the professional learning of teachers at scale. This is not to say that states did not recognize the need for professional development. For example, in California, CLAS was preceded by a state-initiated "replacement units" project and the establishment of nine professional development networks (called "Subject Matter Projects") modeled after the National Writing Project. However, it appears that these prior investments in teacher learning were insufficient to help teachers adopt and integrate the California Content Frameworks along with the CLAS assessments that were designed to assess those new standards. In a survey of about 1,000 elementary mathematics teachers, Cohen and Hill (1998) found that although two-thirds of the respondents reported participating in professional development in one of five curricular areas, half of the respondents reported participating in only one day of professional development, and a little over one third reported participating in 2-6 days of professional development. They also found that only one-third of the teachers reported learning about CLAS and only one third had administered CLAS. Because the survey also included questions related to teachers' enacted classroom practices, the researchers were able to analyze teachers' reports about their professional learning opportunities in relation to their reported classroom practices. Findings (both from the survey

[11] In their initial plans, both consortia agreed to jointly investigate the technology of artificial intelligence (AI) scoring and/or automated scoring to reduce the cost of scoring performance-based items. Currently, only text output can be AI scored by computers that have been taught the characteristics of human-scored essays at different score levels. However, the inability of AI scoring to differentiate the quality of ideas in a text suggests that AI scoring will need further development before it can be used for high-stakes purposes (Markoff, 2013).

analysis as well as fieldwork) suggest that the amount and types of professional development available to teachers matter: "...when teachers' opportunities to learn from instructional policy are focused directly on student curriculum that exemplifies the policy, that learning is more likely to affect their practice" (p.14). Moreover, their findings revealed that learning about or administering the CLAS assessment contributed only modestly to more "reform-oriented" practice, underscoring the idea that an assessment alone cannot drive reforms in teacher practice.

In some states such as Vermont and Connecticut, where educators and stakeholders were initially involved in defining the new content standards, developing the assessments, and/or hand scoring the assessments, teachers were not only more likely to buy in to the new assessments, they were also more likely to understand how the assessments were aligned to the new standards and their implications for curriculum. This was most clearly evident in Connecticut, where teachers were initially involved in determining the goals of the Connecticut Mastery Test (CMT) as well as hand scoring the assessment's constructed-response and performance-based items. The CMT had broad teacher and community support; although there was some initial pushback from urban district superintendents about the challenges for urban schools, there was no organized opposition to the assessments (D. Rindone, interview, April 30, 2013). The low-stakes environment in which the CMT was initially administered allowed time for reflection, revision, and improvement on the part of the state agency as well as teachers (Barron, 1996). Herman, Aschbacher, and Winters (1992) argue that teacher involvement in scoring performance assessments is a valuable professional development experience that "can lead to a reprioritization of classroom goals" and "helps teachers come to a consensual definition of key aspects of student performance" (p. 82). Indeed, high school teachers in Connecticut reported changing their science curriculum to include more inquiry labs similar to those in the Connecticut Academic Performance Test (CAPT), and reported that student lab assignments improved over student work from the past (Kurz, 2001). (For a closer examination of Connecticut's assessment system, see page 77.)

In Kentucky, a study of teacher participation in KIRIS-related professional development and its impact on instructional practices found relatively high levels of participation among close to 400 survey respondents (Stecher, et al., 1998). Among mathematics teachers, 97 percent of teachers had participated in formal

professional development activities that year (1996-97) and their participation rates were similar for the previous two years. Mathematics teachers at both grades 5 and 8 indicated that the professional development helped prepare them for their mathematics teaching (e.g., how to use manipulatives to teach mathematics, how to teach mathematical communication), improved their ability to help students with their mathematics portfolios, and prepared them to administer KIRIS open-response items in mathematics. Teachers also reported on their curriculum coverage, their frequency of instructional practices used, and their orientation toward mathematics teaching. Results showed that teachers were allocating more time to mathematics instruction by integrating the subject with other subject areas, were regularly using both traditional and reform-oriented teaching strategies, and were using standards-based practices more frequently. Kentucky was able to support teachers in this way as a result of the large initial investment the state made in its KIRIS assessment system (McDonnell, 2004). (For a more information about KIRIS, see the case study on page 81.)

The provision of sufficient teacher learning opportunities, and by extension student learning opportunities, is an ongoing issue, as currently witnessed with the emergence of the Common Core aligned common assessments. While some states and districts have made resources available to help teachers understand the Common Core State Standards, to develop curricula aligned to the new standards, and to help teachers make the instructional shifts necessary to prepare students for the Common Core assessments, other states and districts have appeared to remain "on the fence" in a period of policy and leadership transition, perhaps waiting for new funding to kick in to support the assessment changes, or simply maintaining a level of skepticism regarding whether or not the CCSS and the common assessments will be adopted or used in their state. Although an expense to state and local systems, history teaches us that an investment in building teacher capacity to implement new standards, in developing and making available curriculum resources aligned to new standards, and in providing opportunities to participate in developing or scoring assessments aligned to new standards is essential to the success of assessment initiatives. After all, teachers want to prepare their students to do well on assessments – for the students' sake and their own – and students typically do their best when they have been prepared in the kinds of skills the assessments measure. And perhaps not surprisingly, almost every

person that we interviewed regarding lessons learned from past performance assessment initiatives emphasized the importance of supporting the professional learning of teachers as a means to support instructional change.

State Politics and Educational Leadership

Strong leadership and political initiative are necessary whenever public funding and state appropriations for educational programs are part of a reform effort. In all of the state performance assessment initiatives we studied, strong political leaders were requisite to help spur legislative action, generate public support, and procure public funding for the work. Likewise, powerful political leaders also have the power and clout to dismantle performance assessment systems.

In the case of California, it was Democratic State Senator Gary Hart and State Superintendent of Public Instruction Bill Honig who championed the CLAS program, and it was Republican Governor Pete Wilson who ultimately moved away from the CLAS assessment. Wilson had initially supported CLAS based on the understanding that individual student scores would provide a tool for holding teachers accountable. However, when the program failed to produce individual scores in its first two years, and support from various sectors of the public (e.g., some conservative groups) for the program was lacking, Wilson withdrew his support (Kirst and Mazzeo, 1996; McDonnell, 2004).

In Connecticut, Commissioner of Education Gerald Tirozzi and Democratic Governor William O'Neill supported the establishment of a Commission on Equality and Excellence in Education with a $20 million commitment to an education trust fund that would support the development of the Connecticut Mastery Tests (Wilson, 2001). This leadership and support for the state's testing program was sustained by Tirozzi's successors Vincent Ferradino and Theodore Sergi, and lent stability to the state's education programs over time.

Kentucky Democratic Governor Wallace Wilkinson played a large role in sparking general education reform in Kentucky in the late 1980s, setting the stage for KIRIS' introduction. Similarly, Maryland Democratic Governor William Donald Schaefer established the education commission that ultimately led to the creation of MSPAP. Vermont Commissioner of Education Richard Mills and Ross Brewer (Director of Policy and Planning for the Vermont Department of Education) were the driving forces behind the development of the Vermont Portfolio assessment, though the program was also well supported by the governor, state and local school boards, and educator groups

(Koretz, McCaffrey, Klein, Bell, and Stecher, 1992; M. Petit, interview, April 22, 2013).

These leaders' actions initiated changes, but their vision for education was often not sustained due to subsequent changes in leadership or federal policy. In some cases, new governors or state superintendents, representing the opposing party of those they succeeded, dismantled performance-based assessment programs based on ideological or purely political grounds, determined to leave their own imprint on the state's education program. For example, in Kentucky, a Republican takeover of the state senate in 1998 destabilized support for KIRIS, which had become a political bargaining chip (B. Gong, interview, June 8, 2012; NRC, 2010). In Wyoming, there is an ongoing stalemate with the current Republican Superintendent of Public Instruction, who has refused to implement state education laws with which her party disagrees (Celock, 2013; Curtis, 2013). Continuity and discontinuity in state leadership were clearly important factors in whether the state assessment programs of the 1990s were sustained. These were factors that were often out of the realm of control of those running the state assessment programs, and that subsequently resulted in the dismantling of many of the assessment programs we studied.

Other important political actors included courts and legislators. In a few cases, legal action filed against the state (e.g., charging that state funding of education was inequitable), resulted in an influx of funding and jump-started the efforts of state legislators and state education agencies to revamp their education standards and assessment programs. This was true in Kentucky (1989 court case *Rose v. Council for Better Education*), which resulted in the Kentucky Education Reform Act of 1990 and an influx of close to $700 million for public education over two years. Similarly, in 1995 the Wyoming Supreme Court decided that school funding should be allocated in a way that ensures that all students receive equal educational opportunities. This court decision, along with the mobilization of business groups pushing for improvements in the state's education system, was a catalyst for the development of the Wyoming Content and Performance Standards (1998), the Wyoming Comprehensive Assessment System (1999), and the Wyoming Body of Evidence system (2000) (CPRE, 2000; Marion, 1998, 2001).

In many of the cases we studied, state legislators also had a strong role in determining the longevity of a state assessment program. In Nebraska, a lack of understanding of the system by new legislators and concerns of some incumbent legislators about not having a

system for ranking schools on performance, coupled with federal demands for a single accountability system for all schools, contributed to the demise of the Nebraska School-based Teacher-led Assessment and Reporting System (STARS) (D. Christensen, interview, July 26, 2013). Additionally, in Vermont, Kentucky, and Maryland, state legislators withdrew their support of assessment programs when control of state legislatures changed party hands, when reports that critiqued the assessments' technical quality damaged the reputation of the assessment programs, or when it became clear that the systems in their current form could not meet the demands of NCLB.

Public Acceptance and Teacher and Parent Buy-in

A last, but crucial, condition needed to sustain an assessment program is public acceptance and support for the initiative. While educational leaders and many teachers are generally supportive of the idea of including performance-based items in state assessments, oftentimes few parents or members of the public understand the implications of changing the content and nature of an assessment. Some state departments of education sought to engage the public (including educators and parents) in the process of building their new assessment programs or content frameworks (e.g., Vermont, Maryland, Connecticut), but in the remaining cases these efforts were sporadic and insufficient. Because performance assessments of the 1990s were generally more rigorous and difficult than selected-response tests, and were also generally unfamiliar to students, students' performance on the new assessments often took a substantial dip initially. Subsequently, student growth increased following a transition period in which teachers aligned their curriculum and instruction to the learning outcomes of the new assessments. However, the initial dip in scores came as a shock to many parents, especially those who were accustomed to seeing their children attain a certain level of performance on standardized assessments. Without public involvement in the assessment development process and active communication strategies, parents often misunderstood the lower scores, resulting in a backlash against the assessment. This occurred in Maryland in 2001, when school performance scores declined unexpectedly, in some cases by 10 percent or more. This led MSPAP opponents to further question the validity of the MSPAP scores, and it gave challengers momentum that eventually led to the elimination of MSPAP (Ferrara, 2010; Hettinger, 2002; Reilly, 2002).

It is certainly possible that there may be a similar backlash against the Common Core assessments

currently being developed and field tested, especially if and when the first reports of students' scores show depressed performance levels compared to the high proficiency levels achieved on current state assessments. Already, there has been some public backlash against new Common Core aligned state assessments in New York, where parents have voiced concerns about the amount of test-preparation time their children experience, the difficulty and poor quality of items, and the low scores anticipated on the new assessments (Kramer, 2013; Matthews, 2013).[12] Lack of clear, ongoing communication to the public about the rigor[13] of the new Common Core assessments and their impact on student learning is likely to lead to misunderstandings of the purposes behind the assessment, finger pointing, and the weakening of public support for the assessments.

State assessment programs in the 1990s sometimes suffered from a lack of strategic communication efforts designed to educate the public in a clear and accessible way about the nature and content of the new assessments and the rationale for adopting a new assessment system. CLAS, in particular, faced fierce criticism of its content and innovative format. A few of the assessment's released performance items garnered negative public attention from conservative groups that characterized CLAS as a "warm-and-fuzzy exercise of self-expression," claimed CLAS items violated students' privacy, that they were biased and insensitive, and accused the assessment of lacking attention to basic content knowledge and skills (Hanson, 1994). The California Department of Education (CDE) did not help matters by initially refusing to publicly release items for review so that they could be vetted for quality.[14] This lack of transparency led many to question the assessment and provided opposition groups with additional fuel for debate (Hanson, 1994). The California Content Frameworks were also met with criticism from the public, who charged they were unclear and lacked rigor (meaning, a focus on basic skills and content) (McDonnell, 2004).

[12] In fact, the drop in the levels of student proficiency on the new New York tests was significant, as anticipated. The percentage of students in grades 3-8 who met the "proficient" benchmark on the new English language arts test fell from 55.1 percent in 2012 (the old test) to 31.1 percent in 2013 (the new Common Core aligned test). Similarly, in math, the proficiency rate fell from 64.8 percent in 2012 to 31 percent in 2013 (Ujifusa, 2013).

[13] A 2013 analysis by CRESST (National Center for Research on Evaluation, Standards, and Student Testing) indicates that the PARCC and SBAC assessments will reflect significantly higher depth of knowledge (DOK) levels than current states assessments, particularly the performance task components of the SBAC assessment (Herman & Linn, 2013).

[14] The CDE initially chose not release many performance items because of the high cost of developing them. The intention was to re-use the items in subsequent years, which would minimize the cost of refreshing the task bank year after year.

Likewise, there was some controversy over the content of Kentucky's KIRIS assessments and standards. Kentucky's curricular frameworks (developed at the same time as KIRIS) were initially vague, and faced many revisions throughout the KIRIS years (Koretz and Barron, 1998). Teachers reported difficulty in preparing students to take KIRIS because of the wide focus of the content standards (some researchers reported that this led to "rubric-driven instruction") (NRC, 2010; Stecher, 1998). As in California, KIRIS was attacked by a small faction of conservative activists who opposed the assessment's focus on critical thinking, its apparent neglect of "the basics," and the general expansion of the state role in education (McDonnell, 2004). However, Kentucky leaders were better able to quell the debate by engaging the public. The governor and key legislators met with opposition groups, and adults were invited to view KIRIS assessments after signing a non-disclosure agreement (McDonnell, interview, October 8, 2013).

Inherent in the American testing culture has been an underlying belief by the American public that machine-scored, closed-response item formats in which there is only one correct answer (selected-response, true/false, matching) are the most reliable and trustworthy source of evidence about student learning. The public's faith in these types of tests has been continuously bolstered by the field of psychometrics, which has developed a science of assessment based largely on closed-response items and less so on performance items. Public faith in closed-response, machine-scored tests as scientifically valid and reliable measures has also been supported by policymakers and econometric analysts who have harnessed the results of machine-scored tests in high-stakes ways that suggest that they are measures to be trusted (e.g., to measure "teacher effectiveness" through value-added modeling), despite the warnings of measurement experts and scholars about the limitations of these methods. This dominance of selected-response tests is unheard of in other parts of the developed world (e.g., Europe, Asia, Australia) where constructed-response formats and performance items are the norm. Additionally, there has been a long-held general skepticism among the American public and the education policy community about the reliability of test questions in which the quality of a response is judged by a human rater, regardless of the training that raters undergo to ensure adequate inter-rater reliability.[15] Incorporating

[15] There has been some movement since the 1990s to a greater level of acceptance of performance-based assessment formats, such as the use of essay formats in the SAT and Advanced Placement Exams,

performance assessment into state assessment programs, then, is a counter-cultural move that puts the burden of proof for the value of performance assessment squarely in the hands of assessment policy advocates to communicate and forge a consensus around the validity and educational benefits of including performance assessment items in large-scale summative assessment systems.

During the 1990s and early 2000s, this trust in machine-scored tests, along with budgetary constraints, the push for increasing levels of accountability tied to assessment programs, and frequent changes in state leadership, made it politically inviting for legislatures and policy leaders to portray performance-based assessments as unscientific, unreliable, cost-intensive, and inadequate for the new purposes they were expected to fill.

*

The two assessment consortia that have been building common tests that measure the Common Core State Standards (CCSS) have been capitalizing on the initial state support of the CCSS, and have signaled an incremental move toward more performance-based assessments. The item formats used in both the PARCC and SBAC assessments include selected-response, short constructed-response, and technology-enhanced items, as well as essay formats. Both assessment consortia plan to hand score most of the constructed-response and essay items due to the current limitations of AI (artificial intelligence) scoring tools. Because these assessments represent modest and incremental use of performance-based items, and use design methodologies that are likely to strengthen the overall psychometric quality and reliability of the assessments[16], there will be, perhaps, less of a chance of a backlash from the public and policy-makers based solely on misunderstandings about the technical quality of the assessments. (Although, if real problems with the technical quality of the assessments surface, there is certain to be a

portfolios in A.P. Studio Art, and even the use of interview formats in professional board examinations. Today, the use of essays and other types of performance formats that are inherent to particular professions-such as the use of teaching videos for the professional licensure of teachers or the use of interviews for medical board examinations-are accepted as more valid evidence of performance than tests that are comprised of machine-scored items alone, despite the use of scoring rubrics or other hand-scoring methods to produce scores. The strong branding associated with the College Board examinations (SAT, AP Exams, Achievement Exams), many of which include hand-scored components, also signal that there is a growing acceptance of such formats within the higher education field, and that hand scoring can be seen as a reliable way of evaluating such performances.

[16] For example, the use of evidence-centered design methodologies (see Mislevy, R.J., Almond, R.G., & Lukas, J.F. (2003). A brief introduction to evidence-centered design. Educational Testing Service (ETS), Research Report RR-03-16.

reaction from the public and policy-makers.)

Instead, most of the backlash so far has been related to a growing movement against federal control over education in states where political changes have led to a destabilization of support for the Common Core State Standards. This backlash against the Common Core has led several states to consider withdrawing their adoption of the standards (Michigan and Indiana, for example, both unsuccessfully pursued legislation rejecting the standards). Much of this backlash can be attributed to the perennial "states' rights vs. federal control" debate rather than a specific backlash against the content of the Common Core, which has been endorsed by educators, business groups, and higher education groups, among others, and was written in collaboration with teachers, researchers, and community groups. As Lorraine McDonnell and Stephen Weatherford note, the largest challenge for those promoting the CCSS "was to dismantle one of the most deeply entrenched and strongest policy regimes in US education: the tradition of each state and its local districts deciding separately what students should be taught" (McDonnell & Weatherford, 2013a, pg. 11). For the most part, the debate over the Common Core has focused less on what is actually *in* the standards and more on the undemocratic process by which the standards were developed (Ravitch, 2013a), the fact that the standards were developed through the use of federal dollars and required for eligibility for the federal Race to the Top awards, thereby expanding the federal role in education policy, and the fact that the process was spearheaded by the current Democratic presidential administration.[17] There is also growing opposition from the left and progressive groups as well, as noted by Diane Ravitch (2013b), because of the top-down accountability and teacher effectiveness policy framework in which the CCSS is being implemented.

Still, the SBAC and PARCC assessments are a significant improvement over the majority of the currently administered statewide tests for accountability. Both assessments have been found to promote higher order thinking and essential skills, "particularly those related to mastering and being able to apply core academic content and cognitive strategies related to complex thinking, communication, and problem

[17] Opponents of the CCSS include: CATO, the Pioneer Institute, Hoosiers Against Common Core, the Tennessee Eagle Forum, the Republican National Committee, and Tea Party groups. Supporters of the CCSS include: NGA, CCSSO, Achieve, AFT, Campaign for High School Equity, National Association of State Boards of Education, National Parent Teacher Association, Bill and Melinda Gates Foundation (McDonnell & Weatherford, 2013b)

solving" (Herman & Linn, 2013, p. 4). Given the current weakening of the CCSS movement across states, it is now all the more important that state education leaders and the assessment consortia mobilize to strategically communicate and clarify the content and purpose of the CCSS, argue for the strengths and technical quality of the common assessments, and begin to educate and rally the support of the wider public – including parents, teachers, business groups, and especially state legislators and state political candidates, who seem to have the most control over whether an assessment program can survive within and across states.

CHAPTER 2

Technical Quality Issues

A second consistent finding from our study of the performance-based assessment initiatives of the 1990s was that the technical quality of the assessments was generally insufficient for the purposes they were intended to meet. Large-scale assessment programs that use results for accountability purposes are held to a higher standard of technical quality than those in which the results are used primarily for formative purposes. In a changing policy context in which school-level accountability was being significantly intensified and individual scores for students were expected, 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 both defensible and comparable on technical grounds. There was one exception – the case of Connecticut (see the case study on page 77 for more information) – where the assessment program's technical quality and feasibility were sufficient to meet No Child Left Behind's (NCLB) demands for testing all students at nearly all grade levels, with greater stakes associated with assessment results.

There are four main technical quality issues related to performance assessments:

1. Use of matrix sampling and school-level reporting amidst increasing demands for student-level reporting
2. Lack of standardization and comparability of performance assessments
3. Validity and content issues
4. Inter-rater reliability and insufficient item reliability

These technical issues continue to be important considerations in the design of large-scale assessment systems with 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. The science of performance assessment development has seen significant advancements, though these are not well documented in the body of psychometric research, which has focused predominantly on closed response item formats (i.e., selected response and machine-scored items). There is an ongoing need for the publication and dissemination of psychometric research on innovation in assessment development to strengthen the body of evidence that will continue to advance the field. There is much to be learned from current efforts underway by assessment developers (SBAC, PARCC, and other assessment consortia) to develop innovative assessments that measure the higher order thinking skills embedded in the CCSS, as well as assessments of English Language Learners and students with disabilities. However, there is also much to be learned from the performance assessment initiatives of the 1990s and their technical quality limitations to inform current assessment development efforts.

Use of Matrix Sampling and School-level Reporting Amidst Increasing Demands for Student-level Reporting

Many of the performance assessment initiatives of the 1990s used a matrix sampling strategy when administering on-demand or curriculum-embedded performance assessments across schools within a state. These include the CLAS (California), MSPAP (Maryland), KIRIS (Kentucky) and NSP (New Standards Project) exams. A factor that shaped the decision to use matrix sampling was an economic one – it was very expensive to administer complex performance assessments and to hand score responses for every student in a state, especially in a large state like California. Another factor is that these states did not have accountability policies at that time that called for student-level or teacher-level reporting. While it had some disadvantages, matrix sampling was a cost-efficient and technically sound practice that allowed for sampling across a content domain to measure the full range of learning outcomes within a school or district. (Matrix sampling continues to be used by large-scale testing programs to pilot new items and to use different test forms that support greater test security.) Matrix sampling allowed for school-level reporting in a policy context in

which schools were the focus of accountability policies, but by the early 2000s, with the authorization of NCLB, student-level reporting on state tests had become a policy imperative.

In the cases of Maryland's MSPAP, Kentucky's KIRIS (which eventually eliminated its on-demand performance tasks), and the New Standards Project, matrix sampling was used to distribute performance tasks across students within schools. In those cases, matrix sampling was acceptable during the early years of the programs because many state accountability systems were based only on school-level reporting and accountability during the 1990s. However, when NCLB was passed under the Bush administration, this spelled the death knell for all state assessment programs that did not produce student-level results. NCLB required testing at grades 3-8 and 11 in reading and mathematics in the spring of each year, required testing of all students (including special education and English learners), and also required student-level scores to be reported in the summer. Developing, administering, and scoring performance-based assessments at each of the required grade levels would have been too costly and administratively unmanageable for most states, not to mention the technical issues that plagued scoring performance assessment items. Taken together, these issues provided multiple reasons for state policymakers to discontinue support for their performance-based assessment programs.

In California, the legislation authorizing the CLAS program (Senate Bill 662, 1991) called for student-level reporting. However, as an independent expert review panel (the "Select Committee") revealed after the first year of CLAS implementation, there were multiple problems with the CLAS matrix sampling plan (Cronbach, Bradburn, and Horvitz, 1994). Not only did the matrix sampling plan result in no student-level scores being generated, but the insufficient sample sizes of student exams being scored within schools resulted in unacceptably high standard errors even for school-level reports. These high standard errors were the result of both sampling errors and measurement errors. The technical insufficiency of the data produced was compounded by the fact that the state did not have the budget to score every student test (which it knew in advance but failed to inform the public) and other administrative problems that resulted in further reductions in sample size at some schools (problems such as lost test booklets and a lack of bar-coding on some exams linking students to schools). The Select Committee suggested that these problems could have been prevented had the administration, scoring, and score reporting been coordinated by a

single experienced testing contractor rather than the state education agency, which had limited staff and experience managing the logistics of large-scale assessment programs. The committee also suggested numerous ways to reduce standard errors, including increasing the number of exams scored per school, expanding the number of test forms used in a single school, making the test forms more comparable, adding more time for administration so that more tasks could be administered per student, improving the consistency of scoring, and determining a sampling strategy based on a target maximum standard error, among others. Many of these suggestions were adopted by the California Department of Education for the 1994 CLAS, but by then it was too late to fight the rising tide of negative perceptions of the CLAS program. In 1994, Governor Pete Wilson vetoed legislation that would have authorized the renewal of funding for CLAS, citing its inability to produce credible student-level scores (McDonnell, 2004; Stage, 2007).

Only in small states like Connecticut, with a relatively small population of students, was it feasible to administer and score assessments that included performance items *and* produced technically sound student-level scores. Even in Connecticut, however, changes had to be made to meet NCLB requirements. For example, in 2007, Connecticut discontinued its use of science labs that were curriculum-embedded hands-on performance tasks, though its high school exam (CAPT) continues to include five constructed-response items that are linked to science labs completed in 9th- and 10th-grade science classes. Connecticut's assessments were able to survive NCLB because of adjustments made to the assessment design, and also because their original design produced student-level scores. (For more information about Connecticut's assessment system, see page 77.)

Lack of Standardization and Comparability of Performance Assessments

The advent of NCLB also led to the demise of programs in which performance assessments used across classes and schools were not comparable, even when student-level scores were produced. Hand in hand with the demand for student-level results, a related problem that made performance assessment vulnerable was the lack of standardization in performance assessment design, implementation, and resources within some assessment systems. This was particularly true of portfolio assessment systems of the 1990s.

A common critique of portfolio systems was that it was unclear whether students were being held to the same standard when the

content, quality, and difficulty of assignments included in the portfolios could not be said to be comparable. On top of that, much of the student work that was entered into the portfolios was revised and polished, the result of peer assistance and teacher feedback. The question of "whose work is it?" became an issue in an era of increasing accountability in which student scores reflecting unassisted performance were the target of measurement. Student work that was completed with peers or at home with parental assistance became suspect as a trustworthy source of evidence of student learning.

Vermont, in particular, had a portfolio assessment program driven by local teachers designing their own writing and mathematics assignments that were later scored using common (not grade-specific) rubrics (Stecher, 1998). While the state department of education engaged in a concerted effort to build the capacity of teachers to design these assignments, and also provided a teacher-developed bank of assignments, there was little standardization in the assignments that comprised student portfolios (Koretz, et al., 1992). A major reason for this was that Vermont has a long history and culture of independence, local control, and teacher professional autonomy. It would have been seen as an infringement on teachers' professional practice and local autonomy to require a standardized set of writing or mathematics assignments. More importantly, the context in which the Vermont Portfolio Assessment Program was developed and implemented was a low-stakes accountability context. The primary intent of the assessment system was to spur improvements in instruction, not to hold schools and teachers accountable for individual student results (S. Kahl, interview, April 25, 2013; M. Petit, interview, April 22, 2013). And while the Vermont portfolio was also designed to provide achievement data to compare schools and districts, there were no rewards or sanctions tied to performance (NRC, 2010). Because Vermont's portfolio system was not designed for high-stakes purposes, it could not meet the technical demands of NCLB.

After a 13-year run, state legislators ended the Vermont Portfolio Assessment Program in 2004 and joined NECAP (New England Common Assessment Program) to meet federal testing and accountability requirements. Indicative of its important role in supporting instruction, the Vermont portfolio lives on in many schools as a local assessment, and professional networks of teacher leaders that grew out of the portfolio program exist to this day. Deep investments in professional development supported teacher buy-in to the Portfolio Assessment Program (Grant Wiggins, interview, June 11,

2012), and, overall, teachers felt the portfolio was a "worthwhile burden" (Stecher, 1998). Vermont teachers reported an increase in cross-disciplinary work and better alignment of curriculum with standards (Tung and Stazesky, 2010). Teachers felt that the portfolio assessment informed their instructional practice and that it transformed what and how they taught (S. Kahl, interview, April 25, 2013; M. Petit, interview, April 22, 2013; Stecher & Mitchell, 1995).

Similarly, in Kentucky, the writing and mathematics portfolios included assignments that were teacher designed, even while the body of work was scored using a common rubric. In the Kentucky mathematics and writing portfolios, diverse assignments across teachers and schools, completion of assignments as curriculum-embedded tasks, and large variation in portfolio practices presented challenges in determining the assignments' level of difficulty, the extent to which students' work was independently produced, and the comparability of scores (Borko, 1999; Koretz et al., 1996, 1998; Stecher, 1998).

In contrast to Vermont's portfolios, Kentucky's portfolios were developed as one part of a high-stakes "primarily performance-based" assessment system designed to monitor student performance at the school level (Gong, 1996; Wolf, 2000). The Kentucky Instructional Results Information System (KIRIS) included multiple measures: on-demand selected response items (included and excluded in various years), performance-based items, and a year-long portfolio in mathematics and writing at different grades. (See the case study of KIRIS on page 81 for more information.) KIRIS was a high-stakes assessment system; schools faced rewards and sanctions based on performance. The goals of KIRIS were to induce reform and improve instruction, monitor school performance, and serve as a basis of accountability (Koretz et al., 1996; Koretz, 1998). The system also hoped to encourage good teaching practices and the use of performance assessments (Gong, 1996).

KIRIS, however, faced many technical challenges, including variability across portfolio tasks, the lack of comparability between portfolio scores and on-demand essay scores, and the inability to make year-to-year comparisons of the on-demand performance task scores (S. Kahl, interview, April 25, 2013; Koretz, 1998; McDonnell, 2004; Tung & Stazesky, 2010). By 1995, a panel convened by the Kentucky General Assembly concluded that portfolio scores were not appropriate for use in the KIRIS high-stakes accountability system (Koretz, 1998). Based on improved scoring accuracy in the second cycle of testing, however, the program's technical advisory

committee disagreed with the panel, at least with respect to the writing portfolios. Thus, the writing portfolio continued to be used until 2008, even after KIRIS was discontinued in 1998.

Even Maryland's MSPAP, which met standards of technical quality at an acceptable level, had some issues related to standardization (NRC, 2010). MSPAP was composed exclusively of performance assessments and required a week of testing time (spanning a total of about nine hours) (MSDE, 1995; Parke, 2007). (See the case study of MSPAP on page 86 for more information.) Because teachers administered the performance tasks, there were concerns about how student performance might be affected by different group dynamics (e.g., how students were grouped with other students to complete the tasks), access to manipulatives or other resources, and the quality of teachers' implementation (Hambleton, 2000). MSPAP was a high-stakes assessment for schools and districts but not for students. Schools were expected to meet standards for satisfactory or excellent performance by 1996 (later revised to 2000), and failing to meet the expectations could lead to reconstitution (Koretz et al., 1996; Yen, 1997). In fact, State Superintendent Nancy Grasmick did reconstitute several schools that failed repeatedly to meet the standards (S. Ferrara, interview, April 24, 2013). So it was important that the public have confidence in the fairness of the assessment.

Finally, another reason for the downfall of many performance assessment systems of the 1990s was that when different performance tasks were administered across students (as was practiced as part of a matrix sampling strategy), it was unclear whether these assessments could be said to be comparable in difficulty. This is one reason why student-level scores could not be reported. While statistical methods were available to evaluate the comparability of performance tasks to each other (through retroactive scaling based on average performance levels across tasks), it is uncertain whether the small number of performance items administered per student would have produced sufficient levels of reliability in the scores to report defensible student-level data.

The same critiques about comparability and standardization have been made about the Wyoming Body of Evidence (BOE) system and the Rhode Island Diploma System, which were initiated in the 2000s. However, in the case of these two initiatives (which are still operational in both states), the portfolios of students' best work are used as a high school exit requirement, rather than as part of the state's NCLB accountability system. This means that they are

not subject to the same demands for standardization as a statewide test that produces scores used to compare and rate schools. However, because the portfolios are used for the high-stakes purpose of determining high school graduation eligibility, these systems have come under criticism for some of the same issues faced by Kentucky and Vermont.

In the Wyoming BOE, which has been operational since 2001, districts draft their BOE plans and determine the performance tasks that will be used to demonstrate proficiency. They may draw from a state-established task bank (developed by a collaborative of educators from across the state) or create their own tasks. Wyoming high school graduates receive endorsements on their diplomas based on their demonstration of proficient performance across nine content areas as delineated below (WDE, 2013).

- Advanced – Student demonstrates advanced performance in a majority of the nine content areas and proficient performance in the remaining content areas
- Comprehensive – Student demonstrates proficient performance in all nine content areas
- General – Student demonstrates proficient performance in a majority of the nine content areas

To graduate from a Wyoming high school, students must be proficient in five of the nine content areas (Dowding, 2011; WDE, 2013). Students must also meet minimum Carnegie unit requirements to earn a high school diploma. The school-level rewards and sanctions tied to the BOE system are minimal. Rewards include public recognition of schools; sanctions include a mandatory period of revision for the district's BOE plan (Dowding, 2011).

As a state with a strong "local control" culture, Wyoming's delegation of authority to districts to design their own BOE criteria makes sense, but it also makes the BOE system vulnerable to critique. Until recently, the state provided oversight for a peer review process of each district's BOE design every two years, resulting in less room for variability. The review process includes five criteria for evaluation, with a focus on alignment to standards and fairness. However, due to recent changes in state agency leadership and declining political support for the BOE system, the state has not regularly monitored the quality of district BOE plans. Critics have voiced concern over the lack of standardization and variability in district implementation, and several politicians have tried to revoke authorization of the BOE system.

Similar to Wyoming's BOE, Rhode Island's proficiency-based graduation requirements, which are

mandated by Rhode Island's Diploma System, are partially designed by districts. Rhode Island's Board of Regents decided in 2003 that all public high schools would revise their graduation requirements, beginning with the class of 2008, to incorporate measures other than tests (Archer, 2005; DiMartino, 2005). However, more recent revisions have delayed enforcement until the graduating class of 2014 (CEP, 2011). To receive a high school diploma, students must do all of the following (Cech, 2008; CEP, 2011; RIDE, 2013):

1. Pass 20 courses (minimum) in core content areas during 4 years of high school;
2. Successfully complete two performance assessments (chosen by the district or school) during the junior or senior year of high school. The assessments may take the form of a senior project, exhibition/comprehensive course assessment, and/or portfolio;
3. Earn a 2 or above (must be *Partially Proficient*) on the NECAP Reading and Math; and
4. Meet any other locally determined school/district requirements.

Like other portfolio-based assessment systems that came before it, the Rhode Island Diploma System is vulnerable to criticism for the lack of standardization of its performance assessments.

Nonetheless, Wyoming's BOE and Rhode Island's Diploma System represent a promising strategy for including performance-based assessments in a state assessment system where complete standardization of the assessments is not required. We might call this strategy a "locally designed common task approach." Both the BOE and Rhode Island Diploma System assessments measure the important skills and deep knowledge not captured by student scores on traditional standardized tests. These systems put a high premium on local autonomy and standards-based learning outcomes for students, instead of high-stakes accountability for schools and teachers. While the Wyoming BOE system is not designed to produce scores that can be used to judge student progress, nor to produce relative rankings of schools for NCLB purposes, the state's strategy of including a state-monitored peer review process focused on alignment to standards and fairness may partially meet the demand for comparability.

Validity and Content Issues

A third criticism of many of the performance assessment initiatives of the 1990s was a perceived focus on process and "soft skills" instead of core content knowledge. Performance assessments are often scored using a common set of rubrics across tasks within a particular content field and task

genre (e.g., persuasive writing, mathematical problem solving). The evaluative criteria used to score performance assessments in the 1990s could cut across any writing or mathematics task, which allowed for a common scale of measurement even if different tasks had different content foci. As a consequence, performance assessment came to be seen as focusing on cognitive skills rather than on rigorous content knowledge. Vermont's mathematics portfolio rubrics, for example, were criticized for focusing on "process skills" and avoiding the assessment of important mathematical content, even though teachers regularly assessed students' content knowledge using traditional mathematics problems that usually had one correct answer (M. Petit, interview, April 22, 2013).

Not only were performance assessments seen as lacking clear alignment to important content, their validity was challenged because researchers asserted that they showed gains in student performance when other measures did not. For example, in Kentucky, Daniel Koretz and colleagues published studies that examined the relationship between mathematics and writing portfolio scores and the on-demand components of KIRIS (selected response items, constructed-response items, and performance tasks) that showed low correlation between the two (Koretz, 1998; Tung & Stazesky, 2010). In addition, Koretz (2002) and Koretz and Barron (1998) published studies examining the relationship between KIRIS scores and other measures with external credibility (e.g., NAEP's on-demand tasks, the ACT). Their studies found that while scores on the on-demand components of the KIRIS rose dramatically between 1991 and 1998, the rise in Kentucky's NAEP scores was roughly the same as the national increase and statistically indistinguishable from gains in most other states. Likewise, ACT scores between 1992 and 1995 did not see a similar rise in Kentucky (Koretz & Barron, 1998). Koretz and other researchers suggested that large KIRIS score gains were attributed to test preparation and growing familiarity with the testing program, not to learning gains (Koretz, et al., 1996; Stecher, et.al, 1998).[18]

This possibility of score inflation associated with test preparation is a

[18] There is disagreement with these researchers' findings among assessment experts who worked in Kentucky at the time. Stuart Kahl, founder of Measured Progress, which was the testing contractor for KIRIS, considers the findings of Koretz and his colleagues to be inaccurate and their attribution of the rise in KIRIS scores to test prep a mischaracterization. Kahl states, "Initial gains were indeed related to familiarity with format, but the increased scores were a better reflection of the kids' abilities and not the teachers 'gaming the system'... The next NAEP reading results were based on 1998 NAEP, and KY was one of only 5 states whose grade 4 scores increased by five or more points – four states increased by five points, and only CT increased by more than that. KY moved from 18th out of 34 states to 11th out of 39 states in grade 4 reading, and maintained that position."

problem not only with performance assessment systems, but with all assessment formats in which the assessment design and content is fairly predictable. The issue of score inflation, whether on performance assessments or traditional selected-response tests, is important because of the resulting scores' frequent misuse by policymakers and others, who interpret a rise in scores, however modest, as evidence of the effectiveness of the policy, rather than test preparation. The implication is that assessments should be designed to be less predictable in content and format (D. Koretz, interview, May 7, 2013). The notion of reduced predictability, however, runs counter to assessment design principles that support the use of consistent task formats ("task specifications") in order to increase task comparability, as well as the idea that teachers and students should know what to expect on assessments in order to adequately prepare for them. One strategy is to continuously generate new performance tasks to refresh the tasks used in large-scale assessments each year – an expensive option. Another strategy is to produce varied assessment forms that both cover the full range of the content standards and are administered within schools so that teachers cannot predict what kinds of tasks or content their students will be tested on. (Both of these are strategies that SBAC and PARCC have adopted.)

In addition to content validity concerns, a lack of sufficient quality control in the performance assessment initiatives of the 1990s made those programs vulnerable to unfavorable publicity. One or more poor quality performance tasks that had not undergone careful scrutiny by bias and fairness committees would show up on the front page of newspapers, resulting in bad press for the advocates of the performance assessment systems. Or the task might include sensitive or controversial subject matter, leaving parents and skeptical policymakers with ammunition to launch attacks on state assessment programs. The CLAS program, in particular, suffered from accusations from conservatives that the performance tasks measured "feelings" rather than basic skills and content (Chrispeels, 1997; Cohen & Hill, 1998). They also charged that the standards being tested were neither clear nor measurable, that the standards lacked rigor, and that the test invaded privacy and promoted a liberal social and cultural agenda (Kirst, 1996; McDonnell, 2004).

CLAS was not the only program to experience a political and ideological battle. MSPAP also faced accusations from conservative activist groups. These groups claimed that MSPAP and the Maryland Learning Outcomes were

examples of "social engineering," lacked rigorous content, and propagated skills like reasoning and problem solving (skills that were then denigrated as part of a liberal social agenda) (Ferrara, 2010). These perceptions were not helped by administrative mishaps in 1992 (the "MSPAP mishap"), in which some schools lacked the materials necessary for, and many teachers lacked appropriate training for managing group activities associated with, the program's new science and social studies performance tasks. Some parents who became incensed by media coverage of attacks on MSPAP also boycotted the tests in 1996 by refusing to allow their children to attend school for two weeks (Ferrara, 2010).

An ongoing criticism of performance assessments is that other factors that are not measured by the task can impact the accuracy of the scores (i.e., differences in student performance that result from student abilities or assumed background knowledge that are irrelevant to students' learning in the content that is supposed to be the focus of measurement). This is called "construct-irrelevant variance". For example, a mathematics word problem that puts the problem in the context of a complex, real-world scenario might be critiqued for hindering some students with strong mathematics knowledge by making assumptions about the test takers' background or cultural knowledge. A problem that requires students to read a complex scenario might also trip up students with underdeveloped language skills (e.g., English language learners). A problem in which students are required to write a detailed rationale or a description of how they solved a problem might be critiqued as being a measure of writing ability rather than a measure of mathematical content knowledge (Hambleton, 2000).

In contrast to the isolated way in which content understandings have been measured in assessments under NCLB, the Common Core State Standards actually prioritize students' literacy and communication skills across the curriculum, including the ability to communicate in mathematics. This is based on the recognition that the ability to explain and justify one's reasoning in various ways is a key skill that students need for college and careers. Instead of representing construct-irrelevant variance, performance assessments that are framed as applications of content in real-world contexts that students must analyze, and in which students must demonstrate their literacy and communication skills, are key to directly and authentically measuring the learning outcomes. Contextualizing a performance assessment in a relevant, real-world scenario is a purposeful design decision that leads to greater authenticity of the performance assessment, engagement, and

motivation for students to persist in completing the assessment, rather than another source of construct-irrelevant variance.

Of course, questions about content validity and construct-irrelevant variance continue to be important issues for any test-item format, including selected-response items, and in the NCLB era, test developers have become highly sensitized to these validity issues. Developers are now required to put all test items through a battery of content, bias, sensitivity, fairness, and accessibility analytics to improve validity and reduce bias to meet technical quality standards for educational and psychological testing set by AERA, the National Council of Measurement in Education (NCME), and the American Psychological Association (APA).

Both SBAC and PARCC have embedded their work in assessment design frameworks that may lead to greater comparability and technical quality in the design of all assessment components. Evidence-Centered Design (developed by Robert Mislevy)[19] is a design framework that both consortia have adopted to ensure that all assessment components have clear measurement targets and that all important measurement targets are assessed. Coupled with content specifications that ensure alignment to specific grade-level standards and task design specifications that provide detailed guidance on designing particular types of tasks, the framework provides a consistent approach to assessment design that is likely to contribute to greater content validity, consistent quality, and comparability across performance tasks.

Inter-rater Reliability and Insufficient Item Reliability

Another technical quality issue in the 1990s that continues to be a source of critique of performance assessment today is reliability. There are two types of reliability that are of particular concern: 1) inter-rater reliability, and 2) reliability of an item or set of items.

Inter-rater reliability. Because performance assessments are typically scored using a scoring rubric that describes 4-5 levels of performance, the consistency and accuracy of the scores are dependent on the professional judgment of trained raters. This is not an issue with selected-response items where only one possible answer is correct. While it has been demonstrated that with well-designed training protocols, sufficient time to train raters, and mechanisms to monitor scoring consistency, it is possible to achieve acceptable levels of rater

[19] For more information on Evidence-Centered Design (ECD), see Mislevy, Almond, & Lukas (2003) and Williamson, Bauer, Steinberg, Mislevy, & Behrens (2004). See the SBAC and PARCC websites for details regarding how ECD is incorporated into their assessment design processes.

agreement (Measured Progress, 2009; Pearson, 2011), some early reports on rater reliability did not find sufficiently robust levels of reliability. In the case of KIRIS, early external audits revealed that teachers' scores on portfolios were higher than those of second raters/auditors (Koretz, 1998; NRC, 2010; Stecher, 1998; Tung & Stazesky, 2010). In response to this finding, the state revised the design of its scorer training and began to regularly audit schools' scoring. Scoring error was cut in half within one year (Hill, 2000; Stecher, 1998; S. Kahl, interview, April 25, 2013), but a panel of measurement experts still found the scoring insufficiently reliable for use in the state's accountability index (Borko, 1999).

Sometimes, research and auditor findings from early in the piloting of a performance assessment program were reported much later, so that even if a program had improved its inter-rater reliability over time, the early results were widely publicized and perceived as an ongoing problem. For example, Vermont portfolio scores were initially deemed too unreliable to be used for accountability (Hewitt, 2001; Stecher, 1998). Over time, however, score reliability improved, especially after Vermont switched from an analytic to a holistic scoring rubric and teachers received additional professional development[20] (Koretz, 1998). Yet despite these improvements over time, early research findings indicating that scoring was insufficiently reliable for either school- or student-level reporting contributed to an ongoing negative public perception that the Vermont performance assessment could not be scored reliably, even though the scores were not used for any accountability purposes at the time. In fact, critics of portfolios still cite the early Vermont studies as evidence that portfolios are inherently unreliable.

If performance assessment scores are used in a high-stakes context (i.e., to hold an individual student or teacher accountable), rater reliability becomes an even more critical issue. The initial low reliability of local scores that was found in the Vermont and Kentucky portfolio systems is highly related to the training, scoring, and audit systems put in place for the use of hand (human) scoring. Local scoring results can be monitored and reliability can be improved through regular external audits of local scoring, but it suggests that there are limits to the reliability of local scoring, especially under high-stakes circumstances. For large-scale assessment systems, distributed scoring approaches (blind scoring of randomly assigned student responses) that utilize a cadre of trained scores from across

[20] A few years into mathematics portfolio administration, Vermont was able to achieve correlations between raters of 0.8 to 0.9 at the level of one holistic score per portfolio.

a state are likely to lead to greater inter-rater reliability than teachers scoring student responses from their own schools.

This kind of distributed scoring system for scoring constructed-response and essay responses has been utilized by large-scale assessment programs and testing companies over the last 15-20 years, producing sufficient levels of reliability for the inclusion of scores in NCLB accountability measures. However, this does not preclude the involvement of teachers in scoring. Electronic scoring using online systems to distribute student responses for scoring across raters in remote locations has been used successfully for many years by national testing programs (e.g., College Board's AP exams, SAT essays) to score large volumes of student essays. In the last twenty years, since the advent of the Internet, testing companies have developed online systems for training scorers, implementing distributed scoring, and monitoring scorer reliability through regular checks for calibration. Because SBAC and PARCC are both being implemented through contracts with testing companies in the industry (and results are being scrutinized by a national audience), it is highly likely that robust systems of monitoring and auditing test scores will be in place to support sufficient levels of reliability.

Reliability of an item (including a performance task) or sets of items. Because performance assessments take longer to administer than traditional forms of assessment, it is unlikely that a large-scale assessment program will include more than a few performance tasks. Moreover, even though performance assessments take longer to complete than a set of selected-response items, performance assessments are limited in their ability to measure a wide range of content and skills. They usually measure a smaller sample of content and skills within a domain, and usually the content and skills being measured are sampled only once. Thus, performance tasks generate a small number of item scores. Sometimes the tasks are scored holistically (i.e., they receive one overall score; for example, students received a score of 1-4 for the entire Kentucky writing portfolio) and sometimes they are scored using analytic rubrics (generating 3-4 scores according to different dimensions). The small number of scores generated by a single task, as well as the limited number of performance tasks that can be administered to a single student, limits the reliability and generalizability of those scores. In other words, is evidence of student proficiency on one performance task generalizable to performance on a range of other performance tasks? This is called "task sampling variability" (Gao, Shavelson, &

Baxter, 1994). Item sampling variability is also a problem of selected-response tests, in which a small number of items are used to measure a single construct (e.g., the ability to divide fractions); however, selected-response tests are able to sample a domain more broadly and with more items. Shavelson, Baxter, and Pine (1991) found that a single science performance assessment provides unstable estimates of student performance and recommend that 6-8 performance tasks are needed to improve the reliability and stability of performance assessments as a measure of student learning and ability on a given construct. Dunbar et al. (1991) show that in writing, 3 or 4 writing samples may be adequate. Shavelson, Baxter, and Gao (1993) note that this problem is true across subject area domains:

> The findings are remarkably consistent across very diverse studies such as writing, mathematics, and science achievement of elementary students (Baxter et al., 1993; Dunbar et al., 1991; Shavelson, Baxter, and Pine, 1991) and job performance of military personnel (Shavelson, Mayberry, Li, and Webb, 1990; Wigdor and Green, 1991). Interrater reliability is not a problem, but task-sampling variability is. Large numbers of tasks are needed to get a generalizable measure (p. 218)

Given the amount of resources and time needed to administer and score multiple performance tasks to measure a single construct like writing, the issue of item reliability might seem an intractable problem given the time and cost constraints of most state assessment programs.

Some state assessment programs have attempted to address the issue of item reliability by combining performance tasks with selected-response or constructed-response items so that the same learning targets are measured in multiple formats. For example, several states that included performance assessment formats in their assessment programs in the 1990s used a balance of selected-response, constructed-response, and performance-based items within a testing program (e.g., Connecticut's CMT and CAPT; Kentucky's KIRIS; Vermont's Portfolio Assessment Program; the New Standards Project); however, almost none of these systems scaled the scores from across these item formats together. Scaling together the scores from selected-response and performance items to measure a set of common measurement targets could achieve greater reliability of the assessment. This has been done in the Connecticut Mastery Test for the assessment of writing by combining the score from the writing task, the Direct Assessment of Writing (weighted at 60%), and the scores from the Editing and Revising

subtest (weighted at 40%) (Connecticut Department of Education, 2013). Research has also found that different methods of measurement (e.g., hands-on tasks, computer simulations, short answers, observations) do not all converge and that they tell us different things about a student's achievement (Shavelson, Baxter, & Gao, 1993). Shavelson, Ruiz-Primo, and Wiley (1999) found that certain methods of measurement (computer simulations, direct observations, and notebook) converged but that these did not converge with paper and pencil counterparts. These findings support the idea that sampling student performance using a variety of measurement methods for every student could minimize task-level variability and bias, and lead to fairer results for students.

This balanced design, using a strategic combination of selected-response, short constructed-response, technology-enhanced, and longer extended-response items to measure a common set of measurement claims and targets is a common feature of the new PARCC and SBAC Common Core assessments. Performance tasks are not the only item format used to measure a single measurement target. The same measurement targets are assessed multiple times across the range of item types. This design may make it possible to include performance items that assess difficult-to-measure learning targets in large-scale assessments without sacrificing the need for sufficient levels of reliability requisite in high-stakes policy contexts.

*

One common theme that has emerged from our examination of the performance assessment initiatives of the 1990s is a tension between the design of assessments for formative use (i.e., to provide detailed feedback to support instructional improvement) and the use of such assessments for accountability. For example, while portfolios and performance tasks that are scored by teachers provide immediate feedback for instructional action, teacher scoring has been found to be inflated in comparison to external scoring. Similarly, while hands-on performance tasks completed in small groups may serve to support and motivate student learning and performance, they are complex to score and interpret, and it is unclear the extent to which students' scores reflect unassisted performance.

Grant Wiggins, who was involved in the development of the Vermont portfolios and worked on the Kentucky committee commissioned with developing the RFP for KIRIS, reflected: "Assessments without the proper design for high-stakes use are not compatible with that use.

[The portfolios of the 1990s, which were designed for formative use] produced scores with limited reliability, were susceptible to [score] inflation, and did not provide student-level scores. The purpose of accountability is different from the purpose of formative assessment. A one-shot test at the end of the year is absurd formative assessment. It has to be embedded, multiple times" (interview, June 11, 2012). Wiggins' observation about the tension between formative and summative purposes in assessment design underlines the importance of intentionally designing future assessments with performance items in ways that are consistent with their intended purpose (e.g., to produce individual scores, to support accountability policies) and the need to make a distinction between the types of performance assessments that may be more suited to formative versus summative assessment.

Lorraine McDonnell (2004), reflecting on the CLAS experience, noted this same tension between the multiple and conflicting policy goals that have been placed on assessment: "Behind the consensus among the governor, state superintendent, and senate education committee chairman lay different expectations for what the new assessment could accomplish. Each of these men supported CLAS for different reasons, and they expected it to accomplish very different things…the political circumstances that created CLAS led to constraints that would eventually hamper its implementation" (McDonnell, 2004, p. 56).

The lesson learned is that assessment developers must be clear about the intended uses of an assessment when making decisions about the design of the assessment. It is apparent from the experience of the 1990s that performance assessment designs that are more authentic and more likely to support student learning (e.g., personalized performance tasks, curriculum-embedded performance tasks, portfolios of work over time) are unlikely to produce comparable tasks and comparable scores that are viewed as sufficiently credible for high-stakes use. Designs that produce fine-grained feedback that support student learning through highly analytic scoring rubrics may be more useful for teaching, but are less efficient and more costly if the goal is to produce reliable scores (as opposed to holistic scoring rubrics typically found in large-scale assessment programs). These design decisions are important, and make a difference in the credibility and viability of an assessment program.

Another lesson learned is that it takes time to get these decisions right. Expecting testing programs to have resolved all reliability and

validity issues within the first two years of implementation is not a reasonable expectation. This has implications for the timing and phase-in of new assessment programs for the purpose of accountability. While policymakers and the public sometimes have a low tolerance for an accountability vacuum, it would be irresponsible for states to use the results of a new large-scale assessment program for high-stakes purposes before the results suggest that such use is technically defensible. Scale-up and phase-in issues are discussed further in the next chapter.

CHAPTER 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 embed performance assessments into large-scale assessment systems in the 1990s were the practical issues that relate to implementing the assessment systems.

Included in this set of factors are:

1. Costs and burdens associated with developing, administering, and scoring performance assessments;
2. Pressure to quickly scale up and use the assessments for accountability; and
3. Need for a coherent system of curriculum, instructional resources, and professional development.

These factors, along with the political context factors and technical quality issues discussed in chapters 1 and 2, contributed to the viability and sustainability of the assessment programs of the 1990s. They had a strong bearing on the outcomes of the assessment programs – i.e., whether or not implementation of the assessment programs had the intended results for teaching and learning. These practical issues also affected the extent to which the new state or local content frameworks or performance standards became embedded in school practice.

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 a 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.

Costs and Burdens Associated with Developing, Administering, and Scoring Performance Assessments

In addition to technical challenges and political factors, the costs, time, and effort needed to develop, administer, and score performance assessments were also major factors in the demise of many of the performance assessment initiatives of the 1990s. Given the limited resources normally allocated to education and educational assessment, it was only when states procured special funding through legislative action or through the award of grants that state departments of education were able to initiate the performance assessment design process or contract with testing companies to begin development. Because of the start-up costs associated with developing new performance assessment tasks, most states would not have been able to make the transition to a new assessment system without this infusion of special funds. (This is similar to the current situation in which states have received an influx of a significant amount of funding from the federal government through the Race to the Top Competition, which awards funding to the Common Core consortia and to individual states to help with the transition to the Common Core through the development of common assessments and local measures.)

Costs associated with development. The overall cost of developing, piloting, and validating performance assessments that meet a standard of technical quality is almost always much greater than the expense of hiring a testing company to develop and validate machine-scored assessments. There are numerous reasons for this. First, a performance assessment

task that integrates multiple content and skill standards in one unit of measurement is larger and more complex in scope than a straightforward selected-response item that measures one standard. As a consequence, designing a performance assessment that integrates multiple content standards requires careful crafting, revising, piloting, review, and polishing to get the items "just right" for measurement purposes. Also, the kind of expertise needed to develop performance assessment tasks is more extensive and requires more specialized training than that received by developers of selected-response items. Many states in the 1990s did not even need to develop selected-response items; instead they could opt to purchase "off-the-shelf" items from assessment item banks that had already been developed by testing vendors (e.g., the Stanford Achievement Tests, the Iowa Test of Basic Skills). This cost-saving option was not a possibility for states adopting performance assessments, which were designed to measure different kinds of knowledge and skills than those measured by tasks in existing item banks and were also often bound by state-specific content frameworks.

Kentucky's experience designing KIRIS items illustrates the high costs of developing performance tasks. In 1991, KIRIS contracted with Advanced Systems, Inc. to develop 602 performance tasks for grades 4, 8, and 12 over five years at a cost of $3,789,150 or about $6,294 per task. In actuality, the first-year development costs exceeded these initial estimates (Hardy, 1995; Hill & Reidy, 1993).

The New Standards Project (NSP) is another example. The NSP spent approximately $14,480 to $14,780 to develop each of their performance tasks. However, NSP tasks were pretested with 198,000 students. Adjusted to a typical trial size of 5,000 students, the cost estimate is reduced to between $5,400 and $5,500 per task, still quite costly (Hardy, 1995; Monk, 1993). States helped fund the development of these tasks by paying between $100,000 and $500,000 per year to be a member of the New Standards Project (L. Resnick, interview, June 14, 2012). Additional funds to support the project came from the Pew Charitable Trusts and The John D. and Catherine T. MacArthur Foundation (Simmons, 1993). However, when the testing company that bought the rights to continue developing and administering the NSP tasks determined that it would not be profitable to continue, the performance assessment component of the initiative was discontinued in the absence of sustained funding (L. Resnick, interview, June 14, 2012).

Vermont's Portfolio Assessment Program tried to reduce the high

cost of performance assessment development by recruiting and involving educators in the task development process (G. Wiggins, interview, June 11, 2012). While this collaborative approach to development provided important opportunities for teachers to build their own capacity for designing and implementing performance assessments and resulted in strong professional networks across the state that remain to this day, it did not produce comparable assessments across teachers and schools, an important requirement for large-scale measurement of student performance. Thus the cost savings were offset by the lack of task comparability across teachers and schools.

Burdens associated with administration. Performance assessments also take more time and a greater level of skill to administer than traditional assessments. In some instances, teachers were responsible for helping students complete and/or compile responses to the assessments (e.g., portfolios, hands-on science labs). For example, in the cases of the Vermont, Kentucky, and Wyoming portfolio systems, the burden of "administering" the assessments fell on the classroom teachers. These systems relied on the professional judgment of teachers (or district personnel, in the case of Wyoming's BOE system) to develop or select appropriate tasks that students would complete to show the kind of evidence that would meet the state's expectations and criteria for quality. With the exception of some networks of teacher leaders who participated in the design and development of those systems, most classroom teachers in those states had little training in the design of performance assessments or the appropriate selection and administration of those assessments. When states did not use regular classroom teachers to administer the assessments, the costs of administering performance assessments grew even higher. Kentucky sent specially trained personnel to schools to administer on-demand performance tasks at an average labor and travel cost of $5 per student (Hardy, 1995). (The on-demand performance tasks were short-lived in Kentucky, lasting only three years, because of the challenges associated with equating them from year to year.)

The performance assessments of the 1990s varied in the amount of time required for completion. According to a 1992 California field test manual, some science laboratory tasks that were part of CLAS took three consecutive periods of 50 minutes each to complete both the selected response and laboratory portions of the exam (CDE, 1992). In MSPAP, eight to ten performance tasks were administered in an on-demand setting, with constructed-response items, in-class lab investigations,

and extended essays requiring up to nine hours of combined testing time over a week-long period (Hambleton, 2000; MSDE, 1995; Parke, 2007). In contrast, artifacts for student portfolios were constructed and compiled over time, usually over an entire year. Generally speaking, performance assessments required more time than the average selected-response test.

The SBAC and PARCC assessment systems include a variety of item types, including selected-response, short constructed-response, and "technology-enhanced" items as well as extended essays.[21] The SBAC selected-response items are computer adaptive testing (CAT) items that are customized to each student.[22] It appears that both consortia have chosen to scale back the number of performance items to be completed by each student. In the SBAC assessment, students complete only one performance item in mathematics and one performance item in English language arts/literacy (see the SBAC assessment blueprints on the SBAC website), while students in PARCC states will complete one performance item in English language arts and multiple performance items in mathematics on the end-of-year assessment (see the PARCC performance-based assessment blueprints on the PARCC website). Nonetheless, the amount of testing time for a student taking the complete SBAC assessment is estimated to be seven to eight hours in total, and the amount of testing time for a student taking the PARCC assessment is estimated to be eight to ten hours, depending on grade level (PARCC, 2013; SBAC, 2012). In addition, because both of these assessments are delivered through a computer interface, and because many schools have a limited number of computers, it is estimated that in some schools where the ratio of computers to students is low it may take as many as three weeks to complete school-wide administration of the test. Thus, while computer delivery of an assessment may be more cost efficient for the test developer, a computer delivered format can exacerbate the administrative burden to schools. A number of technology challenges associated with administering the SBAC and PARCC tests are anticipated, including test item security, network bandwidth and reliability issues, access to hardware, management and support of hardware, and student response security (Moore, 2013).

[21] For an example of an SBAC technology-enhanced item, see: http://sampleitems.smarterbalanced.org/itempreview/sbac/index.htm

[22] Computer Adaptive Testing adjusts the difficulty of questions throughout the assessment based on student responses. This allows for greater efficiency in measuring student performance with a fewer number of items.

Costs associated with scoring. The cost of scoring performance assessments, which are typically hand-scored by individuals who have received scorer training and have met calibration standards, contributes greatly to the higher cost of implementing performance assessments. In contrast to the cost of machine scoring selected-response items, the cost of hiring, training, and compensating scoring personnel is what comprises most of the cost of scoring performance assessments. In the case of California's CLAS initiative, scoring the assessments, including the performance tasks, cost about $30 per student, much more than the $2-$20 per student that it cost to score the more traditional tests that preceded CLAS (McDonnell, 2004). In Vermont, it cost $13 to score each student portfolio (NRC, 2010). In the New Standards Project, the cost to states included $10 to purchase and score each mathematics exam, $12 to purchase and score each English language arts exam, and $14 to purchase and score each science exam (E. Stage, interview, July 12, 2012).[23] Stecher (2010) estimated that in the 1990s, the cost of scoring performance-based items and on-demand essays ranged from $1.50-$15 per student across different performance assessment programs.

There were other reasons for the high cost of performance assessment in the 1990s. At the time, there was no computer-based delivery of assessments, no computer-delivered scorer training, and no computer scoring of performance assessment tasks, though selected-response items could be machine-scored. Performance assessment scorers had to be gathered at central scoring sites to be trained and then again to score large batches of exams over several days. As a consequence, it also took much longer for performance assessment scores to be generated and reported than selected-response scores.

In sum, the higher cost of performance assessment contributed substantially to the demise of the performance assessment initiatives of the 1990s.

Efforts to share costs. As indicated above, the costs associated with developing, administering, and scoring performance assessments can vary considerably, but historically they have overall been high. Moreover, the costs per student for statewide assessment programs will depend on the size of the states. Although the development costs are usually fixed regardless of a state's size, the administration and scoring costs are dependent on the number of students (e.g., printing, shipping, scanning, human scoring). As a

[23] None of these costs have been adjusted for inflation.

result, many states have made efforts to share costs.

The New Standards Project sought to reduce the cost of developing and implementing performance-based assessments through economies of scale. In that case, states contributed $100,000 to $500,000 annually, depending on the size of their student population, as part of their dues for joining the Project. Over time, however, the annual dues were insufficient to sustain the project. As discussed above, the Project relied on a combination of private grant funding, federal funding, and membership dues. When the initial seed funding was expended, the contracted test vendor discontinued work on the performance tasks.

The desire to minimize costs of developing and administering a high quality assessment was also a motivating factor underlying the cooperation of states that are part of the NECAP (New England Common Assessment Program), which includes New Hampshire, Rhode Island, Vermont, and Maine. These states began implementation of NECAP in 2005, in response to NCLB's demand for student-level scores. The Brookings Institution (Chingos, 2012) estimates that participating in NECAP resulted in significant savings for member states, with an average per-student spending of $33, about half of what 10 other states with similar student enrollments (under 200,000 students) spent on average ($62). (Most of the NECAP states have joined one of the two Common Core assessment consortia and plan to discontinue use of NECAP for ELA and mathematics.)

Other instances of states collaborating to form testing consortia include the four consortia that formed to develop assessments of English language proficiency and consortia that formed to develop alternate assessments for students with disabilities. In response to 2001 NCLB regulations requiring the testing of students with disabilities and English Language Learner proficiency in English, four multi-state assessment consortia were formed to develop and pilot assessments of English language proficiency.[24] Similarly, a multi-state consortium was formed to work on creating a framework for the development of assessments appropriate for students with disabilities. The New Hampshire Enhanced Assessment Initiative (NHEAI) and the National Alternate

[24] These include the Mountain West Assessment Consortium (MWAC) (originally comprising 12 states); the Pennsylvania Enhanced Assessment Group (PA EAG) (originally comprising 5 states), which developed the Comprehensive English Language Learning Assessment (CELLA); the State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO) (originally comprising 20 states), which developed the English Language Development Assessment (ELDA); and the World-Class Instructional Design and Assessment(WIDA) Consortium (originally comprising 9 states) which developed the ACCESS for ELLs assessment (Wolf, et. al, 2008).

Assessment Center (NAAC) partnered to identify key issues in developing technically sound alternative assessments that could be used for NCLB accountability systems (Quenemoen, 2008). Most of these consortia were originally funded through the U.S. Department of Education's Enhanced Assessment Grant program (2002-2012).

A study of assessment systems under No Child Left Behind (Topol, Olson, Roeber, & Hennon, 2013) found that "typical" assessment systems[25] cost an average of $19.93 per student in 2010, while "high quality" assessment systems that incorporate essay formats cost an average of $55.67 in 2010. These "high quality" assessment systems are ones that typically include subject areas that go beyond mathematics and reading/writing, and include a variety of item formats, including selected-response, short constructed-response, and essays. As noted earlier, per student assessment costs also vary depending on the size of the student population within a state. The Brookings Institution developed models for estimating the cost savings for states with large student populations (Chingos, 2012). In their most conservative model, they estimated that a medium-sized state with about one million students in grades 3-9 (such as Illinois) would spend approximately $24 per student on testing, about 35 percent less than a state of about 100,000 students (such as Maine), which would spend $37.

It appears that one of the benefits of joining a multi-state consortium is the economies of scale that can be achieved when states combine resources and share the cost of developing, administering, and scoring assessments. PARCC estimates their summative tests in reading, writing, and math will cost $29.50 per student (PARCC, 2013). The median cost for *current* assessments used by PARCC states is $29.95 per student -*more* than the estimated cost of using PARCC's summative assessment. (Cost estimates for PARCC's diagnostic and formative assessments have not been released.) SBAC estimates its complete system of assessments (summative, interim, formative) will cost $27.30 per student (less than what two-thirds of its 24 states currently spend). For those states only using the summative assessment, the cost will be $22.50 per student (Willhoft, 2013). Within the SBAC consortium, current state spending on tests ranges from as low as $9 per student (North Carolina) up to $69 per student (Maine) (Gewertz, 2013). SBAC's cost estimates include the cost of

[25] These are the costs passed down to states by testing companies through contractual terms. The cost distribution across development, administration, and scoring functions is unknown.

delivering the assessment, providing help-desk services, and hand-scoring performance items; however, the cost of hand-scoring will be taken on by each state, so the actual cost for states may fluctuate depending on how they manage their contracts for those services. The cost estimates of PARCC and SBAC include both machine-scored and hand-scored items.

In addition to cost savings associated with multi-state collaborations, the introduction of computer-based delivery systems appears to play a role in reducing assessment costs at the state level. However, electronic delivery relies on schools and districts having an adequate technology infrastructure and may add to the administrative burden and costs to local districts and schools. While some schools have the resources available for 1:1 computing, many schools, particularly those in districts with few resources, have only one functional computer lab available and/or low computer to student ratios. The implication of administering an assessment program entirely by computer is that groups of students in a school with few computers will need to cycle through the computer lab for multiple components of the summative assessment over time, resulting in an extended testing period. For electronic delivery of assessments to go to scale, states will need to allocate resources to local districts and schools to improve the computer to student ratio, improve broadband quality, and train school staff as well as students on how to use the technology[26]. Both PARCC and SBAC have made paper-based options available at additional cost to states.

PARCC and SBAC's estimates have yet to be tested, so whether the cost of performance assessment will actually be more manageable than in previous instances has yet to be proven. Part of the reason that the costs have been kept under control is that, in both systems, performance-based items play a much more modest role in the overall assessment than was initially proposed. In SBAC's case, the number of performance items in the summative assessment was scaled back from two performance tasks in mathematics and two in language arts to one performance task in each subject area (SBAC, 2012). In PARCC's case, a series of through-course assessments has been scaled back to a summative assessment in each subject area, with two optional non-summative assessments that provide formative information. PARCC's extended performance

[26] Concerns about the readiness of students to successfully complete computer-administered tests may have been overstated. In a survey of over 10,000 Idaho students who participated in the Spring 2014 SBAC field test, only 5% of 3rd-5th graders said navigating the test was difficult and just 2% of 3rd-5th graders said they could not use the keyboard to type their answers (Idaho SDE, 2014).

items are integrated with other item formats, and the current design indicates that each student will complete one performance task in language arts and a few performance tasks in mathematics during the summative Performance-Based Assessment administered 75% into the school year (PARCC, 2013).

Pressure to Scale Up Quickly and Use Assessments for Accountability

A second practical factor that often sabotaged the success of performance-based assessment programs is the political need to get the new assessment system in place faster than is warranted by the established procedures for test development. Assessment systems with performance-based components require time for development, pilot testing, field testing, and to conduct appropriate research to establish their validity and reliability. At least two development years are needed – the first year to develop prototypes and conduct a small-scale pilot test, and a second year to use results from the pilot test to make modifications and test those modifications in a larger field trial. Additional time is needed to conduct the research appropriate to the assessment's intended uses before it is implemented and used for accountability purposes. Moreover, phasing in a new assessment program over a period of time in which there are no consequences for the test takers, schools, and districts allows time to transition to the new assessment before they are held accountable for the assessment's results. Taken together, it may take up to five years from the beginning of development to scale up to operational implementation of a new assessment system.

Because of policy mandates, the actual amount of time given to develop, test, and validate new assessment systems is typically shorter than five years. Unfortunately, in the policy world, support for innovation in assessment can fluctuate with every election cycle, or even more frequently. The pressure to scale up quickly and bring assessment systems to operational use comes both from the desire of politicians and special interest groups to show results for their policy actions, and a seeming lack of tolerance for an accountability vacuum. This means that assessment quality can be compromised when state departments and contractors are forced to rush their work schedules without sufficient safeguards and opportunities for review and revision built into the process. Moreover, the ability to carry out the reliability and validity studies needed to support the assessment process is likely to be also compromised. It also means that assessments come under public scrutiny much earlier than they are ready to be evaluated. A prime

example of this is the case of Kentucky's portfolio system. One psychometric report (Koretz, 1998) published seven years after the inception of implementation reported low levels of reliability based on a first-year audit of local scoring, contributing to a mistrust of the system's scores even after several years of implementation saw increases in inter-rater reliability (S. Kahl, interview, April 25, 2013). Another report by Koretz and colleagues pointed out the inconsistency between improvements in KIRIS scores and Kentucky's NAEP scores, which remained flat through 1994, suggesting that improvements in scores were due less to real changes in student learning and more to "teaching to the test" (Koretz et al., 1996). Although there were legitimate reasons for scores to rise (e.g., teachers and students became more familiar with the non-selected-response formats used in KIRIS), reports such as these, published and disseminated before the assessment had time to be refined and understood, ultimately gave opponents the fuel they needed to overturn support for the assessment program.

Similarly, in California, after only two years of operational use, the CLAS was discontinued following the release of a report by the "Select Committee" enumerating the myriad problems in the first-year assessment design and implementation (Cronbach et al., 1994). Before the California Department of Education had time to correct the problems in the assessment system, the plug had been pulled by Governor Wilson.

The Need for a Coherent System of Curriculum, Instructional Resources, and Professional Development

Assessment-focused policies often overlook the need for curriculum and instructional resources to communicate, clarify, and build understanding of new standards in concrete ways that help teachers translate the standards into high quality instruction.

As noted earlier, the underlying theory of action driving performance-based assessment initiatives of the 1990s was a belief that changing the characteristics of an assessment would change what and how teachers teach, which would in turn lead to improvements in student learning. Khattri, Kane, and Reeve (2012) report that there is a growing body of evidence that the use of performance assessments improves teaching and learning, citing the work of Hilda Borko and colleagues (1993), Beverly Falk and Linda Darling-Hammond (1993), Maryl Gearhart and colleagues (1993), the Kentucky Institute for Education Research (1995), Daniel Koretz and colleagues (1993), and Smith and colleagues (1994). In some instances, performance assessment has been shown to contribute to improved

instructional practices. In Kentucky, Matthews (1995) found that "40 percent of teachers reported that the open-response items and portfolios [of KIRIS] have a great deal of positive effect on instruction, and virtually none reported that about multiple-choice items" (p. 11). A report on the Maryland School Performance Assessment Program (MSPAP) similarly found that "98 percent of school principals felt MSPAP has a positive effect on instruction" (Koretz, et al., 1996, p. 29). Additional research suggests that in Vermont and Kentucky, as a result of the writing portfolios, teachers increased student writing activities and increased the level of group work in their classrooms. This finding was confirmed in a survey study of University of Kentucky college students who had been K-12 students during the implementation of the Kentucky writing portfolios. Researchers found that "almost three-quarters of the students reported writing daily in high school in a variety of disciplines. Approximately one-half rated their writing abilities as "above average" or "excellent" and felt prepared or somewhat prepared to write in college" (Spalding & Cummins, 1998, p.167).

Graduates of the New York Performance Standards Consortium (NYPSC) schools – a coalition of small schools that feature a commitment to performance assessment, inquiry-based learning, and project-based assignments – have been shown to have a lower dropout rate (10.6% vs. 20.3%) and a higher college-bound rate (87.8% vs. 70.1%) than traditional New York City public high school students in 2003 (Foote, 2005). NYPSC graduates attending college were found to earn, on average, a 2.6 college GPA – equivalent to a B minus average (Foote, 2005). Chung and Baker (2003) found that college engineering students engaged in a capstone performance assessment showed "significantly higher content scores" after completing the performance task, and that, most significantly, these gains were highest in the area of deep propositions (p. 25).

Other studies (e.g., Firestone, Mayrowetz, & Fairman, 1998) have found little or no instructional changes or student learning gains attributable to high-stakes performance assessments. There is also evidence that the rewards and sanctions associated with high-stakes tests will lead some teachers to do whatever it takes to improve their students' scores on the assessments, whether that means spending significant amounts of time doing practice tests, teaching only the content that is expected to be on the test, or, in some cases, cheating (Hout & Elliott, 2013; Madaus, 1988; Madaus et al., 1992; Nichols and Berliner, 2005; Shepard, 1990; Smith, 1991).

Policymakers looking for cost-efficient methods to enact school reform have come to rely on assessments and school accountability as a matter of course. However, assessment scholar Lorrie Shepard suggests that the most important way in which states and districts can support the transition to the Common Core State Standards is not through assessment, but by developing, disseminating, and providing professional development for teachers about how to use a common ambitious curriculum (interview, February 25, 2013). Shepard suggests that the main problem with current assessment development approaches is that they are divorced from the development of curricula.

> *"Where I see this being done well in other countries is when performance assessments are part of a curriculum development process where experts and expert teachers are brought together to work through what the instructional units should look like, what the embedded performance tasks – not tests – should be, and then what an intentional set of extensions should be, first within the instructional sequence… But you also have to work out, given that representation of content mastery, how would we test for that on the summative test? There should be conscious working out of those relationships. Instead, what the country keeps investing in is tests, or consortium assessments, and then hoping that someone can backwards translate into curriculum, instructional activities, and formative assessments… people are still separating developing ambitious assessment from the curricula. You can't go deep with assessment if you don't know the particular material that you're learning with. They think these skills can be separated from content, and there's good evidence from psychology that this is not the case" (L. Shepard, interview, February 25, 2013).*

What Shepard has identified is a problem that has been an issue since the beginning of the standards-based reform movement of the 1990s. The movement's underlying theory of action was that high quality standards would make it possible to develop a coherent system of standards, assessment, curriculum, and instruction (O'Day & Smith, 1993). However, what we saw again and again in the performance assessment initiatives of the 1990s is a lack of follow-through on the development of coherent educational systems envisioned by standard-based reform.[27] Lack of resources and

[27] Firestone, Mayrowetz, and Fairman (1998) likewise concluded that the main reason they did not find significant changes in instructional activities related to the implementation of statewide performance

state capacity often meant that the curriculum and instruction components of standards-based reform became afterthoughts (or lost opportunities), rather than essential components.

Educators and scholars alike have expressed concern about the absence of curricular resources that translate standards into instructional practice, and the resulting implications of how educators respond to assessments. Instead of focusing on how to align content and instruction to the standards, teachers exhibit a tendency to solely use the assessment to guide the content of their instruction, assignments, and classroom assessments. Daniel Koretz noted that the more predictable a test is in terms of content and format, the easier it becomes for teachers to "teach to the test" (interview, May 7, 2013). Without sufficient professional development, there is potential for assessment criteria rather than the standards to drive instruction. For example, there was evidence in Vermont that in the absence of clear content frameworks, curricula, and sufficient professional development, "rubric-driven instruction" occurred, meaning that teachers shaped their instruction and classroom assignments more narrowly to meet the criteria embodied in the scoring rubrics used to evaluate mathematics portfolios. "Teachers may emphasize some problem types or response formats over others because they fit the rubrics, or they may discard otherwise appropriate problems that only permit high scores on only four or five of the scoring criteria. To the extent rubrics oversimplify problem solving and fail to represent useful problem-solving skills, teachers may do students a disservice by overemphasizing the rubrics in curricular and instructional planning" (Stecher & Mitchell, 1995, p.31).

Teaching to the test can lead to apparent incremental improvements in student performance on the tests, but not necessarily because student learning has improved (Koretz, 1998). These incremental improvements reflect a greater focus on the content areas and skills that are included on the tests, and come at the expense of less attention to other curricular areas that are no less important or aligned to the standards.

Not only does the lack of common curricular resources make it less likely that there is coherence among assessment, curriculum, and instruction, it also leads to questions about whether students will have sufficient opportunities to learn the kind of knowledge and skills that are

assessments in Maine and Maryland was because teachers lacked aligned curriculum and accompanying professional development.

measured by the assessment. The "opportunity to learn" gap is a serious issue that measurement experts have identified as a threat to the validity of assessments (Airasian & Madaus, 1983; Haertel & Calfee, 1983; Linn, 1994; Schmidt, 1983).

Professional Development. Hand in hand with the need for curriculum resources is the need for professional development for teachers to understand the new standards and how the standards translate to concrete changes in instructional practice. Again and again, one of the major lessons cited by individuals involved in the performance assessment initiatives of the 1990s was a need for teacher professional development. In the past, the adoption of new standards required teachers to implement different kinds of assessments as a matter of course – in order to give students practice with new ways of demonstrating their application of content knowledge and skills. The new performance standards also expected teachers to change their role as teachers, from transmitters of knowledge to facilitators of student learning, empowering students to make decisions and choices in how they complete performance assessments. State education departments had limited capacity to provide professional development that supported these new expectations.

In states where teachers and other educators were integrally involved as stakeholders in the process of developing the assessment items or in scoring the performance assessments (e.g., Vermont, Kentucky, Connecticut, Maryland), the experience of those teachers involved in this kind of assessment-related work supported their ability to transition to the new standards and assessments more quickly and resulted in greater "buy-in." Many of these teachers went on to become teacher leaders responsible for professional networks of teachers engaged in ongoing professional development related to the new standards and assessments.

In other states, where few educators were involved in either the development or scoring of the performance assessments, professional development was largely absent as part of an implementation strategy. These states and districts had limited resources and capacity to provide professional development around the new standards. Or professional development was simply not on the radar of assessment leaders in state education agencies, where personnel overseeing assessment, curriculum, and professional development are often organizationally segregated from one another. In California, because local educators who were implementing CLAS had a limited understanding of the rationale for CLAS or the new standards, they

had a difficult time defending the assessment when it came under attack during its second year of implementation (L. McDonnell, interview, March 30, 2012). Given the burden of administering the lengthy and complex CLAS assessments, low morale when students performed poorly on the new assessments, and the growing pains associated with transitioning to a new set of content frameworks and creating new instructional plans, it became easy for educators and school organizations to back away from the new assessments and withdraw support, even if they had initially supported the new approach.

Researchers have consistently argued that involving teachers in the design, supported implementation, and scoring of performance assessments increases the link between instruction, assessment, and student learning, and encourages reflective teaching practices and active learning (Darling-Hammond & Falk, 2013; Herman, 1997). However, it appears that many of the real-world decisions around the design of assessment and accountability systems have less to do with how assessments support and improve teacher instruction and more to do with the policy goal of holding schools and teachers accountable. If the goal of implementing a large-scale performance assessment system is to improve teaching and learning and support integration of the Common Core State Standards into curriculum and instruction in ways that bolster students' preparation for college and careers, then supporting teachers' professional learning to expand their instructional repertoire, along with the provision of curriculum resources aligned to the intentions of the CCSS, seems to be paramount to ensuring that the desired results are achievable.

Lorrie Shepard, Joan Herman, Grant Wiggins, and others argue that the most important mechanism for improving and aligning instructional practice to the Common Core State Standards is not through assessment alone, but through a coherent system of curriculum and instructional resources and professional development (J. Herman, interview, February 28, 2013; L. Shepard, interview, February 25, 2013; G. Wiggins, interview, June 11, 2012).

CHAPTER 4

What are the Conditions for Sustainability? A Closer Look at Three States' Performance Assessment Programs

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 when student level results are desired, lack of comparability across assessments) that could not withstand the increased demands for assessment-based accountability. 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 overall 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. The story of Connecticut's successful experiment with performance assessment is documented more closely in the following pages.

In contrast to Connecticut is the case of Kentucky, which suffered from attacks both on technical grounds and political grounds. While KIRIS (Kentucky Instructional Results Information System) survived from 1991-98, it could not withstand the increased demands for the use of assessment for accountability purposes, nor strong opposition from politicians. Like Connecticut, Kentucky also had an assessment model that balanced multiple formats – selected-response items, constructed-response items, and a portfolio. The technically weakest component (though seen as having the most instructional impact) was the portfolio, which was comprised of locally developed (e.g., teacher-designed) tasks and scored by local teachers (with a state audit). Ironically, after KIRIS was dismantled in 1998, the one component that remained explicitly in use until 2009 was the writing portfolio (NRC, 2010). On the other hand, Stuart Kahl characterizes the new CATS (Commonwealth Accountability Testing System), which replaced KIRIS, as primarily a "name change" meant to convince opponents that the KIRIS was "dismantled". Kahl asserts that there were only modest differences between the final version of KIRIS and the CATS assessment (S. Kahl, interview, October 11, 2013), and that the shift to the CATS was primarily due to withdrawal of political support, rather than technical quality issues. The story of Kentucky's experiment with KIRIS and the reasons for its ultimate demise are explored in more depth in the following pages.

Maryland's MSPAP (Maryland State Performance Assessment Program) represents a third, relatively successful approach to incorporating performance assessment into a large-scale assessment system. In contrast to Connecticut and Kentucky's assessment programs, which balanced different item formats, MSPAP included only performance tasks. The program survived for 11 years (1991-2002) by virtue of its ongoing focus on technical quality and defensibility. Ultimately, it was defeated because it could not meet NCLB's demand for student-level scores for every student, and it faced mounting political opposition. MSPAP is discussed in more detail in the following pages.

Connecticut

Connecticut Mastery Test (CMT) & Connecticut Academic Performance Test (CAPT)	
Duration	CMT: 1985 – present CAPT: 1994 – present
Grades Tested	CMT: 3–8 CAPT: 10
Content Areas	CMT: Mathematics, reading, writing, science (science added in 2008) CAPT: Mathematics, reading (interdisciplinary), writing (interdisciplinary), science
Description of Assessment	CMT: Selected-response and open-ended items, essay responses CAPT: Selected-response and open-ended items, essay responses, questions related to curriculum-embedded performance tasks, on-demand performance tasks (eliminated in 2007) (See Appendix B, pages 127-135, for sample CAPT items.)
Technical Characteristics	Criterion-referenced. Scale scores within a grade and content area comparable from one year to the next; scale scores not comparable across grade levels.
Timeline	Administered in March, scores released in August
Scoring	Scored by Measurement Incorporated; prior to 1992, CMT scoring was conducted within the state
Score Reporting Level	Individual student score reports; school and district summary results
Accountability System/Purpose of Assessment	CAPT and CMT were designed to be low stakes assessments. Stakes have risen as both tests are now used to meet federal NCLB requirements. CAPT scores are included in district/school graduation criteria, but cannot be the sole criteria for graduation.
State Standards/ Frameworks	*The Connecticut Framework: K-12 Curricular Goals and Standards*
Current Status	CMT and CAPT will be administered for the final time during the 2013-2014 school year; Connecticut will begin using assessments from the Smarter Balanced Assessment Consortium (SBAC) in 2014.

Connecticut provides important lessons about factors supporting the sustainability of an assessment program that incorporates performance assessment as one component. Strong technical quality, flexible responses to a changing policy context, and strong leadership that provided consistent support are characteristics that make Connecticut a distinctive case. Connecticut's assessment system includes a balance of on-demand open-ended items and curriculum-embedded performance tasks, coupled with more traditional selected-response items. Consistently supportive state leadership over a span of nearly three decades, together with the state's willingness to adapt its assessments to meet the demands of NCLB, has allowed the Connecticut Mastery Test and the Connecticut Academic Performance Test to stand the test of time. Strong teacher involvement and public engagement have also led to strong on-going support.

Background

Gerald Tirozzi became Connecticut Commissioner of Education in 1983 and soon after released recommendations to establish low-stakes mastery tests for students in grades 4, 6, and 8. In response, Governor William O'Neill committed $20 million to a trust fund for education and tasked Connecticut's Bureau of Student Assessment and Research with developing a new statewide assessment (Wilson, 2001). The Connecticut Mastery Test (CMT) was first administered in 1985.

Initially only administered to students in grades 4, 6, and 8, students in all grades 3-8 have annually taken the CMT since 2006 as a requirement of NCLB (McAuliffe, 2007). Additionally, students in grades 5 and 8 began taking the CMT Science assessment in 2008 (CSDE, 2012). The Connecticut Academic Performance Test (CAPT), administered to the state's 10th graders, was introduced in 1994 as a performance-based high school component of Connecticut's assessment system.

Strengths

Connecticut's success with CMT and CAPT is largely due to the sustained support of its leadership. Gerald Tirozzi's successors as Commissioner of Education, Vincent Ferradino and Theodore Sergi, along with key staff members at the Bureau of Student Assessment and Research, continued to promote Tirozzi and O'Neill's vision of a comprehensive system of aligned, well-supported, and well-tested assessment policies long after Tirozzi had left office (Wilson, 2001). This continuity of vision among Connecticut's education leaders has produced solid political and public support for CMT and CAPT and has also established the CDE as a learning organization.

Because of the initial low stakes associated with CMT and CAPT, there has been ample opportunity for reflection, multiple revisions, and improvement.

Another key component of Connecticut's success has been the involvement of educators at nearly all levels of the assessment cycle. Both the CMT and CAPT were designed with input from multiple advisory committees, including Connecticut educators (CPRE, 2000; CSDE, 2012). Connecticut teachers were instrumental in developing Connecticut's curriculum frameworks, and were initially responsible for scoring CMT exams, which provided rich opportunities for professional learning (Baron, 1996).

Although the CMT and CAPT faced few technical challenges, the assessments as originally administered did not meet the requirements of the newly introduced NCLB accountability system because they did not test every student, every year, in grades 3-8. Instead of abandoning the assessments, however, state leaders chose to modify their design. All students in grades 3-8 began taking the CMT in 2006, and a science section was added in 2008. Administration was moved from fall to spring, and individual score reports continued to be provided to all students (D. Rindone, interview, April 30, 2013). This adaptability has allowed Connecticut to retain an assessment system that includes performance tasks as opposed to resorting to a more traditional, primarily selected response assessment format.

Challenges

Unfortunately, performance standards of achievement were lowered as a result of NCLB requirements. As initially designed, the state goal for performance on CMT was *Goal* (the five levels of achievement on the CMT are *Advanced, Goal, Proficient, Basic,* and *Below Basic*), but that threshold has been lowered to *Proficient* (CSDE, 2012; McAuliffe, 2007). Furthermore, NLCB's requirement that all students be tested, including those with IEPs and LEP, added to assessment design expenses. The CAPT exams now include fewer performance components as a result of the increased cost of testing at additional grade levels and the higher costs of performance assessments (CPRE, 2000). CAPT's science assessment originally included an on-demand hands-on performance task (a science lab), though that was eliminated in 2007 (Stecher, 2010). Nevertheless, the CAPT science exam still consists of open-ended items related to five curriculum-embedded performance tasks (labs or experiments) that are incorporated into 9^{th-} and 10^{th-}grade science classes (Darling-Hammond, 2010; Stecher, 2010).

Lessons Learned

Connecticut was able to adapt to meet the demands of NCLB by utilizing a mix of selected-response, constructed-response, and performance-based task assessments. The CMT and CAPT survived the policy changes of NCLB because of the state's strong leadership and ability to adapt, coupled with strong buy-in from teacher and education communities.

Kentucky

Kentucky Instructional Results Information System (KIRIS)	
Duration	1991 – 1998
Grades Tested	4, 5, 7, 8, 11[28]
Content Areas	Reading, writing, mathematics, social studies, science, arts and humanities, practical living/vocational studies
Description of Assessment	Selected-response items, open-ended written tasks, performance events, and a portfolio (math, writing) reflecting a student's best work[29] (See Appendix B, pages 136 -137 for sample KIRIS items.)
Technical Characteristics	On-demand test components administered through matrix sampling; portfolios scored holistically
Timeline	Assessment administered in spring with results reported annually; schools formally evaluated every two years (accountability cycle)
Scoring	Portfolios in writing and math locally scored by teachers with a sample sent to the state for rescoring to establish reliability; on-demand components scored by outside testing company
Score Reporting Level	School performance data; individual student scores not released
Accountability System/Purpose of Assessment	KERA (Kentucky Education Reform Act) instituted a school accountability index comprised of KIRIS results and non-cognitive measures (dropout rates, attendance rates, etc.). KIRIS results accounted for five-sixths of each school's score. Each school received an overall score on a scale of 0-140; all schools were expected to meet the long-term goal of at least 100 at the end of 20 years. Schools that reached or exceeded their short-term target score could receive monetary rewards; sanctions for schools that failed to reach their target score

[28] KIRIS originally tested students in grades 4, 8, and 12 in reading, writing, social science, science, mathematics, arts and humanities, and practical living/vocational studies. Assessments were divided between grades 4/5 and 7/8 beginning with the 1996-97 school year. In grades 4 and 7, students completed on-demand assessments in reading, science, and writing, plus a yearlong writing portfolio; in grades 5 and 8, students completed on-demand assessments in math, social studies, arts and humanities, and practical living/vocational studies, plus a yearlong portfolio in math. Testing was moved from 12th to 11th grade in 1995 (Koretz & Barron, 1998; NRC, 2010; Stecher, 1997).

[29] Test composition changed several times and not all task types were included each year. The on-demand open-ended writing task was added in 1997. Multiple choice items were eliminated in 1995 but reintroduced in 1997. On-demand performance tasks across all five subject areas were dropped in 1996.

	included state takeover, mandatory School Transformation Plans, or intervention by a "distinguished educator."
State Standards/ Frameworks	*Transformations: Kentucky's Curriculum Framework, Volume I (1993) and Volume II (*1995), *Core Content for Assessment (*1996)
Current Status	Due to high costs and mounting political opposition, KIRIS was dismantled in 1998 and replaced by the Commonwealth Accountability Testing System, which included writing portfolios and on-demand testing components, including a shorter writing task and selected-response items.

Kentucky was one of the first states to implement a comprehensive state education accountability system. The Kentucky Instructional Results and Information System (KIRIS), the state's assessment system, included a balance of item formats, including selected-response questions, on-demand and curriculum-embedded performance tasks, and a portfolio component. KIRIS was supported by a focus on teacher professional development and a statewide effort to establish clear curricular frameworks. However, KIRIS faced many public challenges to its technical quality and high-stakes usage. In the wake of concerns about its technical quality and costs, combined with state political instability, KIRIS was discontinued in 1998, replaced by the Commonwealth Accountability Testing System (CATS), which included many of the same features of KIRIS, including the writing portfolio, a shorter on-demand writing task, and selected-response items.

Background

Kentucky's 1989 court case *Rose v. Council for Better Education* decided that the state's school system, in which significant numbers of children received an inadequate education, was inherently unequal and unconstitutional. In response, Kentucky began "sweeping education reforms" instituted through the Kentucky Education Reform Act (KERA) of 1990.

KERA reformed education finance, adding close to $700 million to public education over two years and providing funding for 10-12 days of teacher professional development each year (McDonnell, 2004). The Kentucky Instructional Results Information Systems (KIRIS) was designed in response to KERA's call for a high-stakes "primarily performance-based" assessment to account for the state's financial investment (Gong, 1996; Wolf, 2000).

Strengths

Kentucky made a substantial investment in both funding KIRIS

and supporting teachers in its implementation. The state provided funding for up to 12 days of teacher professional development each year, with a focus on supporting teachers in making instructional changes that would align with KIRIS's focus on critical thinking and higher-order thinking skills. State officials both supported technical assistance networks throughout the state and invested in extensive training and calibration for teacher portfolio scorers (McDonnell, 2004).

KIRIS's writing portfolios received strong support from teachers, school leaders, and state legislators (Gong, 1996). Teachers reported that portfolios had a positive influence on students' writing, describing improvement in student performance as "strong to dramatic" (Hill, 2000; Stecher et al., 1998). Although KIRIS was discontinued in 1998, its writing portfolio component survived as a part of the state's accountability system until 2009 (NRC, 2010; Pecheone & Kahl, 2010).

KIRIS was modified several times in response to lessons learned from each test administration: 2 of its 6 learning goals were eliminated in 1994; the high school assessment was moved from 12th to 11th grade in 1996; and selected-response items were removed or introduced in various years. This adaptability allowed the state to keep what worked and try to find alternatives for what did not work, ultimately leading to a stronger assessment system.

Challenges

KIRIS faced many technical challenges. Early external audits of writing portfolio scores revealed that teachers' scores on student portfolios were much higher than those of second raters, causing many policymakers to question the writing portfolio assessment's reliability (Koretz, 1998; NRC, 2010; Stecher, 1998; Tung, 2010). Moreover, external researchers reported that the gains in student performance on on-demand portions of the KIRIS assessments were not reflected in NAEP and ACT scores, and KIRIS portfolio scores were not comparable to KIRIS on-demand scores (McDonnell, 2004; Koretz, 1998; Koretz et al., 1996; NRC, 2010; Tung, 2010).

The large variation in portfolio practices (including the diversity of assignments and varying levels of student independence in completing tasks) made the comparability and technical credibility of the portfolios questionable. Although portfolio score reliability improved due to intensive scorer training, a 1995 panel convened by the Kentucky General Assembly concluded, "portfolio scores are not at this time appropriate for use in the KIRIS high-stakes accountability system" (Koretz, 1998, p. 331; Tung &

Stazesky, 2010). (It should be noted, however, that the program's technical advisory committee disagreed with the panel and the writing portfolios continued for many years with much improved scorer consistency.)

Citing a cost of $29.5 million over 5 years, policymakers and other groups opposed to KIRIS found a wealth of reasons to fight against its continuation (Strong, 1996). A small opposition of mostly conservative groups cited three main problems: a lack of public confidence in the system, KIRIS's insufficient focus on academics (the "basics"), and its technical problems related to design, administration, and scoring (McDonnell, 2004).

In addition, policymakers began calling for individual student score reports, a demand that could not be met under KIRIS's existing design (B. Gong, interview, June 8, 2012; G. Wiggins, interview, June 11, 2012). Changes in Kentucky's political leadership further decreased support for KIRIS as Democrats and Republicans both sided against KERA in their fight for control of the state legislature (NRC, 2010). In the last year of KIRIS, selected-response items were reinstated, allowing for student-level score reports, but by then, it was too late to turn the tide of political forces.

In April 1998, Kentucky Governor Paul E. Patton (D) signed H.B. 53 dismantling KIRIS and implementing CATS (Commonwealth Accountability Testing System) (White, 1998; Wolf, 1999).

Lessons Learned

Kentucky's experience with KIRIS provides valuable lessons that can improve the quality of future performance assessment initiatives. Kentucky supported teacher capacity-building measures such as targeted professional development and investment in teachers as assessment scorers. Similarly, future assessment initiatives should seek to link capacity-building strategies to assessment policy (McDonnell, 1994; G. Wiggins, interview, June 11, 2012).

The limits of high-stakes assessment design were evident within KIRIS. Performance tasks, and specifically curriculum-embedded tasks, come with technical limitations that may not align with high-stakes accountability systems. Assessments without the proper design for high-stakes use are not compatible with that use (G. Wiggins, interview, June 11, 2012).

As became evident in Kentucky, strong political leadership is necessary for the success of any assessment initiative. Kentucky's internal political battles destabilized its education agenda and contributed to KIRIS' downfall. Parents and local communities should be involved in the standards and test development process to

minimize the potential for opposition, and a continuous vision of assessment supported by strong state leadership is essential for success (McDonnell, 2004).

Maryland

Maryland State Performance Assessment Program (MSPAP)	
Duration	1991 – 2002
Grades Tested	3, 5, 8
Content Areas	Reading, writing, language usage, mathematics, science, social studies
Description of Assessment	8-10 on-demand performance tasks (some interdisciplinary), including pre-assessment group activities with manipulatives. No selected-response items. (See Appendix B, pages 138 - 147, for sample MSPAP items.)
Technical Characteristics	At each grade level, 20 tasks were used to assess school performance across the six content areas. Students were assigned on a random basis to one of three test form "clusters." Each cluster included just 8-10 of the grade's tasks, meaning that each student did not complete all 20 of their grade's tasks.
Timeline	Assessment administered in May; score reports released in November
Scoring	Scored by Maryland teachers; teacher scoring procedures were moderated through check sets, accuracy sets, and spot checks
Score Reporting Level	School performance data; individual student scores not released
Accountability System/Purpose of Assessment	Schools were expected to meet standards for satisfactory performance by 1996 (later changed to 2000). A school was rated satisfactory if 70% or more students scored level 1, 2, or 3 on MSPAP's five-point scale.
State Standards/ Frameworks	Maryland Learning Outcomes
Current Status	MSPAP could not feasibly meet the requirements of NCLB; it was replaced by a more traditional on-demand assessment in 2002.

The Maryland State Performance Assessment Program (MSPAP) was an entirely performance-based assessment consisting of interdisciplinary performance activities and extended, multi-part tasks. The assessment met standards for reliability and validity at the school level, but was ultimately discontinued because it could not technically and financially provide the individual student score reports required by NCLB.

Background

The Maryland State Performance Assessment Program (MSPAP), first administered in 1991, was part of a larger state education reform effort known as the Maryland School Performance Plan (MSPP). The MSPP sought to revamp education standards, assessment, and accountability. To show its support of MSPP, Maryland committed to increase education spending to 20% of the state's budget.

Strengths

The relatively strong technical quality of MSPAP, coupled with Maryland's solid state leadership, allowed for its approximately 10 years of success. As demonstrated in numerous studies, MSPAP had the psychometric quality needed for high stakes usage at the school level (Ferrara, 2010; Yen, 1997). MSPAP met standards for reliability and validity (MSDE, 1995; NRC, 2010). Tasks were well aligned with Maryland Learning Outcomes (adopted in 1990), and teacher scoring procedures were moderated to monitor reliability (check sets, accuracy sets, spot-checks) (Hambleton, 2000). Additionally, MSPAP development was a process of constant revision; forms were piloted, reviewed, and revised each year to weed out problems and establish technical validity (Yen, 1997). Teachers were deeply involved in task development and scoring (Ferrara, 2010).

State leaders provided consistent backing for MSPAP. Nancy Grasmick, Maryland Superintendent of Schools from 1991 through 2011, was a strong supporter of MSPAP and a popular education official (S. Ferrara, interview, April 24, 2013; S. Marion, interview, March 19, 2013). Additionally, state education officials supported the creation of the Maryland Assessment Consortium, a district-led forum designed to help teachers learn about performance assessment and create tasks that could be shared throughout the state (S. Ferrara, interview, April 24, 2013; J. McTighe, interview, June 8, 2012).

Challenges

Maryland teachers experienced the tension of merging local and state curricular mandates that developed as part of MSPP. As a result, teachers faced the challenge of preparing students for a state assessment while receiving mixed messages about the curricular frameworks driving the assessment (Goldberg, 2000; Koretz et al., 1996). The state did not invest heavily in building local teacher capacity, and instead invested funds in MSPAP's initial task development (Ferrara, 2010; Goldberg, 2000; NRC, 2010). As a result, nearly all professional development for teachers was provided at the local, not state, level (Koretz et al., 1996).

This increased the tension between local and state control.

MSPAP began to experience a public backlash in 2001 due to decreases in MSPAP scores – 21 of 24 school systems earned unusually low scores on the 2001 test (Reilly, 2002). The Maryland State Department of Education reviewed these "unusual statewide decreases" and ruled the scores accurate. In response, local education leaders and boards of education publicly admonished MSPAP and urged Superintendent Grasmick to halt its administration (Ferrara, 2010; Hettinger, 2002; Reilly, 2002).

In response to opposition and minor design flaws, the state began looking into a redesign of MSPAP that would address its weaknesses and make it more affordable (S. Ferrara, interview, April 24, 2013). Ultimately, however, MSPAP was dismantled because it did not meet the requirements of the recently enacted federal No Child Left Behind (NCLB) legislation. NCLB required testing of every student, every year, in grades 3 through 8 and once in high school, and mandated individual student scores to be reported by August. Yearly MSPAP administration (and creation of tasks for every grade) would have been a financial and time burden for the state, and scores could not feasibly be reported by August (Ferrara, 2010; S. Marion, interview, March 19, 2013; NRC, 2010; Parke, 2007). Most importantly, MSPAP did not have the technical capability to provide *individual* student scores for high-stakes usage as demanded by NCLB (MSDE, 1995; S. Marion, interview, March 19, 2013). Instead of revising the assessment, the state decided to discontinue MSPAP and start anew.

MSPAP operated from 1991 until 2002 when it was replaced by a more traditional assessment that could provide individual student scores (Ferrara, 2010; NRC, 2010).

Lessons Learned

Maryland's experience with MSPAP can provide valuable counsel for states or consortia thinking of implementing performance assessment systems.

First, it is vital that leaders ensure political support and open communication among stakeholders and the public concerning new initiatives (Ferrara, 2010). Maryland education officials faced backlash against MSPAP partly due to a lack of understanding of score results and the Maryland Learning Outcomes.

Additionally, it is essential that state officials involve teachers in assessment development and scoring, *and* make a solid investment in building educator capacity (Ferrara, 2010). Professional development should focus on performance-based

instruction and assessment, and teachers should have ample opportunity for collaboration (S. Ferrara, interview, April 24, 2013; Goldberg, 2000).

Finally, it is essential that future assessments meet the technical qualifications necessary to be used for their desired purpose (e.g., individual scores, school-level accountability). States and consortia must find a balance between building assessments that embody the expectations of their state and/or the Common Core standards and those that are affordable and psychometrically feasible.

CHAPTER 5

Recommendations Based on Lessons Learned

The need to re-structure our schools and classrooms to support the acquisition of higher order thinking skills is becoming more urgent every day as the information age is pressuring our educational system to change or be left behind. A principal means to achieve these ends will depend on states and districts moving beyond No Child Left Behind (NCLB) policies to rethinking the current structure of the state and national accountability systems that focus primarily on core facts and recall to new systems of assessment that are able to support the development of deeper learning skills that promote global competence.

Performance-based assessments require students to use high-level thinking to perform, create, or produce something with transferable real-world application. More than standardized tests of content knowledge, such assessments can provide more useful information about student performance to students, parents, teachers, principals, and policymakers. With the introduction of the Common Core State Standards (CCSS) there is now a renewed interest in the development of more balanced assessments that include performance assessment components that are designed to address higher order thinking skills. However, an examination of past initiatives suggests that performance assessment accountability systems are not adopted and implemented without complication. The following are seven key recommendations for

successful performance assessment initiatives.

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 insufficient. Questions concerning technical quality contributed to the phase out of performance-based tasks in large-scale assessment systems. Developers of performance tasks should focus on ensuring that the tasks are designed to fit the technical requirements of their proposed uses. Assessments to be included in high-stakes accountability systems must be valid, reliable, and produce comparable scores. Lessons from Connecticut 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.

First, performance tasks in large-scale assessments must demonstrate **content validity**. Assessment developers working on designing performance tasks should use research-supported frameworks (e.g., Evidence-Centered Design) and content specifications to ensure that each task measures a clear and appropriate set of measurement targets and standards for the appropriate grade level. External panelists with content, bias/sensitivity, and accessibility expertise should also vet performance tasks to ensure that the tasks measure what they are intended to measure and to maintain quality control. The idea of engaging practitioners in the process of designing performance tasks, while supporting buy-in and building local capacity to enact the new standards, may not be fully compatible with the goal of technical quality. This does not suggest that assessment developers should shut practitioners out of the design process, but that the design of any performance tasks used for high-stakes assessment would need to undergo a rigorous design and vetting process, just as more traditional accountability instruments undergo.

Hand-in-hand with content validity, performance tasks in large-scale assessments must demonstrate that they are **comparable and can produce equivalent scores**. In addition to the use of assessment design frameworks like Evidence-Centered Design, task design specifications and "task shells" or templates should be used by all task designers to support comparability of tasks designed to measure the same targets. Before large-scale use, these performance tasks should also undergo pilot testing among

representative student populations (through random assignment of performance tasks to students) and student performance data should be collected to evaluate the extent to which the performance tasks produce similar results in similar student populations. Such design and piloting strategies, already in use by most assessment developers, are promising examples of the progress that the assessment industry has made to maintain the quality and comparability (equivalence) of items, including performance tasks.

In addition, performance tasks in large-scale assessments must contribute positively to test **reliability** by achieving sufficient levels of **inter-rater consistency** in scores for their inclusion in large-scale assessments. High levels of reliability are most likely to be achieved through a distributed scoring method, in which teacher raters do not score the work of the students in their own school. A rigorous training process that does not cut corners, a high calibration standard, and ongoing calibration checks during scoring are also essential for producing reliable and credible scores.

Lastly, assessments that integrate performance tasks should demonstrate **construct-level reliability**. In the past, research suggested that multiple tasks sampling the same learning targets were needed to provide an accurate and consistent estimate of a student's performance. However, this is neither practically feasible nor desirable. Using a variety of item formats (performance tasks, short constructed-response, and selected-response items) to measure the same learning outcome in different ways may overcome this prior limitation. Instead of relying on a single or small number of performance tasks to assess a measurement target, overall reliability of a set of items (comprised of multiple formats) might contribute to more accurate and consistent measurement of target constructs (learning outcomes). This strategy is part of the design for both the SBAC and PARRC assessments, but the results still need to be borne out. In addition, performance task formats should only be used to assess cognitive skills and abilities that cannot be measured in other ways (e.g., ability to research, to construct an evidence-based argument, explain mathematical reasoning).

The demands for technical quality (task comparability and reliability) are relaxed when performance tasks are not intended to produce student-level scores for comparison. Programs such as Wyoming's Body of Evidence and Rhode Island's Graduation Proficiency System are able to include locally designed (district-selected) performance tasks because their scores are not used for ranking schools or to

compare students, but rather to evaluate whether students have met a proficient standard of performance across the curriculum to qualify for graduation. Even within those systems, however, having a common task bank with common tasks that have been vetted by experts and are scored using common criteria for "proficient" performance would bolster the credibility of the system. Regular state monitoring, a peer review process (as was done in the Wyoming BOE system), and state audits of local scores would also support quality control and boost credibility.

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

While studies have shown that human scorers evaluating complex student performances can achieve sufficient rates of inter-rater consistency with high quality training and moderation (Measured Progress, 2009; Pearson, 2011), 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 systems have been used successfully and have supported the sustainability of performance-based assessments (e.g., Nebraska STARS, New York State Regents[30], and Queensland, Australia[31]). These assessment systems were committed to training and certifying teachers as scorers, incentivized teacher participation, and implemented an audit process. Scoring systems can be structured to optimize teacher involvement while still being designed to ensure that teachers do not score their own students' work by using online distributed scoring models and built-in social moderation processes that regularly check for scorer drift, check borderline scores, and adjudicate conflicting scores. This practice is consistent with the guidance in "Operational Best Practices" (CCSSO & ATP, 2010) regarding real-time monitoring of scoring accuracy. As can be seen in the Vermont and Kentucky portfolio programs, involving educators in scoring can help states minimize the cost of scoring performance assessments. And with proper controls, educator-involved scoring can be technically sound.

In addition to reducing costs, involving teachers in scoring performance assessments is a great way to provide important professional development.

[30] NY State Regents has a rich history of local scoring of the Regents that builds into a teachers' workload the resources and time for teachers to be trained and to score performance items on the Regents examinations.

[31] 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.

Teachers routinely report that scoring performance assessments is one of the "best" opportunities to deepen their professional learning. Scoring workshops in which educators score performance assessments as part of their regular professional responsibilities provide an ideal forum for teachers to systematically discuss student results, develop a common language and lens to evaluate student performance, deepen their knowledge and understandings of the new standards and assessments, and share instructional strategies to improve student learning.

In short, involving educators in scoring performance assessments would make the inclusion of performance tasks in large-scale assessments more technically feasible, cost-efficient, and, when implemented thoughtfully, beneficial for both teachers and students.

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. NCLB's yearly testing mandate that, in effect, required states to develop tests for students at nearly every grade level greatly strained already thin state education budgets. As a result, accountability staff in the state departments of education and their contractors began moving away from performance assessment as a viable method to assess student learning and growth. Few states profiled in this study had the resources, support, and perseverance to maintain their use of performance tasks as part of their state assessment programs. Almost all the states we studied eventually had to scale back or eliminate performance tasks because of the financial drain on their assessment resources as well as the NCLB requirement to test all students at every grade level.

Fortunately, there now exist a variety of cost-saving strategies and a burgeoning number of cross-state networks and consortia seeking to bring down the cost of performance assessment. The magnitude of the effect of economies of scale was illustrated by the Brookings Institution study (Chingos, 2012) that estimated that a state with about one million students in grades 3-9 would spend about 35 percent less on assessment than a state of about 100,000 students. 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. As previously mentioned, SBAC estimates that its complete system of assessments will cost less than what two-thirds of its 24 member

states currently spend on state testing.

As noted above, the cost of hand scoring performance assessments can be reduced by establishing educator-involved scoring systems, but it can also be further reduced by investing in the emerging technology of AI (Artificial Intelligence) scoring. The use of AI scoring is not yet fully operational, and additional research is needed to improve the verisimilitude of AI scoring to hand scoring, but it has the potential to significantly decrease the costs associated with scoring essays. Because teachers benefit from participating in scoring performance assessments, however, we do not recommend completely eliminating all hand scoring.

States and consortia should invest in the further development of AI scoring, support research that will improve its capacity to produce valid and reliable scores, and maintain and support some proportion of hand scoring to promote and deepen professional learning.

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

In the current era of test-driven reform, standardized testing has become the policy lever used to drive changes in school and district practices. This theory of action focuses on establishing high stakes for student learning to ensure that the skills and abilities tested drive changes in curriculum and instruction. Predictably, these policies have led to a narrowing of curriculum and instruction that is focused primarily on the basic skills and rote learning that can be assessed on selected-response tests. To prepare students for summative high-stakes exams, some districts have invested in more testing – interim and benchmark assessments – that are designed to focus the curriculum on the same narrow skills and content tested by the state-level summative assessments.

Alternatively, some districts and states are beginning to invest in new kinds of formative assessment practices that include the development of curriculum-embedded performance tasks to evaluate the complex higher order thinking skills and competencies that are identified in the new Common Core State Standards. A more balanced system of assessment, aligned to *the full range* of the CCSS, would include a continuum of item types – from selected-response to constructed-response to technology-enhanced simulations to rich, curriculum-embedded performance tasks – all designed to be tightly connected to the enacted curriculum adopted by the school and district.

One promising approach that uses curriculum-embedded performance assessment to build teacher and

student capacity around deeper learning and to predict student performance in relationship to the CCSS is Ohio's "Learning Dyad" system of assessments. In this model (initially co-developed by the Ohio Department of Education and the Stanford Center for Assessment, Learning, and Equity), rich, extended performance tasks ("learning tasks") are co-developed by teachers and state design teams to be embedded in curriculum, which are designed to have tight alignment to on-demand summative performance items ("assessment tasks") in terms of both the content and disciplinary skills that are measured. The benefit of this dyad learning system is to move beyond "test prep" to build teacher capacity around the CCSS and to provide students with opportunities to learn content more deeply, while at the same time producing comparable scores on "assessment tasks" that have been designed to be technically defensible, comparable measures. In this process, the formative use of rich "learning tasks" serves to open up the curriculum and support innovative task types rather than narrowing the curriculum by using interim assessments that are the mirror image of what is tested on the state summative tests.

Developing a comprehensive and coherent system of standards, assessment, and instruction to support deep learning should also include the development of the following resources and processes:

- **Curricular resources** aligned to the desired state or local learning outcomes and the assessments. Teachers should have access to instructional resources (and accompanying professional learning opportunities) that will enable them to align their instruction with new standards and assessments, and will allow them to appropriately prepare students for the assessment with a focus on cultivating deep knowledge and skills, not basic test preparation.
- **Protocols and processes to provide "just in time" feedback** to developers of curricula, curriculum-embedded assessments, and instructional modules, including putting in place technology-enhanced peer review processes to evaluate and certify the quality of the newly developed resources.
- **Data reporting systems of student learning** that are structured to include multiple sources of evidence about student learning in relation to the standards. These reporting systems could produce a learning profile that also includes opportunities to capture student self-assessment and reflection on their progress, as well as

teacher observations of learning needs.

One example of a scaled-up initiative to build teacher capacity to implement the CCSS is the Literacy Design Collaborative (LDC), supported by the Bill and Melinda Gates Foundation. LDC, which leverages participation across all levels of the system (classroom, school, district, and state) focuses on building teacher capacity to design rich classroom-based writing assignments, coupled with the development of instructional modules that build students' literacy and writing skills aligned with the demands of the culminating writing task. The LDC system focuses intently on classroom practice through the design and development of technology-embedded templates, tools, and support structures to assist teachers in building CCSS-aligned literacy and writing tasks to support literacy instruction across the curriculum. A parallel initiative – the Mathematics Design Collaborative, also supported by the Gates Foundation – has been working on building the capacity of mathematics teachers to implement and analyze student responses to rich mathematics tasks, and to use that information to guide their instructional decisions.

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

There are two polarized views about the best strategies for implementing new standards and assessments. Due to the high stakes associated with NCLB testing requirements, districts often have resorted to implementing interim and benchmark assessments that are tightly aligned to accountability measures, with the theory of action that these assessments will provide formative information to teachers that will help them make better instructional decisions to support student learning. So far, there is little research that supports the idea that such interim assessments actually improve instructional practice, support student learning, or close the achievement gap. An alternate view of formative assessment is beginning to emerge as part of the CCSS initiative and the Common Core assessment consortia. Both SBAC and PARCC include within their systems a digital library of formative assessment instruments, model curricula, and embedded performance tasks that are designed to rethink and reshape the nature of formative assessment to support CCSS-aligned instruction and improve student performance on the on-demand summative assessments. The creation of digital libraries of formative assessment instruments, 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.

More broadly, to implement the CCSS and to transform teaching and learning in ways that support college and career readiness, we need to incentivize the building of a wide range of vetted CCSS-aligned curriculum, instruction, assessment, and professional learning resources to support and transform teaching and learning at the classroom level. 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 student learning and performance.

To illuminate best practice in new forms of performance assessment and curriculum development, states that have adopted the CCSS should create a cross-state collaborative electronic platform to share resources, information, and best practices. The teaching resource bank should be dynamic, nimble, and flexible to accommodate and harvest high quality work and resources as they are created in real time at the local, state, and/or national level.

Developing a capacity-building learning system that privileges collaboration and fosters and nurtures engagement of networks could include the development of the following tools and processes:

- **Development of a curated, open-source performance assessment task bank.** Performance tasks entered into the task bank should be certified as representing high quality, CCSS-aligned assessments, and should include a continuum of task types that show what students know and are able to do. The task bank should be structured to include performance task bundles that are indexed based on their alignment to grade-level standards, as well as rubrics and task shells to guide the development of new tasks, and curriculum modules that support student learning in relation to the full range of standards.
- **Development of technology-enhanced tools and protocols** to support teacher design of CCSS-aligned classroom assignments, curriculum materials, and resources. The technology would focus on actively engaging teachers as partners in the development of assessments that are integrated into the curriculum

and deepen student knowledge and understandings.

- **Development of technology-enhanced processes to train and certify scorers** using rubrics or scoring guides to evaluate student work. The scoring system should allow for use of a social moderation process that includes recruiting expert teachers and/or expert panels to audit the results of local scoring processes to ensure the validity of the assessment systems.

An electronic platform as described above would allow CCSS networks to carry out rapid prototyping of curriculum and assessment strategies to support local learning and to share results and promising practices with network partners nationwide. Networks that are working on the same set of problems but through different means could utilize the platform to support productive engagement by bringing educators with differing perspectives together to work towards common goals. The momentum of the groundbreaking collaboration among states to establish and adopt the Common Core State Standards and create assessments should be leveraged and maintained through the use of technology as the states move towards implementing and supporting teachers in the adoption of the new standards and practices.

The Innovation Lab Network (ILN)[32], a collaborative of state education leaders and policymakers dedicated to designing assessment and accountability systems resulting in deeper forms of student learning, is ideally positioned to design and develop a dynamic interactive clearinghouse of promising practices and resources by leveraging and expanding upon existing state and school networks. The clearinghouse would be a collection of adaptable tools and resources to support deeper learning that have passed through an exhaustive and comprehensive auditing process to certify the quality and adaptability of performance tasks, curriculum resources, and training/scoring protocols. System portals would provide access to a wide range of vetted tools such as scorer training and hand scoring tools; embedded curriculum modules aligned to the CCSS across content areas; research-based instructional strategies to support student growth; and a performance task bank accompanied by evidence supporting reliability and validity of the assessments.

In past initiatives, performance assessment was more of a cottage industry where both the assessments and tools to support

[32] A working sub-group of the Council of Chief State School Officers.

the work were either not developed or available or shared. Without the development of a science around performance assessment the quality and defensibility of the system could again become vulnerable and untrustworthy. We need, this time around, vetted and research-based tools that are accessible and open-sourced to provide the foundation for building teacher capacity around teaching and learning. For example, Ontario, Canada, has developed a data analysis tool called Ontario Statistical Neighbours (OSN) to assist boards and schools in using multiple data sources to improve student achievement (OME, 2007). These types of tools could be made widely available and improved upon.

Today, if we want to find the best practices in performance assessment, we are relegated to hunting all over the Internet for viable materials, but are often met with materials of variable quality and relevance. There is a better way. There are a few web-based electronic portals that currently exist for educators to access vetted, high quality resources, such as teachinghistory.org, a clearinghouse of expert- and educator-vetted resources for K-12 history teachers sponsored by the National History Education Clearinghouse. Approximately one million visitors browse the site and download resources each year, demonstrating the power and clear demand for a centralized portal of quality tasks and curricular resources (D. Martin, personal communication, September 18, 2013). The use of technology across networks could dramatically increase the quality of performance assessments and, if smartly integrated, could greatly reduce development and implementation costs, as well as build the capacity of teachers and administrators implementing the CCSS.

In summary, as states look to increase their use of performance assessment and implement the CCSS, they should capitalize on cutting edge technological innovations and the power of network collaboration to accelerate the development and sharing of high-quality resources that support high-quality teaching and deeper student learning.

6. *Engage with stakeholders more actively, 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. Anchoring innovation in real images of practice is essential to maintaining long-term support for a performance assessment program.

Unfortunately, another lesson learned that also grew out of the experience of implementing innovative assessment systems is that political leadership supporting these reforms is not enough. Champions move on, leadership changes, and political and educational priorities evolve in ways that can undermine or dilute changes in the reform. The challenge is how to build and sustain the state's organizational capacity to adapt to new policy directions and political challenges that are directed at the reform. Of course, federal or state policies can strike at the heart of the reform as evidenced by the role NCLB played in the phasing out or elimination of performance assessment in many of the innovative assessment systems highlighted in this monograph.

What is needed to sustain innovation is to develop the organizational capacity of educational leaders at all levels of the system – state, district, and school – to engage in on-going communication, evidence gathering, and problem solving. One promising practice that highlights this approach is the "improvement science" framework developed by Anthony Bryk of the Carnegie Foundation for the Advancement of Teaching. Improvement science is grounded in a methodological approach associated with action research that fosters both deep insight into clinical practice and is contextually responsive to local and state policy frameworks. It promotes multiple rapid tests of possible program or policy changes by a range of individuals working on the same problem under different conditions. This approach is designed to capitalize on building "problem solving networked communities" that provide multiple perspectives and evidence to support learning about problems of practice. In the information age, dissemination of evidence-based information that promotes critical dialogue and interaction around substantive issues may be a powerful contributor to achieving more sustainable and scalable policy changes. The challenge is to build an infrastructure and human and social capital in ways that support adjustment to new policies

through the testing, re-testing, and sharing of intelligent practices.

The following approaches for engaging with the public, and for building the capacity of state leaders and policymakers are offered for consideration:

- Develop an interactive web-based portal to provide up-to-date information and resources to support understanding and adoption of new initiatives by policymakers, educational leaders, teachers, and the greater public. One such effort recently launched by the Maine Department of Education is illustrative of a technology-enhanced portal that is specifically designed to inform policymakers, schools and districts, and teacher leaders about new directions in curriculum and assessment. The Maine platform, *Getting To Proficiency: Helping Maine Graduate Every Student Prepared,* provides useful resources, tools, and guidance for administrators, school leaders, educators, parents and community leaders who are working to implement a proficiency-based system to support high school graduation based on the state-adopted learning standards. The website is rich with resources to support policymakers and educational leaders with information, resources, tools, protocols, and images (video and multi-media presentations) of effective practice to use with a wide range of stakeholders and varied audiences. Of particular note is a section identified as "Design for Learning", which provides a wide range of resources that are designed to support the development of a professional learning culture that is both respectful and responsive to local context and state policy and practices. Building a culture of respect does not just happen – it is created through establishing a set of shared knowledge, shared tools and practices, shared experiences, and shared beliefs. Moreover, the Maine portal recognizes that assessment reform is multi-dimensional and engages all stakeholders in understanding new directions in assessment in relationship to curriculum, instruction, and school and district change. Building system coherence around the work is essential to sustaining and deepening understandings around reform initiatives.
- Incentivize and establish an Education Policy Fellows program to support the on-going learning of state policymakers and leaders. Currently the nation is engaged in a state and national dialogue about reforms in teacher

preparation, teacher evaluation, standards and accountability, and the CCSS, just to mention some of the notable topics of the day. Conferences, institutes, and publications (including social media) are the traditional pathways used to discuss and share knowledge, research, and varied perspectives on proposed reforms. A more systematic approach to build capacity and deepen the knowledge of policymakers within states should be considered. An Education Policy Fellows program would regularly bring together key educational and policy leaders within states to network and learn about research-based educational policy options and practices. What distinguishes this approach is the opportunity to engage policymakers and leaders in an on-going way in critical dialogue about reform options related to education policy. Through this forum, Education Policy Fellows would have access to experts and educators with specialized knowledge to deepen their own knowledge and understandings about specific reforms. As policymakers leave their political offices, new members would be nominated to sustain a critical mass of state leaders engaged in a professional learning community around educational reform.

- Incentivize and support opportunities for job-embedded residency experiences that educators and policy leaders can access to deepen their knowledge and understanding of curriculum, instruction, and assessment. "Residency" in this case is conceptualized as on-site deep engagement with schools or organizations to experience first-hand new approaches to curriculum, instruction, and assessment. For example, the Stanford Center for Opportunity Policy in Education (SCOPE) has developed a school-based residency program where educators around the country come together over three days to learn about performance assessment systems that are designed to prepare students for success in college and career. School sites that participate are selected because they have both a proven track record of high performance and a culture of learning that is committed to sharing practice to support more equitable outcomes for all students. These residencies are designed to be highly interactive where participants have access to students, teachers, administrators, and parents to better understand

the nature and effectiveness of the school practices and processes.

A variation of the residency approach is to place educational leaders in residence with professional organizations that are working on problems of practice directly related to policy options that are under consideration. For example, state department leaders of curriculum and/or assessment could participate in a residency where they work side by side with professional organization staff, colleagues, and partners to deepen their knowledge and understandings of the work as well as to contribute to the work. One could imagine residencies that include placement in schools, districts, regional labs, regional service centers, university centers, think tanks, and centers of professional practice (e.g., the Writing Project). Enacting these approaches to broker more embedded learning opportunities for educational leaders and policymakers can enhance the efficacy of educational organizations and provide a forum for developing human and social capital at scale that is both sustainable and adaptable to the ever changing policy landscape.

7. Engage with the public more actively, 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 enhancing student learning and achieving college and career success. 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.

The lesson learned from our examination of performance assessment initiatives of the 1990s is that communication to the public needs to be continuous, responsive, and differentiated to reach a broad

base of constituents that have a stake in the education of our children. The development of a comprehensive communication plan that actively engages the public in dialogue around the new assessment can help avert any confusion or backlash that might occur as states begin using new forms of assessment and teachers begin to change their instructional practices. States should organize to take advantage of the various social media tools presently available to reach the public – online news, editorials, blogs, MOOCs, Twitter, YouTube, Facebook, etc. – to communicate the purpose and vision behind the Common Core State Standards and aligned assessments. Furthermore, we should not ignore the power of building supportive constituencies through on-going face-to-face interactions with key stakeholders to address their concerns, needs, and possible misconceptions.

We must significantly broaden our circle of communication, moving beyond constituencies that share our viewpoint to building a coherent on-going learning culture that has a shared knowledge, understanding, and common language around the meaning, purpose, and value of performance assessments and their implications for learning. In short, communicate, communicate, communicate.

REFERENCES

Abruscato, J. (1993). Early results and tentative implications from the Vermont portfolio project. *Phi Delta Kappan,* 74(6), 474-477.

Airasian, P.W., & Madaus, G.F. (1983, June). Linking testing and instruction: Policy issues *Journal of Educational Measurement,* 20(2), 103-118.

Archer, J. (2005). R.I. Downplays Tests as Route To Diplomas. *Education Week,* 24(31), 1-25.

Bagwell, M.M. (2005). Differences in Connecticut Mastery Test scores and student perceptions of school quality between public and charter schools. (Doctoral dissertation). Retrieved from University of Connecticut Library.

Balakrishnan, A. (2001, August 1). Assessing MSPAP: Who gets scores for what? *The Baltimore Chronicle*. Retrieved from http://baltimorechronicle.com/mspap_aug01.html

Baron, J.B. (1989). Performance testing in Connecticut. *Educational Leadership, 46*(7), 8.

Baron, S. B. (1996). Developing performance-based assessments: The Connecticut experience. In S. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 166-191). Chicago, IL: The University of Chicago Press.

Behuniak, P., & Tucker, C. (1992). The potential of criterion-referenced tests with projected norms. *Applied Measurement In Education,* 5(4), 337-353.

Black, S. (1993). Portfolio assessment. *Executive Educator*, 15(1), 28-31.

Borko, H., & Elliott, R. L. (1999). Hands-on pedagogy versus hands-off accountability: tensions between competing commitments for exemplary math teachers in Kentucky. *Phi Delta Kappan,* 80(5), 394-400.

Briars, D.J., & Resnick, L.B. (2000). *Standards, assessments - and what else? The essential elements of standards-based school improvement*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), CSE Technical Report 528.

Burnham, J.J. (2013, May 24). Thomaston superintendent favors new statewide school testing. *Foothills Media Group*. Retrieved from www.foothillsmediagroup.com

Carlos, L., & Kirst. M. (1997). California curriculum policy in the 1990s: "We don't have to be in front to lead". Paper presented at the Annual Meeting of the American Educational Research Association. Chicago, IL. March 24, 1997.

Cech, S.J. (2008). Showing what they know: In Rhode Island, performance-based assessments are now required for high school graduation. *Education Week,* 24(42), 25-27.

Celock, J. (2013, March 20). Cindy Hill handed setback in Wyoming lawsuit seeking restored education boss powers. *The Huffington Post.* Retrieved from http://www.huffingtonpost.com/2013/03/20/cindy-hill-wyoming-lawsuit_n_2920376.html

Center on Education Policy. (2011). *Profile of State High School Exit Exam Policies: Rhode Island.* Washington, DC: S. McIntosh.

Center on Education Policy. (2012). *State high school exit exams: A policy in transition.* Washington, DC: S. McIntosh.

Chingos, M.M. (2012). *Strength in numbers: State spending on K-12 assessment systems*. Washington, DC: Brown Center on Education Policy at Brookings.

Chung, G.K.W.K., & Baker, E.L. (2003). *The impact of a simulation and problem-based learning design project on student learning and teamwork skills*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), CSE Technical Report 599.

Chrispeels, J. H. (1997). Educational policy implementation in a shifting political climate: The California experience. *American Educational Research Journal,* 34, 453-481.

Cohen, D.K., & Hill, H.C. (1998). Instructional policy and classroom performance: The mathematics reform in California. Consortium for Policy Research in Education (CPRE), Research Report Series RR-39.

Common Core State Standards Initiative (2014). About the Standards. Retrieved on March 25, 2014 from: http://www.corestandards.org/about-the-standards/

Commission on Behavioral and Social Sciences and Education. (1995). New Standards Project (benchmark activities). In *International comparative studies in education: Descriptions of selected large-scale assessments and case studies* (pp. 74-78). Washington, DC: National Academies Press.

Connecticut State Board of Education. (2012). *Connecticut Academic Performance Assessment: Third Generation Program Overview.*

Connecticut State Department of Education. (2013a). Website. Content last modified 1/25/13. Accessed 6/3/13. Retrieved from http://www.sde.ct.gov/sde

Connecticut State Department of Education. (2013b). Connecticut Mastery Test (Fourth
Generation) English Language Arts Handbook. Retrieved from: http://www.sde.ct.gov/sde/lib/sde/pdf/curriculum/language_arts/languageartshandbook-part5.pdf

Consortium for Policy Research in Education (CPRE). (2000). *Assessment and accountability in the fifty states: 1999-2000: Connecticut assessment and accountability profile*. Washington, DC.

Council of Chief State School Officers & Association of Test Publishers. (2010). *Operational best practices for statewide large-scale assessment programs*. Washington, DC.

Cramer, P. (2013, August 6). Arne Duncan steps in to assuage fears about N.Y. test scores. GothamSchools.org. Retrieved from http://ny.chalkbeat.org/2013/08/06/arne-duncan-steps-in-to-assuage-fears-about-n-y-test-scores/#.VDih3PldWSo

Cronbach, L.J., Bradburn, N.M., & Horvitz, D.G. (1994). *Sampling and statistical procedures used in the California Learning Assessment System*. Report of the Select Committee, State of California.

Cumber, C.G. (2002, September 19). State to replace troubled MSPAP. *The Frederick News-Post*. Retrieved from http://www.fredericknewspost.com/archives/article_5167b827-e939-5a69-a2ed-7ddd013c6101.html

Curtis, A. (2013, May 21). Hill talks politics, education policy. *Wyoming Tribune Eagle*. Retrieved from http://www.wyomingnews.com/articles/2013/05/21/news/01top_05-21-13.txt

Danitz, T. (2001, January 27). States pay $400 million for tests in 2001. *Stateline*. Retrieved from http://www.pewstates.org/projects/stateline/headlines/special-report-states-pay-400-million-for-tests-in-2001-85899393054

Darling-Hammond, L. (2010). *Performance counts: Assessment systems that support high-quality learning*. Washington, DC: Council of Chief State School Officers.

Darling-Hammond, L. & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., Adamson, F. (2013). *Developing assessments of deeper learning: The costs and benefits of using tests that help students learn*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L. & Falk, B. (2013). *Teacher learning through assessment: How student-performance assessments can support teacher learning*. Washington DC: Center for American Progress.

DiMartino, J. (2007, April 25). Accountability or mastery? The assessment trade-off that could change the landscape of reform. *Education Week,* 26(34), 36, 44.

DiMartino, J., & Teixeira, D. (2005). Demonstrating Success. *Principal Leadership: High School Edition,* 6(2), 32-36.

Dowding, S.K. (2011). An examination of development of Wyoming's alternative assessment system, the Body of Evidence. (Doctoral dissertation). Retrieved from Montana State University Library.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied measurement in education,* 4(4), 289-303.

Ferrara, S. (2010). The Maryland School Performance Assessment Program (MSPAP) 1991-2002: Political considerations. Paper presented at National Research Council Workshop. Washington, DC. December 10-11, 2009.

Firestone, W.A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis,* 20, 95-113.

Fontana, J. (1995). Portfolio assessment: Its beginnings in Vermont and Kentucky. *National Association of Secondary School Principals Bulletin,* 79(573), 25.

Foote, M. (2005). *The New York Performance Standards Consortium: College Performance Study*. New York: New York Performance Standards Consortium.

Gao, X., Shavelson, R.J., & Baxter, G.P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education,* 7, 323-342.

Gelb, J. (2001, April 17). *Connecticut Mastery Test*. OLR Research Report: Document 2001-R-0389.

Gewertz, C. (2013, July 22). PARCC test cost: High for nearly half the states. [Web log comment]. Retrieved from http://blogs.edweek.org/edweek/curriculum/2013/07/parcc_test_cost_higher_for_half_.html

Goldberg, G., & Roswell, B. (2000). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment*, 6(4), 257.

Gong, B. & Reidy, E. F. (1996). Assessment and accountability in Kentucky's school reform. In S. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 215-233). Chicago, IL: The University of Chicago Press.

Guthrie, J.T., Almasi, J.F., Schafer, W.D., & Afflerbach, P.P. (1994). Policies for integrated reading instruction related to a state-wide improvement program. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA. April 8, 1994.

Haertel, E. (1999). Performance assessment and education reform. *Phi Delta Kappan,* 80(9), 662-666.

Haertel, E., & Calfee, R. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement,* 20(2), 119-132.

Hambleton, R. K., Impara, J., Mehrens, W., & Plake, B.S. (2000). *Psychometric review of the Maryland school performance assessment program*. Psychometric Review Committee, State of Maryland.

Hanson, G.M.B. (1994, August 1). Testing the learning curve in California's schools. *Insight on the News.* Retrieved from http://www.highbeam.com/doc/1G1-15674648.html

Hardy, R. A. (1995). Examining the costs of performance assessment. *Applied Measurement In Education,* 8(2), 121.

Hayes, J. (2010, March 31). *Connecticut Mastery Test's measurement of grade level skills*. OLR Research Report: Document 2010-R-0167.

Herman, J. (1997). Assessing new assessments: How do they measure up? *Theory Into Practice,* 36(4), 196-204.

Herman, J.L., Aschbacher, P.M., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Herman, J.L., Klein, D.C., & Wakai, S.T. (1996). *Assessing equity in alternative assessment: An illustration of opportunity-to-learn issues*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), CSE Technical Report 440.

Herman, J.L., & Linn, R. (2013). *On the road to assessing deeper learning: The status of Smarter Balanced and PARCC assessment consortia*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), CSE Technical Report 823.

Hettinger, J. (2002, March 6). Eighth-graders likely to skip MSPAP tests. *The Gazette.* Retrieved from http://www.gazette.net

Hewitt, G. (2001). The writing portfolio: Assessment starts with A. *Clearing House,* 74(4), 187-190.

Hill, R. (2000). A success story from Kentucky. Paper presented at the Annual National Conference on Large-Scale Assessment, Council of Chief State Schools Officers. Snowbird, UT. June 25-28, 2000.

Hill, R., & Reidy, E. (1993). *The cost factors: Can performance based assessment be a sound investment?* Manuscript submitted for publication.

Hout, M. & Elliott, S.W. (Eds.) (2013). *Incentives and test-based accountability in Education.* National Research Council of the National Academies. Washington, DC: National Academy Press.

Idaho State Department of Education Division of Assessment and Accountability. (2014). *Smarter Balanced operational 2015 updates* [PowerPoint slides].

Jones, K., & Whitford, B. (1997). Kentucky's conflicting reform principles: high-stakes school accountability and student performance assessment. *Phi Delta Kappan,* 79, 276-281.

Kirst, M.W., & 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., McCaffrey, D., Stecher, B., & Koretz, D. (1995). The reliability of mathematics portfolio scores: Lessons from the Vermont experience*. Applied Measurement In Education,* 8(3), 243.

Koretz, D.M. (1998). Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education,* 5(3), 309-334.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources,* 37(4), 752-777.

Koretz, D. & Barron, S. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND Corporation.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., Stecher, B. (1992). *Interim Report: The reliability of scores from the 1992 Vermont Portfolio Assessment Program.* National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/ RAND Institute on Education and Training.

Koretz, D.M., Barron, S., Mitchell, K.J., & Stecher, B.M. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND Corporation.

Koretz, D., Mitchell, K., Barron, S. & Keith, S. (1996). *Final Report: Perceived effects of the Maryland performance assessment program*. National Center for Research on Evaluation, Standards and Student Testing (CRESST)/RAND Institute on Education and Training, CSE Technical Report 409.

Kurz, J. (2001). Open-ended inquiry. *The Science Teacher,* 68(1), 62-66.

Lane, S. (2010). *Performance assessment: The state of the art.* (SCOPE Student Performance Assessment Series). Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Liftig, I. (2006). Do state standardized science tests bring out the Chicken Little in us? *Science Scope,* 30(2), 6.

Linn, R.L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher,* 23(9), 4-14.

Linn, R. L. (1998). Assessment and accountability. Paper presented at the Annual Meeting of the American Educational Research Association. San Diego, CA. April 1998.

Madaus, G. F. (1988). The Influence of Testing on the Curriculum. In Tanner, L. N. (Ed.), *Critical Issues in the Curriculum*, pp. 83-121. 87th Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., Viator, K. A. (1992). The Influence of Testing on Teaching Math and Science in Grades 4-12 (SPA8954759). Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Markoff, J. (2013, April 4). Essay-Grading Software Offers Professors a Break, *New York Times.* Retrieved from: http://www.nytimes.com/2013/04/05/science/new-test-for-computers-grading-essays-at-college-level.html?pagewanted=all&_r=0

Marion, S.F. & Stevens, S. (2001). *The Wyoming assessment handbook*. Wyoming Department of Education.

Marion, S. F., Sheinker, A., Hansche, L., & Carlson, D. (1998). *Wyoming Comprehensive Assessment System design report to the Wyoming state legislature*. Report of the Statewide Assessment Design Team, State of Wyoming.

Maryland State Department of Education (MSDE). (1995). *Technical report: 1995 Maryland school performance assessment program (MSPAP)*.

Matthews, B. (1995). *The implementation of performance assessment in Kentucky classrooms.* Louisville, KY: University of Louisville.

Matthews, K. (2013, April 14). Some NY parents to boycott new, harder state tests (Associated Press). *Wall Street Journal* online. Retrieved from: http://online.wsj.com/article/APb5789c1bb6184ee49df2544f5bd7cb9a.html

McAuliffe, R. (2007). *Defining education proficiency and achievement in Connecticut.* New Haven, CT: Connecticut Voices for Children.

McDonnell, L.M. (1994). Assessment policy as persuasion and regulation. *American Journal of Education,* 102, 394-420.

McDonnell, L.M. (2004). *Politics, persuasion, and educational testing*. Cambridge, MA: Harvard University Press.

McDonnell, L.M., & Weatherford, M.S. (2013a). Evidence use and the Common Core State Standards movement: From problem definition to policy adoption. *American Journal of Education,* 120(1), 1-25.

McDonnell, L.M., & Weatherford, M.S. (2013b). Organized interests and the Common Core. *Educational Researcher,* 42(9), 488-497.

Measured Progress. (2009). *New England Common Assessment Program: 2008-2009 Technical Report*. Dover, NH.

Michaels, H., & Ferrara, S. (1999). Evolution of educational reform in Maryland: Using data to drive state policy and local reform. In G. J. Cizek (Ed.), *Handbook of Educational Policy*. San Diego: Academic Press.

Mills, R. P. (1996). Statewide portfolio assessment: The Vermont experience. In S. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 192-214). Chicago, IL: The University of Chicago Press.

Mislevy, R.J., Almond, R.G., & Lukas, J.F. (2003). A brief introduction to evidence-centered design. Educational Testing Service (ETS), Research Report RR-03-16.

Monk, D.H. (1993). The costs of systemic education reform: Conceptual issues and preliminary estimates. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA. April 4-8, 1994.

Moore, B. (2013). Mitigating bandwidth limitations for online assessments. RJM Strategies, LLC. Retrieved from: http://www.k12blueprint.com/sites/default/files/Mitigating%20Bandwidth%20Limitations%20-%20BMoore%20Analyst%20Report%20v2.pdf

Murphy, S., Bergamini, J., & Rooney, P. (1997). The impact of large-scale portfolio assessment programs on classroom practice: Case studies of the New Standards field-trial portfolio. *Educational Assessment,* 4(4), 297-333.

National Research Council. (2003). *Assessment in support of instruction and learning: bridging the gap between large-scale and classroom assessment*. Washington DC: Committee on Assessment in Support of Instruction and Learning.

National Research Council. (2010). *State assessment systems: Exploring best practices and innovations: Summary of two workshops*. Washington, DC: Beatty.

Nichols, S. and Berliner, D. (2005). The Inevitable Corruption of Indicators and Educators through High-Stakes Testing. Tempe, Arizona: Education Policy Studies Laboratory, Arizona State University. Retrieved from: http://nepc.colorado.edu/publication/the-inevitable-corruption-indicators-and-educators-through-highstakes-testing

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

O'Day, J. & Smith, M. (1993). Systemic reform and educational opportunity. In Fuhrman, S. H. (Ed.), *Designing Coherent Education Policy: Improving the System,* pp. 250-312. NY: Jossey-Bass.

O'Neil, J. (1990, September). New curriculum agenda emerges for the '90s. *Association for Supervision and Curriculum Development: Curriculum Update*, 1-8.

O'Neil, J. (1993). On the New Standards Project: A conversation with Lauren Resnick and Warren Simmons. *Educational Leadership,* 50(5), 17-21.

Ontario Ministry of Education. (2007). *Ontario Statistical Neighbours: Informing our strategy to improve student achievement*. Ontario, Canada: The Literacy and Numeracy Secretariat.

Parke, C. S., Lane, S. (2007). Student's perceptions of a Maryland state performance assessment. *The Elementary School Journal*, 107(3), 305-324.

Partnership for Assessment of Readiness for College and Careers (PARCC). (2013). Website. Accessed 8/27/13. Retrieved from parcconline.org/assessment-cost-estimates

Pearson. (2011). *National Board for Professional Teaching Standards technical report: Technical quality of scores*. New York, NY.

Pecheone, R., Kahl, S., Hamma, J., Jaquith, A. (2010). *Through a looking glass: Lessons learned and future directions for performance assessment.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Porcello, D. & Hsi, S. (2013). Crowdsourcing and curating online education resources. *Science, 341*(6143), 240-241.

Quenemoen, R. (2008). A brief history of alternate assessments based on alternate achievement standards (Synthesis Report 68). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from: http://www.cehd.umn.edu/NCEO/OnlinePubs/Synthesis68/index.htm#current status

Ravitch, D. (2013a, Feb 26). Why I cannot support the Common Core State Standards.
Retrieved from: http://dianeravitch.net/2013/02/26/why-i-cannot-support-the-common-core-standards/

Ravitch, D. (2013b, July 22). Rethinking schools: The trouble with the Common Core. Retrieved from: http://dianeravitch.net/2013/07/22/rethinking-schools-the-trouble-with-common-core/

Reilly, T. (2002, February 18). Board of Education reserves judgment on MSPAP. *The Herald-Mail.* Retrieved from http://www.herald-mail.com

Reitz, S. (2011, February 22). *Connecticut loses 'No Child Left Behind' legal challenge*. Associated Press. Retrieved from

http://www.nbcnews.com/id/41723439/ns/us_news-crime_and_courts/t/connecticut-loses-no-child-left-behind-legal-challenge/#.Uh-ROndRx8E

Resnick, L.B., Nolan, K.J., & Resnick, D.P. (1995). Benchmarking education standards. *Educational Evaluation and Policy Analysis,* 17(4), 438-461.

Rhode Island Department of Education. (2013). Website. Accessed 5/15/13. Retrieved from http://www.ride.ri.gov/StudentsFamilies/RIPublicSchools/DiplomaSystem.aspx

Schmidt, W.H. (1983, June). Content biases in achievement tests. *Journal of Educational Measurement,* 20(2), 165-178.

Shavelson, R., Baxter, G.P., & Pine, J. (1991). Performance Assessment in Science. *Applied Measurement in Education, 4*(4).

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments.*Journal of Educational Measurement,* 30(3), 215-232.

Shavelson, R.J., Ruiz-Primo, M.A. & Wiley, E.W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement,* 36(1), 61-71

Shepard, L.A. (1990). Inflated Test Score Gains: Is the Problem Old Norms or Teaching the Test? *Educational Measurement: Issues and Practice* 9 (3), 15-22.

Shepard, L.A., Flexer, R.J., Hiebert, E.H., Marion, S.F., Mayfield, V., & Weston, T.J. (1996). Effects of introducing classroom performance assessments on student learning. *Educational Measurement: Issues and Practice*, 15, 7-18.

Simmons, W., & Resnick, L. (1993). Assessment as the catalyst of school reform. *Educational Leadership*, 50(5), 11-15.

Smarter Balanced Assessment Consortium (SBAC). (2012). Website. Accessed 8/26/13. Retrieved from http://www.smarterbalanced.org

Smith, M. L. (1991). Put to the Test: The Effects of External Testing on Teachers. *Educational Researcher,* 20(5), 8-11.

Spalding, E. (1995). The New Standards Project and English language arts portfolios: A report on process and progress. *Clearing House,* 68(4), 219.

Spalding, E. (2000). Performance Assessment and the New Standards Project: A story of serendipitous success. *Phi Delta Kappan,* 81(10), 758-764.

Spalding, E. & Cummins, G. (1998). It was the best of times. It was a waste of time: University of Kentucky students' views of writing under KERA. *Assessing Writing,* 5(2), 167-199.

Stage, E.K. (2007). Perspectives on state assessments in California: What you release is what teachers get. *Assessing Mathematical Proficiency,* 53, 357-363.

Stecher, B. (1998). The local benefits and burdens of large-scale portfolio assessment. *Assessment in Education,* 5(3), 335-351.

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Stecher, B.M., & Mitchell, K.J. (1995). *Portfolio-driven reform: Vermont teachers' understanding of mathematical problem solving and related changes in classroom practice*. RAND Institute on Education and Training, CSE Technical Report 400.

Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey of Kentucky teachers of mathematics and writing*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/RAND, CSE Technical Report 482.

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student and school variables. *Applied Measurement in Education,* 16(1), 1-26.

Strong, S., & Sexton, L. C. (1996). Kentucky performance assessment of reading: valid? *Contemporary Education,* 67, 102-106.

Strong, S., & Sexton, L. C. (1997). Kentucky performance assessment of mathematics: do the numbers add up? *Journal of Instructional Psychology,* 24, 202-206.

Topol, B., Olson, J., Roeber, E., & Hennon, P. (2013). *Getting to higher-quality assessments: Evaluating costs, benefits, and investment strategies*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Tung, R. & Stazesky, P. (2010). *Including performance assessments in accountability systems: A review of scale-up efforts.* Boston, MA: Center for Collaborative Education.

Ujifusa, A. (2013, August 19). N.Y. Test-Score Plunge Adds Fuel to Common-Core Debate. Education Week online. Retrieved on August 30, 2013 from: http://www.edweek.org/ew/articles/2013/08/21/01newyork.h33.html?qs=New+York+tests

Unknown Author. (1994, March 9). O.C. California learning assessment system scores: School tests: A guide to gables. *The Los Angeles Times.* Retrieved from http://articles.latimes.com/1994-03-09/local/me-33694_1_california-learning-assessment-system

Unknown Author. (1994, May 14). State to release controversial CLAS tests. *Los Angeles Times*. Retrieved from http://articles.latimes.com/1994-05-14/news/mn-57654_1_clas-test

United States Department of Education (USDOE). (2008, February 4). Fiscal year 2009 budget summary. Retrieved from: http://www2.ed.gov/about/overview/budget/budget09/summary/edlite-section1.html

United States Department of Education (USDOE). (2013, October 30). Education department budget history table: Fiscal year 1980-2014. Retrieved from: http://www2.ed.gov/about/overview/budget/history/edhistory.pdf

United States Department of Education (USDOE). (2013, March 13). Grants for enhanced assessment instruments. Retrieved from: http://www2.ed.gov/programs/eag/awards.html

Vogler, K. E. (2002). The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. *Education,* 123(1), 39.

Vu, P. (2008, January 17). Do state tests make the grade? *Stateline*. Retrieved from http://www.pewstates.org/projects/stateline/headlines/do-state-tests-make-the-grade85899387452

Weick, K. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly,* 21, 1-9.

White, K. A. (1998). Ky. bids KIRIS farewell, ushers in new test. *Education Week,* 17 (32), 16.

Wiley, D.E., & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis,* 17, 355-370.

Willhoft, J. (2013). *The future is (almost) now: Implementing Smarter Balanced assessments in 2014-15* [PowerPoint slides]. Retrieved from *https://ccsso.confex.com/ccsso/2013/webprogram/Presentation/Session3700/3%20Willhoft_201321_NCSA_SmarterBalanced%20Sustainability%5B2%5D.pdf*

Williamson, D.M., Bauer, M., Steinberg, L.S., Mislevy, R.J., & Behrens, J.T. (2004). Design rationale for a complex performance assessment. *International Journal of Testing,* 4(4), 303-332.

Wilson, M., & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement,* 19, 51-71.

Wilson, S.M., Darling-Hammond, L., & Berry, B. (2001). *A case of successful teaching policy: Connecticut's long-term efforts to improve teaching and learning.* Seattle, WA: University of Washington, Center for the Study of Teaching and Policy.

Wolf, S., & McIver, M. C. (1999). When process becomes policy: the paradox of Kentucky state reform for exemplary teachers of writing. *Phi Delta Kappan,* 80(5), 401-406.

Wolf, S., Borko, H., & Elliott, R. L. (2000). "That dog won't hunt!": Exemplary school change efforts within the Kentucky reform. *American Educational Research Journal,* 37(2), 349-393.

Wolf, M.K., Kao, J.C., Griffin, N., Herman, J.L., Bachman, P.T., Chang, S.M., & Farnsworth, T. (2008, January). Issues in Assessing English Language Learners: English Language Proficiency Measures and Accommodation Uses - Practice Review. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

Wolk, R. A. (2007). The Real World. *Teacher Magazine,* 18(4), 54.

Wyoming Dept of Education. (2013). Website. Accessed 5/28/13. Retrieved from http://edu.wyoming.gov/programs/district_assessment/boe.aspx

Yen, W. M., & Ferrara, S. (1997). The Maryland School Performance Assessment Program: Performance assessment with psychometric quality suitable for high stakes usage. *Educational and Psychological Measurement*, 57(1), 60-84.

Yoon, B., & Resnick, L. (1998). *Instructional validity, opportunity to learn and equity: New Standards examinations for the California Mathematics Renaissance*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), CSE Technical Report 484.

APPENDIX A

Interviewees

California	
Edward Haertel	Former Technical Advisor for CLAS, Education Researcher and Professor of Educational Measurement
Joan Herman	Former President of California Educational Research Association, Currently Technical Adviser to the Smarter Balanced Assessment Consortium
Bill Honig	Former California State Superintendent of Public Instruction (1983-1993)
Kate Jamentz	Former Director of the California Assessment Collaborative
Michael Kirst	Former Co-director of Policy Analysis for California Education (PACE), CA State Board of Education President, Professor of Education Policy
Lorraine McDonnell	Political Scientist and Education Researcher
Richard Shavelson	Former Technical Advisor for CLAS, Professor of Educational Measurement
Elizabeth Stage	Education Researcher, Mathematics and Science Advisor for CLAS
Connecticut	
Pascal Forgione	Former Chief of the Office of Research and Evaluation of the Connecticut State Department of Education
Douglas Rindone	Former Chief of the Bureau of Student Assessment and Evaluation for the Connecticut State Department of Education
Kentucky	
Brian Gong	Former Associate Commissioner for the Office of Curriculum, Assessment, and Accountability of the Kentucky Department of Education
Stuart Kahl	Co-founder and former CEO of Measured Progress, Inc. (previously known as Advanced Systems, Inc.)
Daniel Koretz	Education Researcher and Professor of Educational Measurement
Grant Wiggins	Member of KIRIS RFP Committee
Maryland	
Steve Ferrara	Former Maryland Director of Assessment
Jay McTighe	Former Director of the Maryland Assessment Consortium
Scott Marion	Education Researcher, National Center for the Improvement of Educational Assessment
Nebraska	
Douglas Christensen	Former Education Commissioner for the State of Nebraska
New Standards Project	
Robert Marzano	Education Researcher, Formerly Researcher for Mid-continent Research for Education and Learning (McREL)

Lauren Resnick	Co-founder of the New Standards Project
Elizabeth Stage	Former Math and Science Advisor for the New Standards Project
Rhode Island	
Brian Gong	Executive Director of the National Center for the Improvement of Educational Assessment
Vermont	
Stuart Kahl	Co-founder and former CEO of Measured Progress, Inc. (previously known as Advanced Systems, Inc.)
Daniel Koretz	Education Researcher and Professor of Educational Measurement
Marge Petit	Former Deputy Commissioner of Education for the Vermont Department of Education (1996-2000); Former Assessment Specialist at the Vermont Institute for Science, Math, and Technology (1993-1996)
Grant Wiggins	Education Researcher, Former Consultant to Vermont Department of Education
Wyoming	
Scott Marion	Former Director of Assessment and Accountability for the Wyoming Department of Education
Robert Marzano	Education Researcher, Formerly Researcher for Mid-continent Research for Education and Learning (McREL)
Performance Assessment Experts	
Lorrie Shepard	Education Researcher and Professor of Educational Measurement

APPENDIX B

Sample Tasks and Standards

Connecticut Mastery Test and Connecticut Academic Performance Test

- CAPT Released Items: www.csde.state.ct.us/public/cedar/assessment/capt/released_items.htm#7

Kentucky Instructional Results Information System

- 7th Grade Science Sample Tasks: http://www.martin.k12.ky.us/Assessment%20Items/KIRIS/96-97_M-Choice/Science/SG7.pdf
- 5th Grade Mathematics Sample Tasks: http://www.martin.k12.ky.us/Assessment%20Items/KIRIS/96-97_M-Choice/Math/MG5.pdf
- 8th Grade Mathematics Sample Tasks: http://www.martin.k12.ky.us/Assessment%20Items/KIRIS/96-97_M-Choice/Math/MG8.pdf

Nebraska School-based Teacher-led Assessment and Reporting System

- STARS Summary, 2006: http://www.education.ne.gov/assessment/pdfs/STARSbooklet.2006.pdf

New Standards Project

- Mathematics Performance Standards and Sample Items: http://schools.nyc.gov/offices/teachlearn/documents/standards/science/index.html
- English Language Arts Performance Standards and Sample Items: http://schools.nyc.gov/offices/teachlearn/documents/standards/ELA/index.html
- Performance Standards: http://www.ncee.org/publications/archived-publications/new-standards-2/

Rhode Island Diploma System

- English Language Arts Common Tasks: http://www2.ride.ri.gov/HighSchoolReform/DSLAT/pdf/por_070402.pdf
- Mathematics Common Tasks: http://www2.ride.ri.gov/HighSchoolReform/DSLAT/pdf/por_070403.pdf
- Common Task Resources: http://www2.ride.ri.gov/HighSchoolReform/DSLAT/comtask/ct_intr.shtml

Wyoming Body of Evidence

- Wyoming Assessment Handbook: http://edu.wyoming.gov/sf-docs/publications/Wyoming_Assessment_Handbook_Spring_2008.pdf
- Sample District BOE Plan #1 (Common assessment approach): http://edu.wyoming.gov/sf-docs/publications/BGH1_Overview_of_BOE_Plan_Clean_Copy.pdf?sfvrsn=0
- Sample District BOE Plan #2 (Course-based common assessment approach): http://edu.wyoming.gov/sf-docs/publications/BOE_F1_Overview_of_BOE_Plan_Clean_Copy.pdf?sfvrsn=0

CAPT Science Performance Task

Testing for Vitamin C

The recommended intake of vitamin C is 60 mg per day, which can come from different food sources. To determine the presence and the amount of vitamin C in different foods, there is a need to perform simple chemical tests. In this task, you will use a purple indicator to test for vitamin C.

Your Task

First, you and your partner will test a series of vitamin C solutions with known concentrations using the vitamin C indicator. Next, you and your partner will design and conduct an experiment to compare the amount of vitamin C in various fruit juices. Then you will determine the concentration of vitamin C in the juice that was identified as containing the most vitamin C.

During this activity you will work with a partner (or possibly two partners). However, you must keep your own individual lab notes because after you finish you will work independently to write a report about your investigation.

You have been provided with the following materials and equipment. It may not be necessary to use all of the equipment that has been provided. You may use additional materials and equipment if they are available.

CAUTION: **The vitamin C indicator will stain clothes and hands.**

Vitamin C solution (1 mg/mL)	**1 Small plastic cup (for holding test tubes)**
Vitamin C indicator	**5 Labeling dots**
Apple juice	**5 Plastic test tubes**
Pineapple juice	**8 Plastic measuring cups**
White grape juice	**5 Eyedroppers**
Graduated cylinder	**Access to tap water**
Paper towels for clean-up	
Splash-proof safety goggles and an apron for each student	

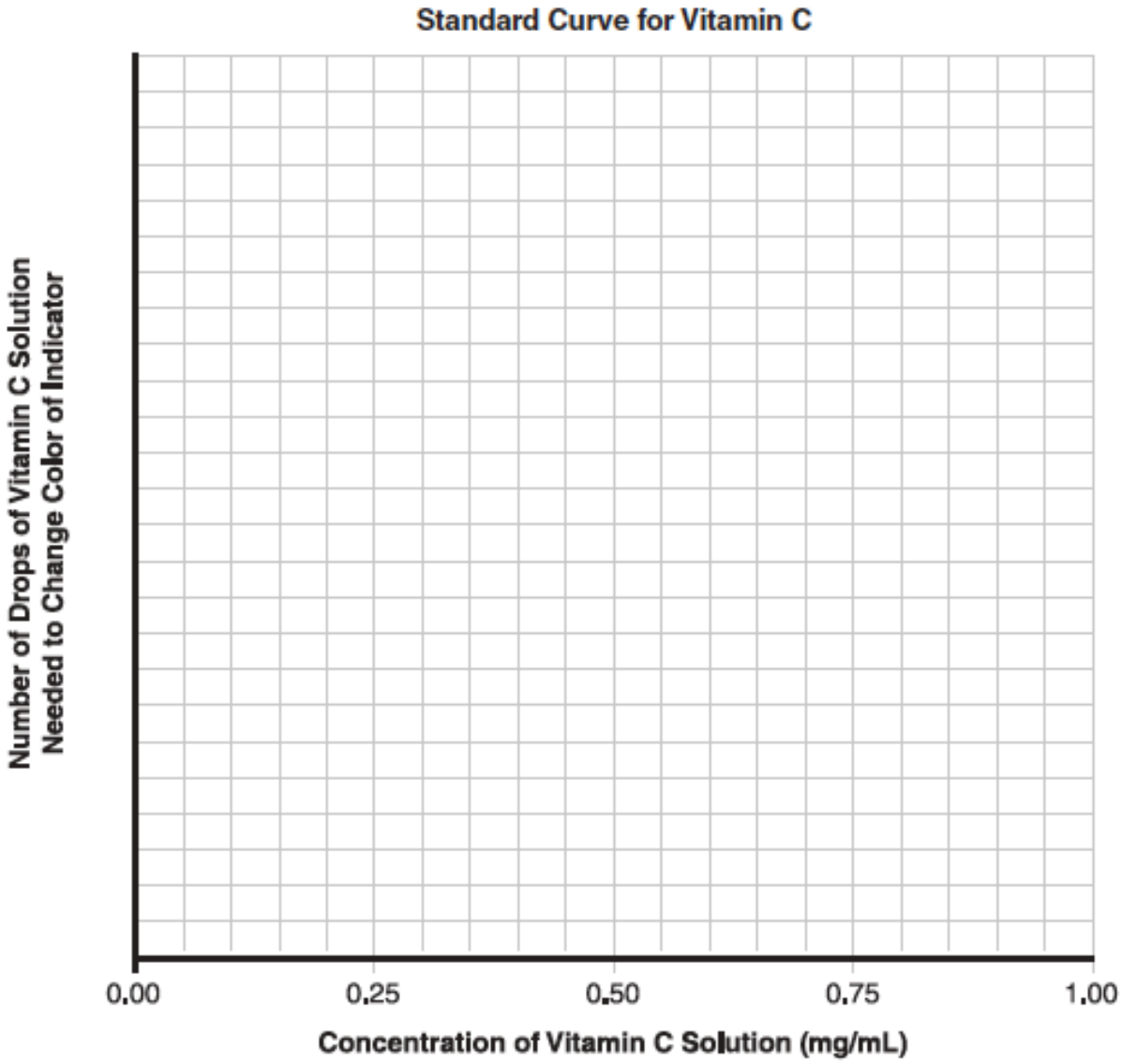
Part I: Testing a Vitamin C Solution

First, you will find out how many drops of a vitamin C solution (with a known concentration) it takes for the indicator to lose its purple color. You will investigate vitamin C solutions with varying concentrations and one solution (water) that has no vitamin C added. The higher the concentration of vitamin C in the solution, the fewer drops it will take for the indicator to lose its purple color.

You have been given a solution containing 1.00 milligram (mg) of vitamin C per milliliter (mL) of water.

1. Using the table below, create vitamin C solutions with different concentrations by mixing the 1.00 mg/mL vitamin C solution with water in plastic cups.
2. Add 10 drops of the purple indicator to a clean test tube.
3. Add drops of the 1.00 mg/mL vitamin C solution, one at a time, to the test tube containing the indicator. Shake the test tube gently after adding each drop.
4. Keep adding drops of the vitamin C solution until the indicator loses its purple color. Record your results in the table below.
5. Repeat steps 3 and 4 using the other vitamin C solutions you created in step 1.
6. On the next page, create a line graph of your results.

Drops of 1.00 mg/mL Vitamin C Solution	Drops of Water Added	Concentration of New Vitamin C Solution (mg/mL)	Number of Drops of Vitamin C Solution Added to the Indicator
40	0	1.00	
30	10	.75	
20	20	.50	
10	30	.25	
0	40	0.00	


Standard Curve for Vitamin C
Number of Drops of Vitamin C Solution Needed to Change Color of Indicator
0.00
0.25
0.50
0.75
1.00
Concentration of Vitamin C Solution (mg/mL)

Part II: Comparing the Amount of Vitamin C in Three Fruit Juices

Now you and your partner will design and conduct an experiment to compare the amount of vitamin C in various fruit juices.

1. **In your own words, clearly state the problem you are going to investigate.** Include a clear identification of the independent and dependent variables that will be studied. Write your statement of the problem on page 6.

2. **Design an experiment to solve the problem.** Your experimental design should match the statement of the problem, should control for variables, and should be clearly described so that someone else could easily replicate your experiment. Include a control if appropriate.

   Write your experimental design on page 6. Show your design to your teacher before you begin your experiment.

3. **After receiving approval from your teacher, work with your partner to carry out your experiment.** Your teacher's approval does not necessarily mean that your teacher thinks your experiment is well designed. It simply means that, in your teacher's judgment, your experiment is not dangerous or likely to cause an unnecessary mess.

4. **While conducting your experiment, take notes on the attached pages.** Include the results of your experiment. Tables, charts, and/or graphs should be used where appropriate and should be properly labeled. Space for your data is provided on pages 9 and 10.

   Your notes will **not** be scored, but they will be helpful to you later as you work independently to write about your experiment and results. You must keep your own notes because you will not work with your partner when you write your lab report.

5. **Use your results from Part I to determine the concentration of vitamin C in the juice that contains the most vitamin C.**

CAPT Experimentation Open-Ended Questions: *Testing for Vitamin C*

Testing for Vitamin C

Students in a science class carried out the *Testing for Vitamin C* performance task.

Group A carried out the following experiment.

1. Take three test tubes and add the purple vitamin C indicator to each of the test tubes.
2. Place 15 mL of apple juice, pineapple juice and grape juice into three different cups.
3. Take one of the juices and add one drop at a time to one of the test tubes with the indicator.
4. Keep adding drops of juice until the indicator turns from purple to clear.
5. Repeat steps 3 and 4 for the other two juices.

Our results are shown below:

Test Tube	Type of Juice Added	Number of Drops of Juice Added
1	Apple	15
2	Pineapple	7
3	Grape	8

1. Group A concluded that "Apple juice has the most vitamin C of the three juices we tested because it took the most drops to turn the indicator from purple to clear." Is this conclusion valid? Explain your answer fully.

Write your answer in your answer booklet.

Group B carried out the following experiment.

Part I

1. Add 10 drops of the indicator for vitamin C to three test tubes.
2. Add drops of grape juice to one test tube until the indicator loses its purple color.
3. Add drops of pineapple juice to the second test tube until the indicator loses its purple color.
4. Add drops of apple juice to the third test tube until the indicator loses its purple color.

Part II

5. Take the Vitamin C solution and make concentrations of 1.0, .75, .5 and .25 mg/mL.
6. Place 10 drops of indicator into four test tubes.
7. Test each concentration of Vitamin C by adding one drop at a time to the indicator in each test tube. Add drops until the indicator loses its purple color.

Our results are shown below:

Part I

Drops of Indicator	Type of Juice	Drops of Juice
10	Grape	8
10	Pineapple	10
10	Apple	17

Part II

Drops of Indicator	Concentration of Vitamin C Solution	Drops of Vitamin C Solution
10	1.00 mg/mL	5
10	0.75 mg/mL	11
10	0.50 mg/mL	16
10	0.25 mg/mL	20

2. **a.** Draw a line graph of the results from **Part II** of Group B's experiment.

   b. Use the graph in your answer booklet to determine the concentration of the juice with the most Vitamin C from Group B's experiment. Explain how you got your answer.

Write your answer in your answer booklet.

3. What is the problem that Group B is trying to solve in Part I? Explain your answer fully.

Write your answer in your answer booklet.

4. What are the variables that need to be controlled in Part II of Group B's experiment? Explain why it is important to control them.

Write your answer in your answer booklet.

CAPT Science Multiple-Choice Items

Red Cabbage pH Indicator Investigation

Red cabbage contains a water-soluble pigment. In a highly acidic solution, the pigment turns bright red, and in a moderately acidic solution, it turns pinkish. In a highly basic solution, the pigment turns yellow, and in a moderately basic solution, it turns bluish.

A student makes a pH indicator from red cabbage that has a reddish-purple color with a pH of approximately 7. The student pours the same amount of the cabbage solution into each of four different beakers. He then adds a different household solution to each of the four beakers until a color change is obtained. His results are shown in the table below.

Household Solution	Color of Mixture
lemon juice	bright pink
club soda	light pink
window cleaner	light blue
drain cleaner	greenish-yellow

The student finds the following chart online from someone else's cabbage pH indicator investigation.

pH	2	4	6	8	10	12
Color	red	pink	purple	blue	green	yellow

1. Based on the information in the passage, which household solution has the lowest pH?

a. club soda
b. lemon juice ✪
c. drain cleaner
d. window cleaner

Strand II: Chemical Structures and Properties

Expected Performance: D 12. Explain the chemical composition of acids and bases, and explain the change of pH in neutralization reactions.

2. Comparing his results to the chart above, what logical conclusion can the student make regarding the substances he tested?

   f. Lemon juice has a pH between 0 and 2.
   g. Drain cleaner has a pH between 10 and 12. ✪
   h. Window cleaner and club soda are both neutral solutions.
   j. Window cleaner can be used to completely neutralize drain cleaner.

Strand II: Chemical Structures and Properties

Expected Performance: D INQ.9 Articulate conclusions and explanations based on research data, and assess results based on the design of the investigation.

3. What should be held constant in the student's investigation to make sure he obtains valid results?

   a. the pH of the household solutions that are tested
   b. the amount of household solution placed in each container ✪
   c. the brand of the household solutions added to the cabbage pH indicator
   d. the final color of the cabbage pH indicator after the household solutions are added

Strand II: Chemical Structures and Properties

Expected Performance: D INQ.4 Design and conduct appropriate types of scientific investigations to answer different questions.

KIRIS
1994-1995 Grade 4 Mathematics On-Demand Items

1. Mr. Miller's class is having a bake sale to raise money for a class trip. Each student offered to bake one batch of a different kind of cookie for the sale. By the day of the sale, the students had baked 550 cookies. The class is thinking about charging 15¢ per cookie. They hope to raise $100.

   a. If Mr. Miller's class sold every cookie they baked for 15¢, how much money would they earn? Would they make as much money as they hoped to make? Explain your answer.

   b. Some students think the cookies should be sold for 20¢ each. How much money would the class earn if they sold every cookie they baked for 20¢? Explain what you did to get your answer.

   c. Imagine that Mr. Miller's class could predict that they would sell only 400 cookies. What price would they have to charge to raise $100? Explain how you got your answer.

   BE SURE TO LABEL YOUR RESPONSES (a), (b), AND (c).

2. The libraries at Lincoln School and King School both charge their students fines for overdue books.

Fines at Lincoln School

Number of Days Overdue	1	2	3	4	and so on
Fine	15¢	23¢	31¢	39¢	and so on

Fines at King School

Number of Days Overdue	1	2	3	4	and so on
Fine	1¢	2¢	4¢	8¢	and so on

   a. If you returned a book to Lincoln School that was 5 days overdue, how much would your fine be? Explain how you figured this out.

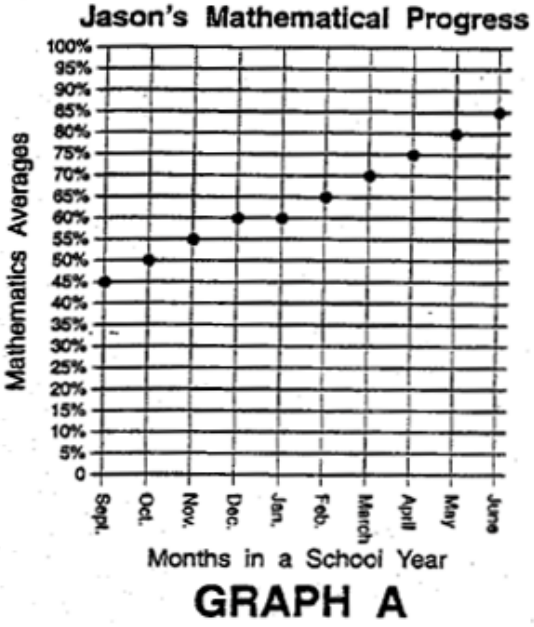
   b. If you returned a book to King School that was 5 days overdue, how much would your fine be? Explain how you figured this out.

   c. Kendra and Brian are each returning a book that is 8 days overdue. Kendra's book is from the Lincoln School library, while Brian's book is from the King School library. Who will pay the greater fine? Explain how you figured this out.

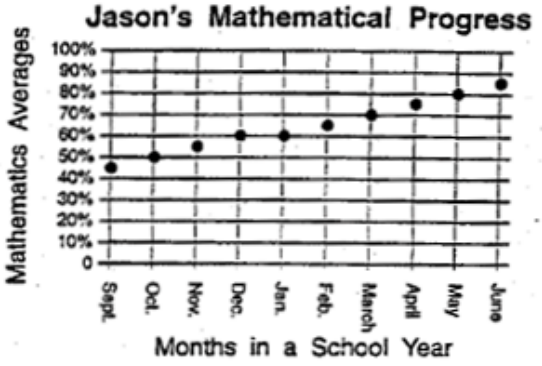
   BE SURE TO LABEL YOUR RESPONSES (a), (b), AND (c).

KIRIS
1994-1995 Grade 8 Mathematics On-Demand Items

5. Last December, Jason's parents said that if his grades in mathematics significantly improved by June, he could go to Disney World during summer vacation. Jason decided he would figure his average for each month. He tried to find different ways to graph the information. He graphed the information on two different graphs as shown below.



GRAPH A



GRAPH B

a. Does each graph show the same information? Explain your reasoning.
b. Which graph would Jason show to his parents to convince them that his mathematics grades had improved? Explain your reasoning.
c. Give several reasonable claims that Jason could use to convince his parents that his grades improved.

6. A hot new rock group is coming to Kentucky for two performances. The concerts will be held at Rupp Arena, which has a seating capacity of 24,000, and Freedom Hall, which has a seating capacity of 19,000. The group needs to make $150,000 from each concert to cover their expenses. They would like to make a total profit of at least $110,000. It is predicted that both concerts will be sold out. The cost of the tickets needs to be the same at both arenas. What would be the minimum cost of the tickets, to the nearest dollar? Show how you arrived at your answer.

KIRIS Released Items reprinted with permission from the Kentucky Department of Education

1996 MSPAP
Public Release Task

Deserts

Grade 3

Reading
Writing
Language Usage
Social Studies

Maryland State Department of Education
March 1996

Tuesday, Task 1
Title: Deserts

DIRECTIONS

You will have 20 minutes to complete Activities 1 through 4 by yourself.

1 What do you already know about deserts? Look at the chart below. Write one thing about a desert that fits under each heading.

A DESERT

Plants	Animals	Climate

2 In this activity, you will give the location of a desert in three different ways. Look at the map titled "Deserts of the World" on pages 6 and 7 in your Resource Book. Letter "D" shows the location of the Sahara Desert. You could give the location by writing:

This desert is in Northern Africa.
It is east of the Atlantic Ocean.
It is in the Northern Hemisphere.
It is south of Europe.
It is on the continent of Africa.

You might even think of other ways to describe the location of this desert region.

Step A From the map, choose a desert other than the Sahara and write the letter of the desert you have chosen on the line below.

_____

Step B Using what you know about location, write the location of this desert in three different ways.

_____

3

Step A Think about the deserts and what they are like. Give one reason people might choose to live in a desert community.

Step B Give one reason people would not choose to live in a desert community.

4 In the United States today, a large number of people are moving to communities in states, such as Arizona, Nevada and Utah, where there are large areas of desert. List two things that could happen to the desert environment as more people move there.

DIRECTIONS

Today you will be reading to be informed. When reading to be informed, you may want to use the following strategies to help you:

- Underline or highlight important facts.
- Pay careful attention to illustrations (pictures), boldface print (darker print), captions (words near pictures, maps, or charts), and aids the author has provided.
- Read the material carefully, and reread it if necessary.

You will have 20 minutes to read the passage "Crossing the Sahara" and to complete Activities 5 through 7. You may look at the passage and the list of words on the chalkboard as often as you like. Open your Resource Book to page 8 and read "Crossing the Sahara."

5 What information about traveling in the desert could someone learn by reading "Crossing the Sahara"?

6 Explain why Geoffrey Moorhouse's trip might have been a good experience for him. Use information from what you have read when you write your answer.

7 Use information from the passage to write about one of the major events during Geoffrey Moorhouse's trip.

DIRECTIONS

Now that you have read about Geoffrey Moorhouse's trip, read "Living in the Desert–I" to learn more about desert life. Open your Resource Book to page 10. You will have 25 minutes to read and to complete Activities 8 through 12.

8 Use the information from this passage to tell why you think Geoffrey Moorhouse selected a Tuareg for a guide.

9

Step A Put an "X" next to the passage that gave you more information about desert life.

_____ "Crossing the Sahara"

_____ "Living in the Desert–I"

Step B

Write a note to your teacher explaining why the passage you selected gave you more information about desert life. Use information from both passages to support your answer.

Dear Teacher,

10 Look at the last picture of "Living in the Desert–I" on page 11 in your Resource Book. How does it show the Tuaregs using the environment to meet their needs?

11 How are your needs and wants the same as or different from those of the Tuaregs in the desert?

12 How are the lives of the people described in the passages different from those of people living in the deserts of the United States today? Write a paragraph for your classmates explaining your answer.

MSPAP

Wednesday, Task 1
Title: Deserts

WRITING PROMPT: WRITING TO PERSUADE

> You have heard of someone who is thinking about traveling across the Sahara Desert, the way Geoffery Moorhouse did. However, this person is not sure whether to take the trip. Write a letter to the traveler to persuade him or her either to go on the trip or to stay home. You may use information from your reading to help support your point of view.

PRE-WRITING

- Think about the problems Geoffrey Moorhouse had on his trip.
- Think about things another person could do differently today to avoid those problems.
- Think about whether or not the traveler should take this trip.

As you write, you may try making a list, a web, or a diagram on the lined paper provided to arrange the ideas you want to share in your letter.

DRAFTING

Use your ideas as you write a first draft of your letter on the lined paper. You will have 30 minutes to plan and to write your first draft.

Thursday, Task 1
Title: Deserts

REVISING

Yesterday you wrote a first draft of a letter. Today you will take 5 minutes to read your draft and think about what you have written. Imagine that you are the traveler reading the letter. Think about the answers to the questions below.

1. Does the letter persuade the traveler to do what is best?
2. Does the letter give reasons that support that advice?
3. Does the letter make sense?

After you have thought about how well your letter answers these questions, you will get some ideas from a partner to help improve your writing.

PEER RESPONSE

You have had the chance to ask yourself questions about how well you have composed your writing. In order to determine if your writing says what you want it to say, it is usually helpful to get someone else to react to your writing. This is called "peer response." You will work with your partner to do your peer response. Your Peer Response Form is on page 36 of your Answer Book.

1. Decide with your partner who will go first.
2. Follow the instructions on the Peer Response Form, and be sure to allow enough time for both of you to read and take notes about the answers to the questions.

PEER RESPONSE FORM

Directions:

1. *Ask your partner to listen carefully as you read your rough draft aloud.*
2. *Ask your partner to help you improve your writing by telling you the answers to the questions below.*
3. *In the space provided, jot down notes about what your partner says.*

1. What did you like best about my rough draft?

2. What did you have the hardest time understanding about my rough draft?

3. What else can you suggest that I do to improve my rough draft?

Use this space to write additional comments.

Friday, Task 1
Title: Deserts

WRITING THE REVISED DRAFT

Write your revised draft on the lines provided in Activity 1. You will have 28 minutes to revise your draft, to write it in this book, to proofread, and to complete Activity 2. Make sure that you correct all of your revised draft in your Answer Book, because only the writing that is in your Answer Book will be scored.

PROOFREADING

After you have written the revised draft, look over your writing to make sure it is clear and complete. Open your Resource Book to page 16 to find the Proofreading Guidesheet. Make any necessary corrections on your revised draft. Use the suggestions on the Proofreading Guidesheet which are appropriate for your letter.

1

2 Draw a circle around the number below that shows how easy or how hard it was for you to write to persuade.

1	2	3	4	5
Very easy	Somewhat easy	About average	Somewhat hard	Very hard

Public release task reprinted with permission from the Maryland State Department of Education