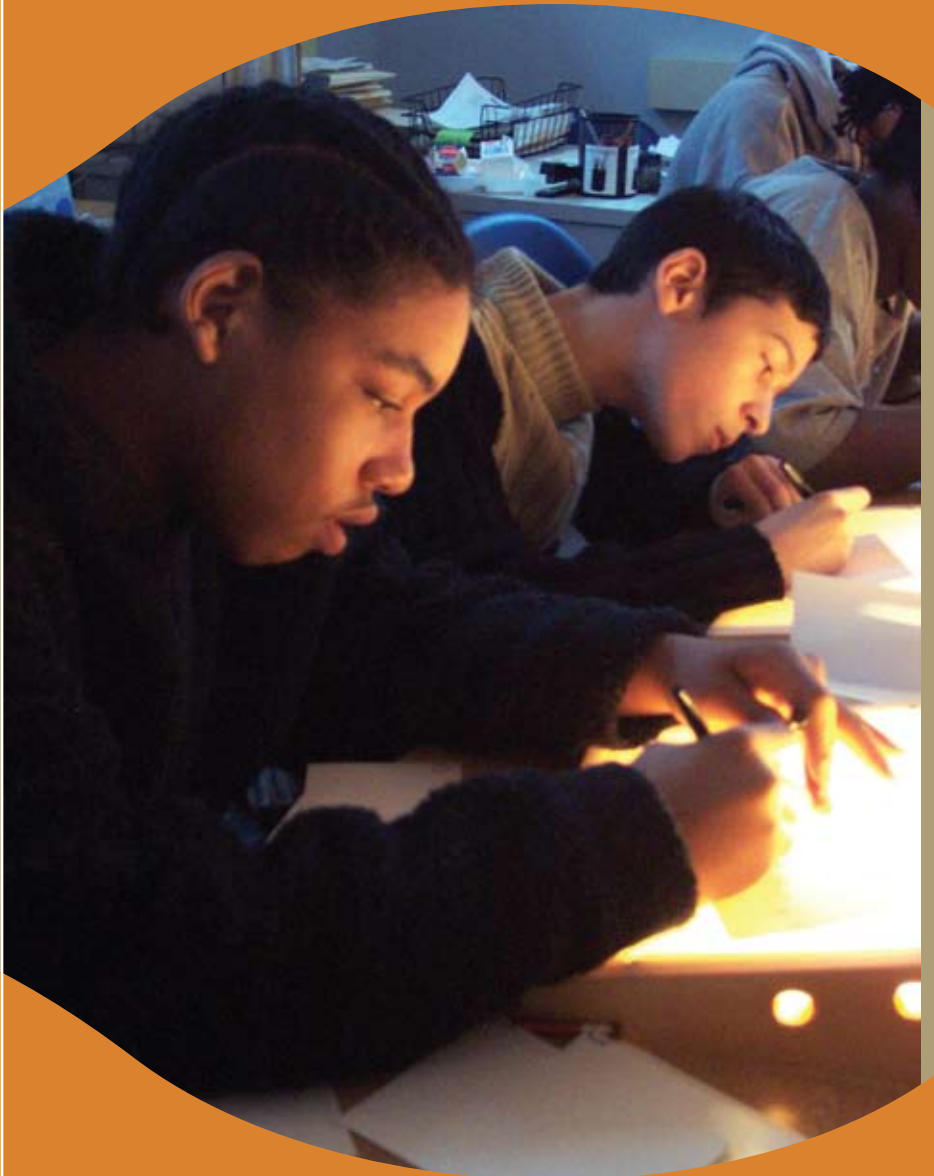


# **Including Performance Assessments in Accountability Systems: A Review of Scale-up Efforts**

*January 2010*



# Including Performance Assessments in Accountability Systems:

---

## A Review of Scale Up Efforts

**Rosann Tung and Pam Stazesky**

**January 2010**

# Including Performance Assessments in Accountability Systems: A Review of Scale Up Efforts

---

## Contents

<b>INTRODUCTION</b> .....	<b>2</b>
WHAT ARE PERFORMANCE ASSESSMENTS? .....	2
WHY USE PERFORMANCE ASSESSMENTS? .....	3
WHY USE PERFORMANCE ASSESSMENTS IN ACCOUNTABILITY SYSTEMS? .....	4
HOW CAN PERFORMANCE ASSESSMENTS BE USED IN ACCOUNTABILITY? .....	6
METHODS .....	7
LIMITATIONS .....	8
<b>FINDINGS</b> .....	<b>8</b>
DESCRIPTIONS OF INDIVIDUAL SCALE-UP EFFORTS .....	9
<i>Queensland, Australia</i> .....	9
<i>Vermont</i> .....	13
<i>Kentucky</i> .....	15
<i>New York Performance Standards Consortium</i> .....	18
<i>Los Angeles Unified School District, California</i> .....	21
<i>Nebraska</i> .....	23
<i>Rhode Island</i> .....	26
SYNTHESIS OF INDIVIDUAL SCALE-UP EFFORTS .....	29
THE USE OF PERFORMANCE ASSESSMENTS IN OTHER FIELDS .....	35
<b>DISCUSSION</b> .....	<b>42</b>
SUCCESSSES IN EFFORTS AT SCALING UP PERFORMANCE ASSESSMENTS.....	42
CHALLENGES IN EFFORTS AT SCALING UP PERFORMANCE ASSESSMENTS .....	42
LESSONS AND IMPLICATIONS FOR STATES, DISTRICTS, SCHOOLS, AND INTERMEDIARY ORGANIZATIONS .....	43
<b>GLOSSARY</b> .....	<b>50</b>
<b>REFERENCES</b> .....	<b>54</b>

# Including Performance Assessments in Accountability Systems: A Review of Scale-Up Efforts

---

## Abstract

The purpose of this literature and field review is to understand previous efforts at scaling up performance assessments for use across districts and states. Performance assessments benefit students and teachers by providing more opportunities for students to demonstrate their knowledge and complex skills, by providing teachers with better information about student progress, and by encouraging schools to build professional collaborative cultures through integrating curriculum, instruction, and assessment.

Despite these benefits, performance assessments in accountability systems are still relatively rare in the U.S. This review of such systems found that the main challenges facing educators interested in using performance assessments at scale, and for accountability purposes, are the technical quality of performance assessments, the sustainability of implementation through strong professional development models, and political buy-in. Through the wealth of their collective experiences, the knowledge and capacity to tackle each of the challenges exists, and those lessons should be applied to the next generation of accountability systems.

Through a systematic description of seven performance assessment scale-up efforts, as well as descriptions of analogous efforts from teacher certification, medicine, and law, this review concludes with implications for future endeavors to integrate performance assessments into accountability systems. Some of these implications include:

- Develop and provide schools and districts with criteria and guidelines for achieving technical quality in performance assessments.
- Use external partnerships and experienced teachers to create professional development models in assessment literacy, assessment design, and scoring.
- Use state and district policy incentives to support the regular collection of technical quality evidence for each assessment.
- Build public and political support for including performance assessments in accountability systems.

## Introduction

### What are Performance Assessments?

Using the term “performance assessment”<sup>1</sup> with educators inevitably elicits the question, “What do you mean by a performance assessment?” Many definitions can be found in the research literature (Darling-Hammond & Pecheone, 2009; Johnson, Penny, & Gordon, 2009; Zane, 2009). Most educators agree that performance assessments set forth expectations for students and require them to:

- Create an original answer or product
- Use higher order thinking and 21<sup>st</sup> century skills<sup>2</sup>
- Demonstrate thinking processes
- Evaluate real world situations

We provide here the accepted academic definition (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999):

*Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied.*

Despite this academic definition, the field does not have a common understanding of what performance assessment means, mostly because there are so many variations of performance assessment in practice; many diverse teacher-assigned tasks are called performance assessments. Many considerations must be taken into account when designing, administering, and revising performance assessments, such as purpose, format, administration window, and length of task. Because terms referring to these nuances are used repeatedly throughout this paper, we review them here.

One reason the term is so difficult to define is that the “measurements” which emerge from performance assessments are from student work products that can take many forms. Essays, speeches, exhibitions, and projects are all performance assessments. Portfolios are collections of student work, some of which are products of performance assessments. Performance assessments require students to do more than choose a fixed response on a test (multiple choice is the most common example of this). Students may be asked to construct a short response, write an essay, or create a product based on a project or experiment.

---

<sup>1</sup> Underlined terms are defined in the Glossary.

<sup>2</sup> 21<sup>st</sup> century skills is the term used for skills that matter in the 21<sup>st</sup> century, that take into account the global economy, technology, and changing workforce requirements. These skills include complex thinking, analytical skills, computer skills, creativity, media literacy, and cross-cultural skills .

Performance assessments serve different purposes. They may be formative, used to help teachers know where students are in their learning, or summative, used at the end of a unit or course to report on student learning.

A final reason that the term eludes common definition is that educators have used the same term to refer to both on-demand as well as extended-time administration windows.

**An extended time performance assessment in science:** A small group of students are directed to make observations of a set of materials to understand gravity’s effect on a ball on a ramp. Students work together to create observations. They then work individually to answer a set of questions, such as ones about predicting future events or analyzing data from past events.

On-demand means students are given the assessment in controlled situations (typically classrooms) on the same day throughout a school, district, or state. Students must complete the assessments in one sitting. Extended-time means that assessments are given to students in a period of weeks or months, depending what best fits the teacher’s or school’s curriculum. As a result, they are given in less controlled and secure situations; some teachers may give the assessment a month earlier than other teachers within the same school.

With both types of administration windows, the amount of time it takes to complete an assessment should be considered. However, educators often do not specify the lengths of time that performance assessments require. Some may take a few minutes

to complete. Others may take hours to weeks to months. In this review, a distinction is made based on the length of time it takes to complete an assessment. Assessment items or tasks that can be completed in one sitting usually take less than 30 minutes to complete. We call these short performance assessment tasks. Those that require more time, group work, or the creation of a substantial product are called extended time assessment tasks.

For the purposes of this review, the term “performance assessment” is used in conjunction with terms describing the purpose of an assessment, the conditions under which it is administered, and the type of product that students are asked to create, because the scale up implementations of performance assessments varied along these dimensions.

**An on-demand, short performance assessment:** Third grade math task from Nebraska. Students demonstrate their ability to count back change up to \$20.00.

### Why Use Performance Assessments?

K-12 teachers have always used formative and summative performance assessments to inform their instruction. Performance assessments benefit students and teachers in ways that fixed response assessments, such

as multiple choice or fill-in-the-blank items, can not. Performance assessments provide students with more ways to show what they know and can do, allow students with different learning styles more opportunity to succeed, and engage students more in their own learning and interests, because they include reflection and demonstration of thinking processes. Performance assessments are more closely aligned with real world skills that students will need, such as public speaking. Others have discussed the benefits of performance assessments as formative feedback:

“...performance assessments enable students to more readily acquire the subsequent skills because they require study and practice by the student and (ideally) additional support and feedback from the teacher, compared with multiple-choice tests.”  
(Goldschmidt, Martinez, Niemi, & Baker, 2007; Pellegrino, Chudowsky, & Glaser, 2001)

Outcomes on traditional standardized tests often show gaps in achievement by race/ethnicity. Some have shown that the use of performance assessments, which assess a broader range of skills, reduces ethnic differences in test scores and college admissions while allowing a greater diversity of students to achieve at high levels (Stemler, Sternberg, Grigorenko, Jarvin, & Sharpes, 2009). Thus, research points to the ability of such assessments to highlight student strengths not evident through traditional routes.

In addition, classroom-based performance assessments benefit teachers because the results are available in real time. Teachers who administer performance assessments and score them receive information about their students' progress and what students know and can do (not just what they remember) and can use them immediately. An additional benefit of embedding performance assessments into curriculum is that through sharing their assignments and looking at student work together, teachers have the opportunity to develop more collaborative practices and school cultures (Darling-Hammond & Wood, 2008). Finally, with common agreement about performance levels for student work, teachers' expectations for the quality of student work increase.

### **Why Use Performance Assessments In Accountability Systems?**

In the 1990s, education reform at the state level called for greater accountability by schools and districts. As a result of findings of inequitable funding and access, many states developed accountability systems to evaluate how schools and districts were serving students. These accountability systems were designed to be low stakes, providing comparative information about the relative performance of schools and districts, but not having consequences for them or for students.

Because information generated from performance assessments provides additional information about what students know and can do, information that standardized tests can not provide, policy makers and educational leaders have long sought to include

performance assessments in accountability systems. For example, in 1993, the Massachusetts Education Reform Act called for such a system of evaluating districts and schools, suggesting that multiple measures, or assessment instruments, should be included:

*“The system shall be designed both to measure outcomes and results regarding student performance, and to improve the effectiveness of curriculum and instruction... The system shall employ a variety of assessment instruments... Such instruments shall include consideration of work samples, projects, and portfolios, and shall facilitate authentic and direct gauges of student performance...”*

In addition, standard 13.7 from the Standards for Educational and Psychological Testing (American Educational Research Association, 1999) supports the assertion that a decision, one that will have a major, high stakes impact on a student, should not be made on the basis of a single test score.

Despite wisdom from the field of measurement, the No Child Left Behind Act of 2001 (NCLB) required states to document educational progress of districts, schools, and student subgroups in order to calculate “adequate yearly progress” in multiple subjects in grades 3-8 and high school. While NCLB called for “multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding” (NCLB, Sec. 1111, B, I, vi), this part of the legislation has been ignored or narrowly interpreted to focus solely upon a one-size-fits-all, paper and pencil, on-demand standardized test to make high-stakes decisions about a school or district. As a consequence of NCLB, states adopted standardized, mostly multiple choice tests to administer to public school children annually. By the 2005-06 school year, all but one state (Nebraska) administered standardized tests in multiple subjects in multiple grades (“Grade-by-Grade State Testing Policies,” 2006). Standardized, on-demand tests at the high school level have become high stakes for students to graduate in 26 states (Viadero, 2009). As a consequence of NCLB requirements, some states reduced their effort at implementing performance assessments at scale.

In 2009, there has been a resurgence of interest in and support for assessment and accountability systems that include performance assessments. Recent advances and changes in technology, the economy, and information access require that schools teach new content and skills, termed “21<sup>st</sup> century skills,” to prepare students for post-secondary education, the workplace, and meaningful civic engagement (Partnership for 21<sup>st</sup> Century Skills, 2007). A vibrant national conversation about the need to equip our students with 21<sup>st</sup> century skills sheds light on the inability of multiple choice tests to adequately assess the acquisition of these skills (Silva, 2008). The MA Governor’s Readiness Project subcommittee on MCAS and Additional Assessments called for the development of more authentic assessments, “common tasks” to accompany benchmarks

for student performance, and the integration of 21<sup>st</sup> century skills (MCAS and Additional Assessments Subcommittee, 2008). At the federal level, a new administration and unprecedented stimulus funds are focusing attention on multiple measures of what students know and can do (Duncan, 2009; FairTest, 2008). The policy agenda of the Obama administration as outlined during his campaign included “a broader range of assessments” that evaluate 21<sup>st</sup> century skills and “an accountability system that supports schools to improve, rather than focuses on punishments...Such a system should evaluate continuous progress for students and schools all along the learning continuum and should consider measures beyond reading and math tests” (FairTest, 2008). There is a growing consensus that improving upon NCLB and state accountability systems will require the assessment of content and skills beyond the traditional core content knowledge.

### **How Can Performance Assessments Be Used In Accountability?**

Despite the benefits of performance assessments to students and teachers, and multiple efforts to include them in accountability systems, very few places have a well-documented, sustainable accountability system that carries high stakes for students and schools and uses performance assessments. A key challenge has been that accountability systems of the NCLB years have constrained how much performance assessments are used. As part of the renewed national conversation about multiple measures in state accountability systems, educational leaders have encouraged the use of assessments that both provide information for improving instruction as well as Adequate Yearly Progress information (Darling-Hammond & Pecheone, 2009; Forum on Educational Accountability, 2007; Stiggins, 2008). Performance assessments can tap 21<sup>st</sup> century skills in ways that on-demand, fixed response standardized tests can not. Therefore, accountability structures that support the use of performance assessments should be developed.

In order to design an accountability system that will provide quality information for both instruction and reporting, it is important to understand all previous efforts at using multiple measures, particularly performance assessments. Performance assessments have historically been included in two major ways: as common assessment tasks across schools and in local (school or district level) assessment systems. While all 50 states currently use standardized tests for student for accountability, only a few states use the results of performance assessment for decisions about promotion or graduation.

Why were standardized tests so widely adopted for accountability reporting while performance assessments, despite their advantages to teachers and students, were not? The purpose of this review is to examine efforts at using performance assessments at a scaled up level for state accountability in order to learn from those efforts. The following specific questions guide the review:

- How have performance assessments been included in state accountability systems?
- What was the nature of the professional development provided to support assessment literacy and to successfully implement performance assessments in state accountability systems?
- How was the technical quality of each accountability system assured?
- What lessons can we learn from the reviewed scale up efforts?

Through answering these questions, recommendations emerge for the design of a sustainable, technically sound accountability system that uses performance assessments. While others have summarized the states and countries that have used performance assessments at scale, not necessarily for accountability, this paper reviews and compares state-wide or large-scale efforts at using combinations of common performance assessments, local assessment systems which include performance assessments, and standardized tests at scale. In addition, this review evaluates reliability and validity evidence gathered for performance assessments initiatives.

## Methods

In order to learn about assessment scale-up efforts which included common performance tasks and/or school-based performance assessments systems, available documentation on scale-up efforts was collected. Scale-up efforts were assessed for whether or not they met the following criteria:

- The scale-up effort must be described in writing and have accessible documents to review
- Performance assessments must be based on local or state standards
- The scale-up effort must use performance assessments for information about K-12 student achievement across content areas like English Language Arts (ELA) and math

It is important to note that this review considers the use of performance assessments in several forms, listed below. No two locations used performance assessments in the same combination:

1. In local (school- or district-based) assessment systems
2. As extended time, common tasks administered to all students in a grade and subject across a district or state during an extended administration window
3. As short common tasks administered to all students in a grade and subject across a district or state on demand
4. In portions of statewide, on-demand standardized tests (writing prompts)

Both research and opinion articles were found in ERIC, key education journals, Education Week, and the bibliographies of key reviews. Most searches were conducted via on-line databases.

In addition, we interviewed more than 15 people who were directly involved in the scale-up efforts included in this review, other experts in the field of measurement and assessment, and authors of key research and review articles. Documents and articles, many not available on the internet, were also identified through these interviews with leaders in the assessment field.

We found that our document and interview approach continually led us to the same experts and resources. In other words, there is a consensus in the field about who represent expertise in performance assessment scale-up efforts and where and when legitimate performance assessment scale-up efforts occurred. This redundancy in identification of resources strongly suggested that significant scale-up efforts were not omitted from this review.

## **Limitations**

This review has a number of limitations to bear in mind. Because of the limited number of locales meeting our criteria for inclusion, the findings and recommendations may not generalize to other settings.

Several other states implemented performance assessments at scale for accountability reasons, including Maryland and Wyoming. However, they were not included in the review because little to no information on their efforts was publicly available for review, our first criterion for inclusion in this paper.

Many authors have criticized the high stakes nature of the standardized testing movement for its role in narrowing the curriculum, in reducing teacher morale and positive school culture, for hindering implementation of best classroom practices, and for score inflation. They hypothesize that using performance assessments ameliorate some of these negative effects. The research reviewed in this paper does not address whether or not they do.

## **Findings**

Our findings are divided into three sections.

1. Descriptions of each scale-up effort, organized by context, design, professional development and implementation, and technical evidence of quality
2. Summary of important details about the scale-up efforts, allowing for at-a-glance comparisons across sites about key features of the assessment systems
3. Lessons from other disciplines (besides K-12 public education) in which performance assessments have been brought to scale.

## Descriptions of Individual Scale-Up Efforts

Our documentation search and interviews identified seven scale-up efforts which met our criteria. Out of the seven, all but one were in the United States. Within the US, all but two were states. Of the two that were not state-wide, one was the second largest school district in the US, Los Angeles Unified School District (LAUSD) with a student population larger than some states and the other was a group of schools across districts in New York state, the New York Performance Standards Consortium. We describe each scale-up effort (in order of the start date of the scale-up effort).

Table 1: Scale-Up Locales, by years, subjects, grades

Locale	Years	Subjects	Grades
Queensland, Australia	1970-present	English, Math, Science	4, 6, 9; 11+12
Vermont	1988-1996	Writing Math	Writing, 5 & 8; Math, 4, 8, 10
Kentucky KIRIS	1990-1999	Writing Math	Writing 4, 7, 12; Math 5, 8, 12
NYPSC	1995-present	Literature, Social Studies, Math, Science	12
Los Angeles Unified School District	1996-2003	Writing, Math	Evolved each year; ended with performance tasks grades 2-9
Nebraska STARS	2001-2008	Reading, Math, Social Studies, Science	All grades, reported at 4, 8, 11
Rhode Island	2004-present	ELA, Math, Science – state; Social Studies, Technology, Arts – district	12

### Queensland, Australia

#### *Context*

Queensland, a state in the northeast of Australia, is the only scale-up effort in this review that is outside of the US. Since 1970, the Queensland education system has used data to inform decisions about how to build its assessment system.

Entrance to college has been determined by multiple measures of students' performance in grades 11 and 12 since the 1970s. Their scores in courses, which are performance based, and on the Queensland Core Skills test, which includes writing, multiple choice,

and short answer items, combine to determine whether students qualify for college entrance.

More recently, efforts to bring this assessment framework to the primary grades has occurred in two forms. From 2001-2005, an ambitious initiative called New Basics was implemented in a small number of schools. This effort was an attempt to encourage teachers to view assessment as aligned with and integral to curriculum and instruction. The main vehicle for New Basics was Rich Tasks, which were a set of interdisciplinary projects to be completed over a span of 2-3 years by students. Rich Tasks were examples of extended administration window, extended time performance assessments. New Basics was not ever taken beyond a pilot stage to scale across Queensland because implementation was burdensome on the teachers in the last year of each grade span. These teachers had to make sure the in-depth, time consuming Rich Tasks were completed (J. Drazek, personal communication).

The latest revision to the assessment system in Queensland, which began in 2006, is called the Curriculum, Assessment, and Reporting Framework (QCAR) and was designed by the Queensland Studies Authority, a government agency that oversees assessment policy for the state and develops syllabi, assessment, and certification (Queensland Department of Education, 2005). Its charge is to provide consistency for students from school to school and district to district. The new framework responds to the public's and educators' concerns about changes in the state's student population. It also addresses the changing nature of the workplace and different employer demands placed on high school graduates. The public recognizes that as Queensland becomes more ethnically diverse through immigration, and businesses need more employees with 21<sup>st</sup> century skills such as teaming and problem-solving, K-12 education must respond with curriculum, instruction, and assessment that embrace diversity and encourage higher level skills. The senior assessment process and the parallel QCAR framework are elaborated upon below.

### *Design*

QCAR assessment is based on the state's Essential Learnings, which are streamlined content standards due to concerns about the breadth of the curriculum. Essential Learnings are published for grades 3-9 and supported by accompanying documents that describe their three types of components, the assessable elements and descriptors of quality (a rubric), scope and sequence guides, and unit templates for each subject. At the senior level (grades 11 and 12), course syllabi are published for each subject.

The components of senior assessment in Queensland include:

- Queensland Core Skills Test, centrally scored
- Course assessments, locally scored and externally scored

- Local performance assessment systems (graduation portfolios) for high school graduation, locally and externally scored

Paralleling the senior assessment, QCAR for primary and middle grades includes:

- National on-demand, standardized test for literacy and numeracy, in grades 3, 5, 7, and 9 centrally scored, for state and district accountability
- Statewide common performance tasks, Queensland Comparable Assessment Tasks (QCATs), in grades 4, 6, and 9, locally scored
- Local performance assessment systems, locally and externally scored.

### *Professional Development*

In Queensland, a principle of the assessment policy is that good assessment is a key element of teacher professional development (Queensland Studies Authority, 2009). The QCATs are not simply assessment tasks; they are packages of information for teachers, including the student performance tasks. There are teacher administration and marking guides which provide great detail about what prior knowledge students must have, what content knowledge and skills students must be exposed to, and how scoring should occur. Rubrics and samples of student responses are included. These packages therefore both inform teachers and assess students.

In addition to the QCAT guides, there is an internet- and password-accessible assessment bank of past items and instruments similar to the QCATs, that teachers can download and use, modify, or study.

Most additional professional development is provided by staff from the Queensland Studies Authority. The QSA provides workshops throughout the state on assessment literacy for pre-service and current teachers.

**QCAT for 7<sup>th</sup> grade science:**  
 90 minutes over 1-2 days;  
 Given some contextual information, students must analyze and construct food webs in two environments. Through multiple prompts, students must show an understanding of food chains and the impact of environmental disruptions on populations.

### *Technical Quality*

The QSA facilitates the professional development required for teachers to achieve inter-rater reliability. They rely on moderation studies, a process by which teachers serve on curriculum review panels and judge student work from other teachers and schools. It is through moderation that greater reliability is achieved. The discussions about the quality of the student work lead to a common language and understanding of expectations and proficiency. The QSA facilitates the process by which it is determined whether seniors have met the college entrance requirements.

Freebody (Freebody, 2005) discusses the types of validity that are desirable in the state's assessment system. They include content validity, or the extent to which the task assesses the standards it was designed to measure; concurrent validity, the extent to which the task aligns with other assessments given at the same time; and predictive validity, the extent to which the task aligns with other assessments given some time in the future; and generalizability, the extent to which the outcomes can be generalized to other groups or situations. The reliability and validity evidence for the first pilot of QCATs will be publicly available in Fall 2009.

For the senior assessments of courses, evidence of the technical quality of the course assessments was collected both within schools and through external review. These individual course reviews are published in State Panel Reports for every subject at the senior level. They are aggregate reports of the review panels and address the quality of assessments. However, primary evidence of technical quality is not publicly available.

#### *Successes in Design, Professional Development, and Technical Evidence*

The successes of the Queensland Curriculum, Assessment, and Reporting Framework are in large part due to the support and oversight by a state agency separate from the State Department of Education, called Queensland Studies Authority. Several aspects of design make the implementation of QCAR more sustainable, including streamlined competencies, called Essential Learnings, which guide curriculum and instruction, and the fact that standardized tests and common tasks are administered in different school years to reduce burden on teachers. The common performance tasks, or QCATs, are published with detailed administration and scoring guides, so that scoring can be done locally, reducing costs and increasing opportunities for ongoing professional development.

#### *Challenges in Design, Professional Development, and Technical Evidence*

The assessment initiative preceding QCAR was called New Basics. New Basics was based on the idea that assessments should drive curriculum and instruction. Rich Tasks, the mainstay of New Basics, were extended administration window, extended time performance assessments. They were not brought to scale across the state because of the lack of will and resources at the government level to provide the extra professional development for teachers in all subject areas and grade levels that would have been required.

The current QCAR framework project is a compromise on the more ambitious Rich Tasks – QCATs provides more clarity for teachers about what is being assessed, the introduction of streamlined standards, and support for teachers to implement quality assessment tasks. However, these QCATs are shorter performance tasks than Rich Tasks, taking an average of 19 minutes to complete in an extended time administration window.

With shorter performance tasks, there is less time for students to create, perform, and/or exhibit their learning.

## **Vermont**

### ***Context***

Vermont is a state in which local control has historically been important. In education, teachers used their voice and authority to support a highly decentralized system where schools and districts, not the state, developed accountability mechanisms. In 1988, teachers were concerned with score inflation in testing and grades and improving the use of assessments to drive instruction (Mills, 1996). The state's commissioner of education, along with his policy advisor, set out to develop an assessment framework which was teacher-driven and responsive to public demand, while also amenable to accountability requirements. What emerged after a series of public meetings and interviews with stakeholders was a portfolio-based performance assessment system and uniform tests in math and writing. This effort ended in 1996, when the state adopted the Vermont Comprehensive Assessment System and ended the use of portfolios for accountability purposes. The current statewide standardized test is the New England Common Assessment Program (NECAP) in math, reading, writing, and science and is based on the state's Framework of Learning Standards and Opportunities.

### ***Design***

Vermont's assessment system from 1988-1996 was composed of:

- Statewide standardized on-demand exam, centrally scored
- Local, portfolio-based assessment systems, locally scored
- On-demand writing assessment

From 1988-1996, the state-wide Uniform tests in math included multiple choice and constructed response items. The Uniform test in writing was an on-demand essay with a common prompt. Tests were administered in 4<sup>th</sup> and 8<sup>th</sup> grades in writing and math (Koretz, 1998).

Districts and schools were given broad guidelines for their local assessment systems in the same grades, which were to be portfolios of student work. The guidelines did not specify what work should be included, nor that the work be standardized or performance based. Students were to include not only their best work but also supporting work. Holistic rubrics for portfolios were developed. Rubrics included 4-5 scoring levels and were not tied to individual performance tasks.

### ***Professional Development***

Professional development for the assessment system in Vermont during the portfolio years was largely created by teachers. A teacher network was formed to support the

learning about and implementing student portfolios. Volunteer teachers from around the state conducted training and support as needed, by phone or in person. In that way, they could be responsive to the needs of teachers as they presented themselves. Teachers also developed committees to guide the portfolio development process in their regions.

Even with this responsive teacher network, the implementation of portfolios varied widely, raising questions about the validity of the assessment (Koretz, Stecher, Klein, & McCaffrey, 1994b). Not only were the assignments leading to the student work included in portfolios widely varying, from worksheets to collaborative projects, but also, the amount of teacher guidance or student revision was different from classroom to classroom.

### ***Technical Quality***

#### **Reliability**

Portfolios were scored by teachers who were not the students' own in statewide meetings of teachers (Koretz, et al., 1994b). A random sample of portfolios was evaluated by these second scorers. Given the small scales used in the portfolio assessments (4 to 5 points), score agreement due to chance was high, and inter-rater reliability was low. In addition, given the different types and difficulties of tasks, score reliability was also low. In the second year of implementation, the math portfolios scores showed improved reliability across raters, possibly because of increased training and calibration for raters (Koretz, Stecher, Klein, & McCaffrey, 1994a).

#### **Validity**

Portfolio scores and Uniform test scores for individual students were compared. Evidence of concurrent validity of portfolio scores with uniform test scores was low (Koretz, et al., 1994b). No other types of validity evidence were reported.

#### ***Why did it end?***

The portfolio development and double scoring created an administrative burden that was difficult for the state to sustain, although portfolios continue to be used at the local level (B. Gong, personal communication).

### ***Successes in Design, Professional Development, and Technical Evidence***

Vermont's teachers pioneered portfolio assessment and demonstrated a number of successes and strengths. For example, the system was created and driven by teachers, with support from the public. Teachers and principals reported a positive impact of portfolio work on their instruction, including more time with cross-disciplinary units and projects (Koretz, et al., 1994a). As a result of performance assessment, they spent more time with students on higher order thinking and skills (Stecher, 1998). Teachers also reported that their curriculum was more aligned with standards (Stecher, 1998). Students

benefited in exposure to more group work in the classroom (Stecher, 1998). Vermont evaluators also collected evidence that with professional development, the reliability of scoring improved over time (Koretz, et al., 1994a).

### *Challenges in Design, Professional Development, and Technical Evidence*

While the portfolio initiative was teacher-driven, leadership for implementation was not sustainable. More educators were needed to administer the initiative. In the grades in which both the on-demand State Uniform tests and portfolios were required, teachers bore the burden of preparing their students for both types of assessment in the same year.

Professional development in assessment literacy took time and resources in an ongoing way. Teachers needed to understand assessment, create performance tasks, and find time to prepare, implement, and score portfolio-related tasks. These tasks took over 30 hours per month, in addition to the time needed to change curriculum and instruction to prepare students for portfolio products.

Achieving technical quality for the portfolios was difficult, given the wide variation in implementation of the portfolio program from teacher to teacher and across schools. While there was improvement in inter-rater reliability over time, and teachers reported changes in their instruction, the process was resource-intensive. In terms of scoring, reliability was low for several possible reasons: the scoring rubrics were not tied to individual tasks, the scales were too small, the difficulty of assignments varied, and the amount of guidance and revision allowed by teachers varied.

## **Kentucky**

### *Context*

The large-scale assessment system implemented in Kentucky from 1990 – 1999, called Kentucky Instructional Reporting and Information System (KIRIS), was the accountability approach mandated by the Kentucky Educational Reform Act (KERA) (Fontana, 1995; Gong & Reidy, 1996). KERA was passed by the legislature due to inequitable funding for public education in the state. The backbone of KIRIS was a portfolio in writing and in math. Educators in the state used KERA and KIRIS as a motivation for curriculum reform (Pearson, Calfee, Webb, & Fleischer, 2002).

### *Design*

At different times, KIRIS included all four components of an assessment system for learning and for accountability:

- A statewide standardized test, centrally scored
- Local performance assessment systems, locally scored
- Short performance tasks, centrally scored

- Extended time performance tasks, centrally scored

The standardized, on-demand test in Kentucky was given for writing and math, and included multiple choice, constructed response, and an essay prompt.

The extended time performance tasks in Kentucky were called “performance events” and took one hour to several hours to complete. During these assessments, which occurred eight times per year in the subjects and grades assessed, students worked collaboratively and individually to complete a task. Performance events were changed each year so that they would remain authentic and of high interest to teachers and students.

The locally developed writing portfolios, the longest lived part of KIRIS, included six different pieces/types of writing from students. As with Vermont portfolios, they were judged holistically rather than piece by piece and given a single score. Writing portfolios were developed in 4<sup>th</sup>, 8<sup>th</sup>, and 12<sup>th</sup> grades.

### *Professional Development*

The Kentucky Department of Education partnered with universities and nonprofits to provide professional development to teachers (Stecher, 1998). Prior to the state’s adoption of writing portfolios for assessment and accountability, many Kentucky teachers were trained through the National Writing Project (NWP), an initiative that supported teachers to share writing practices among each other. The NWP groups were all based at universities. Due to the scope of NWP training, the implementation of writing portfolios was relatively smooth.

During KIRIS, the external audit of portfolios revealed large discrepancies between local teacher scores and the outside auditors’ scores. Two other forms of professional development emerged from these findings. Both were designed to improve scoring consistency. For schools that had the largest differences between its own portfolio scores and external reviewer scores, targeted professional development resulted in both more accurate scoring and in increases in student performance, likely attributable to improvements in instruction. Teachers who participated in intensive scoring sessions reported making the connection between the assessments and instructional approaches (Gong & Reidy, 1996). Another means for improving scoring consistency was through a self-audit of portfolio scores. Schools that chose to self-audit under the guidance of the Department of Education achieved higher accuracy and inter-rater agreement. This form of professional development, chosen by the school staff, allowed them to analyze their instructional program simultaneously. The teachers became engaged in applying the standards to their instruction and assessment (Stecher, 1998).

## *Technical Quality*

### Reliability

Evidence of the reliability of scoring between teachers and over time was collected in Kentucky for the writing portfolios. Data from three types of reliability tests were collected. These tests included teachers scoring anchor papers of each level of student portfolios, inter-rater reliability data between pairs of reviewers, and examination of scorer drift over time. Early in KIRIS implementation, when portfolio scores were compared between students' own teachers and by external reviewers, teachers' scores were higher. Over time, with tight training and well-developed rubrics, reliability of scoring improved (Koretz, 1998).

### Validity

When portfolio scores were compared with other measures such as the on-demand essay, there were significant differences in student scores (Koretz, 1998). Portfolio writing pieces were scored higher and could be revised. As with Vermont, unreliability undermined evidence of validity.

### *Why did it end?*

Different components of the KIRIS accountability system ended for different reasons, although all of the reasons related to the tensions in achieving technical quality, sustainability, and political support. The use of performance events ended during KIRIS. Not only were new performance events difficult to equate with old ones, their scoring was costly. In one year when the data showed that performance events were not of equivalent difficulty from one year to the next, the state legislature became involved in a successful battle to end their use.

Data showed that the use of portfolios faced challenges regarding technical quality in the early years. However, with professional development, over time their technical quality improved. KIRIS was replaced by Commonwealth Accountability Testing System (CATS), which included the writing portfolios and added an on-demand standardized test.

### *Successes in Design, Professional Development, and Technical Evidence*

Kentucky is the only locale included in this review which attempted to implement all four types of assessment components. While the performance events were short-lived, the writing portfolios lasted until 2008. Part of the success of the writing portfolios may be attributed to a longstanding partnership with a nationally recognized organization that supported communities of teachers in sharing practice around teaching writing. Through this partnership and the professional development offered through KIRIS, scoring and instruction both improved. Researchers (Stecher, 1998) showed that teachers reported

increased expectations of students (Stecher, 1998); they spent more time on higher order thinking tasks; and they grouped students more heterogeneously as a result of KIRIS.

### *Challenges in Design, Professional Development, and Technical Evidence*

The original intent of KIRIS was to cover seven core academic subjects; however, this ambitious design was scaled back such that the assessment components were implemented in writing and math. Portfolios and performance events both faced challenges in demonstrating evidence of technical quality – both reliability and validity evidence started out as weak and improved over time. However, both also suffered from public and political disapproval at key moments in KIRIS, resulting in the elimination of performance events and valuable resources directed at defending the intent of portfolio assessment.

### **New York Performance Standards Consortium**

#### *Context*

The New York Performance Standards Consortium schools came together in 1995, when the state’s Regents exams became a requirement for high school graduation. Through the efforts of then-Commissioner Tom Sobol and leaders in a small group of progressive schools in New York City, Consortium participants obtained a waiver from four of the five Regents exams; the schools take only the English Language Arts exam. Instead of taking the standardized tests, students in the 28 participating schools complete standards-based performance assessments in four core subjects as graduation requirements; individual schools may add additional tasks in such areas as art, foreign language, and internships. All of the Consortium schools are small schools, and many are members of the Coalition of Essential Schools, a national school reform approach built upon ten common principles.

#### *Design*

The Consortium’s assessment design has seven components: active learning, formative and summative documentation, strategies for corrective action, graduation level performance tasks aligned with learning standards, external evaluators of student work, and professional development (<http://performanceassessment.org/performance/components.html>). The graduation requirements are streamlined and involve four common, extended time performance tasks, one per core academic subject. The tasks are both skills- and content-based,

Research paper: In order to graduate, students must complete a research paper in social studies. The paper is judged on a holistic rubric which evaluates viewpoint, use of evidence, organization, implications, voice, and presentation.

and require both written work and an oral presentation, defense, or panel. The four tasks are:

1. Literary analysis
2. Application of higher math
3. Extended science project or original experiment
4. Research paper in social studies

Tasks, both written and oral, are scored by the students' and other teachers as well as external evaluators—some of them experts in their fields— using skills-based rubrics for each task. Participating schools are encouraged to design curriculum and instruction throughout high school which leads students to proficiency in the graduation tasks by the eleventh and twelfth grades.

In addition, many of the Consortium schools also use portfolios to document what students know and can do. The Performance Assessment Review Board, whose members are educators and assessment experts, reviews member schools every five years to determine whether or not the school has developed structures which support the use of performance assessment in a sustainable way.

### *Professional Development*

The Center for Inquiry supports the assessment work of NYPSC schools through a series of workshops on curriculum, instruction, and assessment for teachers and administrators. Some workshops feature Consortium teachers and others are facilitated by Center staff. The workshops are both content-specific and assessment focused. The Center also organizes cross-Consortium conferences and school visits for participating teachers.

### *Technical Quality*

#### Reliability

The Consortium schools have participated in moderation studies in recent years. These studies are designed to generate evidence of reliability of scoring and have been conducted in three phases: 1) Consortium teachers score student work individually and come to consensus as a group; 2) External reviewers re-score student work; and 3) Schools sponsor site-based moderation studies to compare with the moderation work done cross-Consortium. Each year, staff at the Center for Inquiry collect student work for each graduation task from each school and create packets of student work. In the fall, groups of teachers across the Consortium use the rubrics individually and in small groups to discuss scoring and come to consensus. Center staff report high inter-rater reliability scores (P. Tashlik, personal communication) and have created and publicly posted collections of exemplar papers at each performance level.

## Validity

The NYPSC considers predictive validity to be the most important evidence of validity. One study documented a subset of graduates of NYPSC schools from the classes of 2001 and 2002, who were more likely to be low income students and students of color than other NYC public school students and found that graduates attend college at a higher rate, persist into their second year of college at a higher rate, and have higher college GPAs than the national average (Foote, 2007).

Students in the Consortium schools take the state standardized test in English only, so validity evidence regarding a comparison of performance assessment outcomes with other testing outcomes is not possible. The NYPSC and Center for Inquiry believe that concurrent validity studies, in which outcomes on performance assessments are compared with outcomes on other exams, give credence to high stakes standardized testing and therefore do not plan to undertake these studies. Their performance tasks are rooted in their local curricula, and validity studies should emerge from alignment of these tasks to the curriculum and standards.

### *Successes in Design, Professional Development, and Technical Evidence*

With the political and community advocacy of a third party organization, this consortium of schools succeeded in obtaining a waiver from four of the five state on-demand, standardized accountability tests. The waiver has been in effect since 1995. With the waiver, teachers in Consortium schools do not need to “teach to the test.” They can focus their teaching and learning on the performance tasks and college readiness. One reason the Consortium has been successful is that it is composed of a network of like-minded schools, many of them affiliated with the Coalition of Essential Schools reform model. As a network, teachers in these schools find support for and collaboration about their teaching practice. Many of the schools have a culture of including projects and common performance tasks in other grades, before the senior year, as scaffolding for graduation tasks. However, the high stakes nature of performance assessment is focused on high school exit, rather than spread out among all of the grade levels. Finally, research has shown that students in Consortium schools not only earn good grades in college, they also persist in college at higher rates than the national average.

### *Challenges in Design, Professional Development, and Technical Evidence*

One challenge for the Consortium’s approach is that most districts operate in a context where a waiver from testing would not be possible to obtain. Therefore, this work applies only to this small group of Consortium schools. Interviews with Consortium staff confirm that the teacher led professional development is both time- and resource-intensive, partly because each session is tailored to participating schools’ specific needs and interests. A final challenge is the collection of technical evidence about reliability. Moderation

sessions have been held; the data is not yet publicly available. However, the Consortium has shown a credible assessment system that does not depend on standardized exams is possible and provides students with the skills, knowledge, and attitudes they will need to succeed in college.

### Los Angeles Unified School District, California

This scale-up initiative is the only one included in this review that is a district, Los Angeles Unified School District (LAUSD). This initiative was also not intended for using performance assessments in accountability. Rather, its purpose was to build teachers' capacity to use performance assessments. However, given the size of the district and the details of its implementation, the initiative has been included because it informs other performance assessment scale-up initiatives.

#### Context

The designers of LAUSD's performance assessment scale up effort from 1996-2003 were the district staff and UCLA's Center for Research on Evaluation, Standards, and Student Testing (CRESST). CRESST staff created performance assessments based on a history of research on assessment development and were interested in studying the results of this scale-up endeavor. The goals of the partners were to (Niemi, Baker, & Sylvester, 2007):

- Provide a clearer focus for standards-based instruction in mathematics, language arts, history and science than existing large-scale multiple-choice tests did.
- Provide information that teachers could use to implement more effective instruction in areas targeted by state testing plans.
- Obtain information that could be used to improve professional development across the district.
- Provide student achievement information consistent with state plans to incorporate performance measures in the state assessment system, and specifically to predict performance on state writing tests that would be developed in the future.

CRESST uses model-based assessment (MBA) to support student learning. In MBA, the short items or tasks assess and measure domain independent skills and knowledge, domain specific big ideas, and the cognitive demands of both.

**9<sup>th</sup> grade math Performance Assignment:** The assessment is an investigation of diameters and circumferences. Students are given data about ten types of cylindrical cans in the form of a table. They use graph paper to plot the relationship between diameter and circumference, explain aspects of their graphing, and apply their analysis to a new situation.

## *Design*

In LAUSD, the original plan to include performance assessments district-wide in multiple subjects and grades changed from year to year. Therefore, CRESST was asked to develop different assessments in different years of the scale-up effort as other district and state assessment requirements were phased in and out. During the majority of the initiative, students were assessed with:

- State standardized, on-demand tests, centrally scored
- State on-demand writing tasks, centrally scored
- On-demand Performance Assignments in ELA and Math in multiple grades, locally scored

## *Professional Development*

As a partner in this initiative, the district provided strong buy-in and support for professional development. Each school had a part-time standards-based assessment coordinator and each local district had a full time standards-based education administrator. By the third year of the project, CRESST had built the capacity of the district to handle all training, scoring, and reporting (Niemi, et al., 2007).

The effort allowed the district to build a focus for professional development on assessment literacy. The district developed administrative and local scoring training, professional development on the use of assessments to improve instruction, and scoring processes for reliability. All teachers participated in this professional development.

In addition, volunteer teachers participated in the development and refinement of the Performance Assignments, including pilot testing and scoring. For these 3000 teachers, the iterative process was added professional development in assessment literacy.

## *Technical Quality*

### *Reliability*

In order to analyze the inter-rater reliability of scoring for Performance Assignments, schools reported on the percent agreement between local scorers (students' own teachers) and external scorers. Because of the emphasis of professional development on scoring training, reliability was strong. Plans were in place to support teachers with less reliable scoring. Reliability of scoring improved over time when teachers received professional development (Niemi, et al., 2007); (Martinez, Goldschmidt, Niemi, Baker, & Sylvester, 2007).

### *Validity*

CRESST has conducted a number of studies to collect evidence of the validity of Performance Assignments, including criteria validity, predictive validity, transferability, and fairness. They found that the Performance Assignments showed convergent and

predictive validity with CAHSEE, California Standards Tests, and Stanford Achievement Test (Goldschmidt, et al., 2007). They showed fairness, in that students performing better on the Performance Assignments performed better on both SAT-9 and on CAHSEE, regardless of student background (Martinez, et al., 2007).

### *Why did it end?*

The funding for the initiative ended. Simultaneously, the district moved on to other priorities.

### *Successes in Design, Professional Development, and Technical Evidence*

This initiative had the funding to hire district level support for staffing and professional development. It was designed and supported by academic leaders in assessment from a local renowned university. Because of the strong research interest of the university, technical evidence was gathered throughout the initiative leading to primary evidence of the technical quality of the Performance Assignments and the consistency of scoring. Not only were the Performance Assignments designed with deep knowledge of content, standards, and cognition, the evidence of their technical quality “supports the possibility that high-quality learning-centered assessment may again be a practical option for large-scale assessment and accountability” (Niemi, et al., 2007).

### *Challenges in Design, Professional Development, and Technical Evidence*

One troubling aspect of this initiative was that the district’s wants and needs changed every year, affecting the selection and implementation of grades and subjects for the Performance Assignments. Initially, the project was to span a range of content areas. In the end, only two subjects were covered.

## **Nebraska**

### *Context*

In 2001, in response to concerns about high stakes standardized testing, Nebraska Department of Education (NDE) developed “School-based Teacher-led Assessment and Reporting System,” or STARS. Teachers in the state wanted an assessment system that supported curriculum and instruction rather than accountability, they wanted to be the ones to develop the system, and they wanted the system to include multiple measures, not just an on-demand standardized test that many other states were adopting. For seven years, with the state Department of Education guidance, the state’s 500 districts designed their own assessment systems. In most states, funding for accountability went to a testing contractor to develop, field test, and administer a standardized test. In Nebraska, that analogous funding supported the staffing and professional development for meeting the state guidelines for quality assessment systems.

## *Design*

During Nebraska STARS implementation, districts developed their own assessment systems based on local standards or ones they adopted. The state developed broad guidelines for district assessment systems and communicated those guidelines both on-line and in supporting documents, including “Guidelines and Requirements for Documenting Assessment Quality for STARS” (Nebraska Department of Education, 2007). These “Guidelines” were written to guide districts’ development of local assessment systems. They also allowed the state to communicate its requirements for evaluating each district’s system for technical quality. Districts submitted documentation of their systems and samples of student work in grades 4, 8, and 11 for evaluation of technical quality along six criteria:

1. Assessments align to state or local standards
2. Students have opportunity to learn
3. Assessments are free from bias
4. The level is developmentally appropriate
5. Scoring is consistent and reliable
6. Mastery levels are appropriate to subject and grade levels

Some districts used corporate standardized tests, others used locally developed multiple choice tests. However, some districts took the opportunity to develop multiple measures, including portfolios and other performance assessments. In some districts that used performance assessments, these assessments were district-wide. For example, in one district, students in grades 4, 8, and 11 were given at least four common performance assessment tasks in each of four subject areas. These were given in an extended administration window.

During STARS, the state also implemented a statewide writing test as a common, on-demand performance assessment. The student writing was centrally scored by at least two trained teachers.

## *Professional Development*

Because of the focus on local assessment for learning, not for accountability, professional development on assessment literacy gained prominence during STARS. Schools developed cultures of collegiality and learning about curriculum, instruction, and assessment. In one district, all new teachers received a one day assessment workshop. In addition, teachers participated in summer workshops about assessment, group scoring and double scoring sessions, and on-site collaborative time. One research study showed that the attitudes towards STARS and the efficacy of local, performance based assessment extended beyond the reporting grades 4, 8, and 11 (Isernhagen & Mills, 2008).

Many districts have formed partnerships with the NDE, colleges and universities, and other school districts in order to support assessment literacy and professional development. In Nebraska, Educational Service Units are state agencies, accredited by the NDE, which support school districts through leadership, research, and development services. ESUs provided support for STARS in several districts. An independent consulting company, Buros Center for Testing, conducted all the district trainings for meeting technical quality criteria.

### *Technical Quality*

Each district was required to compile a District Assessment Portfolio annually. The guidelines for this portfolio were explicit about how to document evidence of each assessment quality criterion. Based on this evidence, portfolios were rated on a common scale. Each year, more districts achieved the top 3 of 5 ratings, so that by the end of STARS, the quality of all districts' local assessment systems was "good," "very good," or "exemplary."

The District Assessment Portfolios are not available publicly. Two district assessment leaders described their portfolios as including evidence of each of the quality criteria, including validity and reliability, for each common assessment in each grade and subject. All technical evidence for each district was gathered either by district personnel or by contracted agencies.

### *Why did it end?*

During the STARS years, Nebraska students scored among the top in the United States on national measures of academic achievement. However, policy-makers and some members of the public wanted district-to-district comparability, which a purely local system could not provide. Due to these concerns about comparability across districts, Nebraska legislators abandoned STARS and its local assessments in 2008 and became the last state in the nation to implement on-demand, state standardized tests in grades 3-8 and high school. In response to that changed legislation and its implications for education, the NDE Commissioner of Education left his role. Starting in 2009, standardized tests were being administered statewide and STARS was phased out.

### *Successes in Design, Professional Development, and Technical Evidence*

STARS was well funded, because all testing dollars from the federal government for NCLB were used for district staffing, partnerships, and professional development in assessment literacy to support the local assessment systems. These resources allowed the state to develop guidelines for District Assessment Portfolios, which provided explicit rationale and directions for documenting evidence of quality criteria for all districts. In addition, most districts employed either a district assessment coordinator or a local Educational Service Unit to facilitate the collection of data for the annual District

Assessment Portfolios. As a result of this support and the successful early implementation of STARS, school cultures changed and extended beyond the reporting grades, because teachers saw the effects and felt the benefits of performance assessments.

### *Challenges in Design, Professional Development, and Technical Evidence*

One challenge to STARS was that there was variability in how districts interpreted the quality guidelines. With autonomy for decisions about assessment, some districts chose standardized tests and did not choose to use multiple measures, including performance assessments.

In a state where technical quality and sustainability were largely achieved in STARS, the political will to maintain the proven system wavered in its final years. The pervasive climate created by NCLB, the policy makers' desire for comparability across districts, and elected officials acting upon the wishes of some constituents all contributed to questioning of STARS. The cost of moderation and external validation of assessment systems across districts, which would have addressed some of the comparability issues, was too high (Wood, Darling-Hammond, Neill, & Roschewski, 2007). As a result, legislators voted to end STARS and implement state-wide testing beginning the 2009-10 school year.

## **Rhode Island**

### *Context*

Rhode Island Department of Education wanted to develop a high school pathway using multiple measures that better prepared students for college and work, acknowledging the skills and knowledge that students bring to school. They also wanted to end the practice of counting courses taken as the requirement for high school graduation. Their initiative is called the Proficiency Based Graduation Requirements (PBGR). Starting in 2004, this scale-up effort is the most recently initiated in this review. Processes were developed to help schools support students in the performance assessment aspects of this graduation requirement. The first class to experience this form of accountability was the class of 2008.

### *Design*

The Rhode Island Department of Education has designated six core subjects in which students must be proficient by graduation. The content is guided by state and national standards. Local schools and districts develop their own assessment systems with guidelines about technical quality and other educational goals. In order to graduate from high school, students must pass their courses and complete the following:

- On-demand state standardized tests in ELA and math (students must pass)

- In each of six subjects, two of the following locally developed and scored performance assessments: exhibition, graduation portfolio, comprehensive course assessments

Exhibition is a public demonstration of in depth exploration and application in a content area. A graduation portfolio contains entries in each of the content areas, reflections on the body of work in each content area. Exhibitions and graduation portfolios allow students some choice in how they choose to demonstrate proficiency and in special interest areas. Comprehensive course assessments are end-of-course exams that must be at least 50% performance-based. According to the department’s documentation (S. Lee, personal communication):

The purpose of the performance-based component is to allow students to demonstrate the ability to apply several of the applied learning skills and demonstrate knowledge of a set of the larger concepts, “big ideas”, or unifying themes of the course content.

### *Professional Development*

Some technical assistance to high schools has been provided by field experts trained by a contractor and the state so that school staff understand the new graduation requirements and the performance-based components.

Each secondary school has common planning time built in to its schedule. During this time, teachers develop performance assessment tasks and associated rubrics and collaboratively look at student work together. Through this teacher-led work, teachers’ knowledge of and expectations from performance assessments have improved.

In addition, summer workshops are provided by the Rhode Island Department of Education on various aspects of assessment. Schools have formed networks with other schools based on their region to discuss work on the new graduation requirements across school sites. Some school staff participated in an earlier, pilot initiative to develop the assessment toolkits for common performance tasks and to develop portfolio and exhibition processes. These teachers have taken leadership roles in their schools and networks and will continue to provide technical support in their schools.

### *Technical Quality*

Each high school undergoes a peer review process administered by the state Department of Education, in which schools share their graduation requirement design and demonstrate that they have a rigorous and balanced assessment system. This review, which includes a portfolio and a day-long site visit, is scheduled to occur every two years. A report on the quality of the assessment system is submitted to the commissioner, with recommendations for improvement if the system is deemed insufficient.

The peer review includes the following criteria:

- Access and opportunity to learn
- Alignment to standards
- Sufficiency
- Fairness
- Standard setting
- Data use

Reliability and validity are aspects of the peer review and fall under the category of data use. The peer review process and studies about schools meeting these criteria are just being designed and developed. The first peer reviews are scheduled to be completed in 2010.

It is too early in Rhode Island's initiative to discuss successes and challenges.

## Synthesis of Individual Scale-Up Efforts

The previous section describes each locale’s efforts to develop performance assessments for use in accountability systems. This section brings these efforts side-by-side to compare and contrast elements of design, professional development, and technical quality.

### Design

#### *Standards, Subjects, and Grades*

Each scale-up effort chose the subjects, grade levels, and standards for accountability. All of the efforts based their assessment systems on their own state’s standards, except Nebraska, whose districts had the option to develop local standards. Most districts in Nebraska chose to adopt the Nebraska state standards. State standards provide a guideline for expectations for both content and skills.

The subject areas of reading/writing and math were included in all of the scale-up efforts. Queensland, NYPSC, Nebraska, and Rhode Island efforts extended beyond reading/writing and math to cover other core academic subjects. In some cases, like LAUSD, the original intent was to include more subjects. However, the original intent was reduced to two subjects when the district decided to focus on technical quality and professional development in greater depth.

Locales differed in their emphasis on grade levels. NYPSC and Rhode Island focus their performance assessment efforts on high school graduation requirements. Other locales included a selection of elementary, middle, and early high school grades, leaving accountability reporting to those grades. The most common grades for performance assessments were grades 4 and 8. What is clear is that none of the locales tried to implement their accountability reporting at all grade levels. In fact, the most that any state required for accountability was four grades.

### *Components*

Table 3: Components by Locale

Locale	State or national standardized on-demand test	Local performance assessment systems	Common extended time performance tasks	Common short performance tasks
Queensland, Australia	√ (National)	√	√	
Vermont	√	√		√
Kentucky	√	√	√	√

New York		√	√	
LAUSD	√			√
Nebraska		√		√
Rhode Island	√	√		

#### State standardized test, centrally scored

All but two sites use a large-scale, on demand standardized test, composed mostly of multiple choice items, as a component in their accountability system. Nebraska was the last state to add a standardized test, with the tests being phased in starting in 2008. One of the efforts, New York Performance Standards Consortium, is in a state that has a long history of standardized testing with the Regents exam. The schools in the Consortium, however, have obtained a waiver from state testing and therefore only use local assessment systems and common, extended time performance tasks.

#### Local assessment systems that include performance assessments

Local assessment systems can differ along a number of dimensions – how prescriptive the elements are, the diversity of assessments included, and number of assessments included. We summarize the local assessment systems from most prescriptive to least prescriptive.

The most simple local assessment system was in Kentucky during KIRIS, where every student across the state was responsible for compiling a writing portfolio with specific pieces to be included. For accountability, these portfolios were scored holistically, but not on the individual pieces.

Vermont also used student portfolios as the basis of their local assessment systems. However, the individual elements of the portfolios were not prescribed. Rather, guidelines are given to districts on what a quality portfolio contains. Portfolios did not necessarily include performance assessment tasks in Vermont.

The next level of prescriptiveness for local assessment systems is found in Rhode Island, where individual schools can choose two of three assessment options, all of which include performance assessments. The state provides quality guidelines but does not specify what an exhibition, graduation portfolio, or comprehensive course assessment should contain. The state plans to organize and carry out peer reviews of each school’s local system for rigor and quality.

Finally, NYPSC and Nebraska leave decisions about local assessment systems completely up to the schools and districts. NYPSC design components and external

review process encourage schools to embed performance assessments throughout their grade levels. Nebraska developed six technical quality criteria, leaving the onus on districts to prove the technical merit of their systems. In the case of Nebraska, some district local assessment systems did not include performance assessments.

**Common, extended time performance tasks with extended administration windows**  
Four locales incorporate(d) common, extended time performance tasks into their assessment systems: NYPSC, Nebraska, Kentucky, and Queensland. In the case of the New York Consortium schools, these common assessments form the basis of their students' graduation requirements – one assessment task each in English, history, math, and science. Students must pass all four performance assessments, in addition to their school requirements, to receive a New York diploma. In Nebraska, all student completed the statewide writing assessment, which was completed in two sessions over a two-week period.

Kentucky's common extended time performance tasks of KIRIS were called performance events. Performance events had several phases, such that some of the assessment was conducted in groups of students and some individual work. Performance events were changed each year in order for the assessments to remain authentic.

The Queensland common, extended time performance tasks are called Queensland Comparable Assessment Tasks (QCATs). They are both centrally and locally developed according to design briefs, which are available on-line and provide guidelines and requirements for the assessments. Teachers administer these tasks by certain deadlines in the school year and score them locally with moderation sessions. The sample QCATs available on-line take from 90 minutes to 3 weeks to complete.

**Common, short performance tasks administered on-demand**

On-demand assessments require that all students complete the assessment at the same time/date and thus ensure that items have not been seen by teachers or students beforehand. Three sites had common on-demand performance tasks of a writing prompt, centrally scored. However, they differed in the mode of delivery. In Vermont, the statewide writing assessment was only one writing prompt. In Kentucky, the writing prompt was part of the state test, which also included multiple choice items.

In LAUSD, performance assignments in writing involved reading, analyzing, and responding to literature in writing. These on-demand performance tasks were locally scored with a sample randomly scored by external scorers.

### **Professional Development**

The primary implementers of any assessment policy are teachers. Therefore, professional development to build their capacity to use performance assessment to inform their curriculum and instruction is a major part of any accountability system that includes

performance assessments. In the seven sites studied, professional development was conducted in a number of ways. We describe both the formats in which teachers were supported in building assessment literacy as well as the content of professional development work.

### ***Process and Format for Professional Development***

In all of the locales reviewed in this paper, external partnerships were formed to provide professional development and support the implementation of performance assessments. External partners included the respective state departments of education (or district assessment office in the case of LAUSD), higher education institutions, and non-profit organizations. Some sites used both departments of education and higher education partners.

In most locales, some of the professional development was facilitated by teachers for teachers. In most cases, teachers volunteered to support other teachers in assessment. Professional development took place during common planning time, in meetings of teachers collaboratively looking at student work, discussing tasks and scoring. Several sites conducted professional development workshops, both during the school year and during the summer. Finally, several sites formed networks among schools and teachers, a way to share practice beyond the classroom or the school.

### ***Content of Professional Development***

The content of professional development was to build basic assessment literacy, to design performance tasks, and to ensure that teachers across classrooms, schools, and districts showed consistency in scoring common performance tasks.

In Queensland, Vermont, LAUSD, and Rhode Island, teacher professional development also involved development of standards-based performance tasks. In some districts in Nebraska, every teacher new to the district participated in a daylong assessment literacy workshop.

In Kentucky, Queensland, and NYPSC, professional development included teachers working together to compare, discuss, and come to consensus on the scoring of student work from performance assessments.

In one site, Queensland, web-based technology is used to provide information and professional development to teachers through an assessment bank, which included guidelines for administration and scoring, assessment tasks, and student work.

### ***Technical Quality***

The most prominent aspects of technical quality in the literature were reliability and validity. Most of the locales studied reported upon or had plans to report upon evidence of reliability and validity of performance assessments. We note that other technical

qualities are also important. Some locales, such as Nebraska and Rhode Island, collect evidence of other technical qualities.

### Reliability

Reliability evidence is “the degree to which scores are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual student” (American Educational Research Association, et al., 1999). Evidence of reliability was reported for all of the locales that have completed reliability studies, but the primary data was only available for examination in Vermont and LAUSD. The primary data in other places resides in technical reports, which are not available to the public for the initiatives that are no longer running.

Table 4: Reliability Evidence Available for Review, by Locale

Locale	Reliability Evidence
Queensland, Australia	No for senior assessment; Evidence being gathered for QCATs
Vermont	Yes
Kentucky KIRIS	No
NYPSC	Evidence being gathered
Los Angeles Unified School District	Yes
Nebraska STARS	No
Rhode Island	Evidence being gathered

Across the locales studied, the reliability (Linn & Baker, 1996) of performance assessment scoring was approached in two main ways: scoring by more than one person and moderation sessions.

#### *Scoring by more than one person*

Performance assessment results were scored in some locales by two raters – one local (though not necessarily the student’s teacher) and someone else. The second rater is often called the external rater and can be another teacher in the student’s school, another teacher from a different school or district, or an expert at a central scoring session/location. Second raters typically score a random selection or proportion of student work rather than each piece. Vermont, Kentucky, Nebraska, and LAUSD used external raters to collect data on reliability. In three of the four cases, reports are/were that initial reliability of scoring between raters was low but improved over time. In Vermont, reliability of scoring between teachers not the students’ own and second raters was low, although it improved over time. In Kentucky, reliability of scoring between students’ own teachers and second raters was low, with teachers scoring their students higher than external raters prior to professional development in scoring. In LAUSD, score agreement

between local and central scorers was low at first and improved over time. Reliability of scoring of the district portfolios according to technical quality guidelines was done in Nebraska; scorer reliability was high (Isernhagen & Mills, 2008).

*Moderation sessions*

Two locales, Queensland and NYPSC, are using an approach called “moderation,” a multi-step process that promotes and documents consistency in teachers’ judgments about whether or not students have met the expectations. In Queensland, moderation is used for all of the high school exit requirements, and in NY, for the common performance tasks required for graduation. Moderation processes bring groups of teachers together to come to consensus on standards, assessment components, and the judging of student work. Moderation relies on individuals’ judgments rather than on statistical tests. One reason it is used less widely in the US might be that those in charge of accountability policy must trust that teachers who are doing the judging are rigorous and knowledgeable (Brian Gong, personal communication).

*Validity*

Validity is the extent to which an assessment gives information about what the assessment was designed to “test” (Linn & Baker, 1996). Evidence of the validity of an assessment may be gathered in a variety of ways; in other words, there is no one test or procedure for validating an assessment. Validation of assessments should be designed to collect evidence about (American Educational Research Association, et al., 1999):

- Content of the assessment (content validity)
- Relationship of the assessment to other variables (convergent and predictive validity)
- Generalizability of the assessment
- Consequences or fairness of the assessment

An assessment is not valid or not valid, but valid for making an inference or giving a consequence. The table below summarizes whether or not evidence was found for the validity of making inferences from and consequences for the assessment systems in each locale, along with the type of evidence published.

Table 5: Validity Evidence by Locale

Locale	Validity Evidence
Queensland, Australia	Content validity
Vermont	Convergent validity: Weak relationship to National Assessment of Educational Progress
Kentucky KIRIS	None

NYPSC	Predictive validity: Evidence of positive relationship to college going and college grades
Los Angeles Unified School District	All four types of validity evidence shown in peer reviewed journals
Nebraska STARS	Content validity Convergent validity: evidence of relationship to other tests
Rhode Island	Evidence being gathered

Only in LAUSD were all four types validity evidence collected for the Performance Assignments, and of the rigor required for publication in a peer reviewed journal. In four other locales, one type of validity evidence was cited for each – content validity for Queensland, convergent validity for Vermont and Nebraska, and predictive validity for NYPSC.

### Current status

The three scale-up efforts included in this review that are currently still operating are Queensland, Australia, New York Performance Standards Consortium, and Rhode Island. The other four scale-up efforts ended for various reasons, such as political pressure for centralized accountability (Nebraska), difficulty demonstrating reliability and validity (Kentucky and Vermont), and funding issues/priority changes (LAUSD). Despite the vision, expertise, and resources invested in scaling up the use of performance assessments, achieving the optimal mix of public support, technical quality, and resources is difficult to maintain.

## The Use of Performance Assessments in Other Fields

### Introduction

Because of the long history of performance assessments in fields outside of public K-12 education, considerable information can be gleaned from reviewing this literature. Through the administration of licensure exams, many professions ascertain the skills and abilities of those planning to join the profession. Fields such as medicine and law administer extended examinations. Other fields such as real estate, nursing, teaching, accounting, psychology, and truck driving also administer exams routinely.

Some of these licensure exams carry high stakes, i.e., require a passing score to obtain a license to practice in that profession. For example, one cannot become an attorney without receiving a passing score on the bar exam. However, in other fields, the stakes are lower. While graduating with a college degree in accounting will permit one to

become an accountant, a passing score the CPA exam is required to become a certified public accountant (CPA).

Some professional exams are on-demand exams composed entirely of multiple choice items. However, others employ formats such as essays and open response items in addition to multiple choice items. A substantial amount of information exists regarding several professional licensure exams that contain performance assessments. We summarize relevant information from teaching, medicine, and law.

## **National Board Certification**

### *Context*

The National Board for Professional Teaching Standards (NBPTS) was formed in 1987 by a coalition of teachers, policymakers, and academic and corporate leaders. Their work is guided by five core propositions.

1. Teachers are committed to students and their learning.
2. Teachers know the subjects they teach and how to teach those subjects to students.
3. Teachers are responsible for managing and monitoring student learning.
4. Teachers think systematically about their practice and learn from experience.
5. Teachers are members of learning communities.

This coalition wanted to provide a national, voluntary system for certifying teachers who meet rigorous standards for what accomplished teachers should know and be able to do. To fulfill this part of their mission, National Board Certification was created.

### *Design*

National Board Certification is a process in which teachers demonstrate their teaching practice through portfolio entries and six computerized assessment exercises. Because this process is designed to be a powerful professional development experience, the entire process typically takes 1-3 years to complete. As of November 2004, slightly more than 40,000 educators had earned National Board Certification. To be eligible to participate, educators must meet three criteria:

1. Possess a bachelor's degree from an accredited institution,
2. Have completed three years of teaching experience, and
3. Held a valid state teaching license for each of these three years.

The portfolio portion of the certification process contains four different entries. Three of these entries require evidence of teaching, such as a video clips from a lesson or actual student work samples, as well as a written reflection describing and analyzing the portfolio entry. One of the entries may focus on student work samples, and two of the

entries must include a video of a lesson showing interactions between the teacher and the students. The fourth entry must capture the teacher's work with families and the larger community as well as with colleagues and the profession at large. All evidence for classroom-based portfolio entries must be gathered during the 12 months prior to submitting the portfolio.

In addition to the portfolio, all candidates must demonstrate their content knowledge by completing six computer-based exercises. Candidates are given 30 minutes to complete each exercise. These exercises are specific to the chosen area of teacher certification (e.g. early adolescent mathematics). For example, for one exercise in mathematics, a candidate will be required to demonstrate knowledge of data analysis of a given data set. The expectation is that the candidate will create appropriate graphical representations of the data as well as analyze and interpret the data provided. These assessment exercises are developed and scored by practicing professionals.

### ***Professional Development***

Unlike professional development for teachers in assessment literacy in K-12 education, professional development in teacher certification refers to the training of people who score the teacher portfolio entries. Raters are teachers or school counselors who have been recruited, have completed an intensive training workshop, and have qualified as an assessor prior to scoring (National Board for Professional Teaching Standards, 2009). Carefully trained raters use detailed four point rubrics to score these portfolios.

### ***Technical Evidence***

While evidence regarding the reliability or validity of the National Board Certification was not publicly available, NBPTS conducts analyses on an annual basis to assess inter-rater reliability. The 2009 Guide to National Board Certification states "the results of these reliability analyses have consistently indicated that NBPTS assessors are making reliable, accurate, and fair evaluations of candidates' responses." However, some research studies challenge the impact of National Board Certification on student progress (Sanders, Ashton, & Wright, 2005; Vandevort, Amrein-Beardsley, & Berliner, 2004).

## **United States Medical Licensing Examination (USMLE)**

### ***Context***

Oral exams have been used for hundreds of years to assess the skills of those completing medical training. However, due to psychometric and logistical issues, oral exams were eliminated in the U.S. as a component of the medical licensing exams more than 40 years ago (Hubbard, 1971).

Since then, in response to dissatisfaction with traditional multiple-choice tests, other performance components have been embraced in the health professions. One form of

performance assessment is using the standardized patient, an individual trained to portray a real patient with a set of symptoms or health problems. While standardized patients have been used in a number of residency programs throughout the U.S. since the mid-1980s, they did not become part of the US Medical Licensing Exam (US MLE) until 2004. In 2006-07, approximately 17,000 examinees took the US MLE.

### *Design*

Passing the U.S. MLE is required to practice medicine in the U.S. This exam is completed over multiple days and includes three steps. While each step contains multiple choice items, steps 2 and 3 also include a performance-based assessment component.

The performance assessment in step 2 uses standardized patients to assess clinical skill proficiency through 12 physician-patient encounters. Raters observe live encounters or review digital recordings to score the performance on three domains: a) clinical skills, b) communication and interpersonal skills, and c) spoken English proficiency.

Step 3 assesses proficiency through nine computer-based case simulations (CCS) which include contextual information about the patient. Because the assessment utilizes interactive software, individuals may select from the full range of diagnostic tests and therapeutic options found in laboratory, hospital, or other clinical settings. As a result of these choices, the computer-simulated patient may improve or regress. The patient may also develop complications that will require an appropriate follow-up response.

### *Professional Development*

In order to minimize measurement error, well-defined rubrics and carefully trained raters who are physicians are important aspects of the assessments. Physician raters are also trained to score the candidate's patient note summarizing the clinical information obtained during the visit with the standardized patient.

A critical aspect of professional development involves the recruitment and training of standardized patients. Standardized patients learn to portray the personality, emotions, body language, and, in many instances, physical symptoms of a specific patient. Part of the training also involves the case specific checklists that standardized patients complete, recording the history-taking questions asked and the physical exam maneuvers performed.

### *Technical Evidence*

Over time, evidence regarding the technical quality of these performance-based assessments has been documented. With attention to assessment design and professional development, acceptable levels of reliability have been achieved with performance-based clinical exams such as those that utilize standardized patients (Carraccio & Englander, 2000). For example, generalizability studies have been conducted to determine the

sampling design needed (e.g., number of tasks, number of raters). Studies have shown that increasing the number of standardized patients increases the generalizability exponentially (Swanson, Clauser, & Case, 1999). Furthermore, there is evidence that assessments utilizing standardized patients may test a component of clinical skills not evaluated using other measures. Therefore, the use of standardized patients in the medical field is viewed as an important contribution to the assessment of clinical competence (Dupras & Li, 1995; Papadakis, 2004).

## **Bar Examination**

### *Context*

Historically, the development, administration, and scoring of bar exams have fallen under the domain of states. Early in the 20<sup>th</sup> century, due to concerns about the variability in bar admission standards from state to state surfaced, the National Conference of Bar Examiners (NCBE) was formed to help states develop their bar exams. Beginning in the 1970s, the NCBE developed the Multistate Bar Exam (MBE), a six hour multiple choice exam, to cover a broader range of topics than can be adequately addressed in a few essay questions. Currently, nearly every state and U. S. territory uses the MBE as one component of their state bar exam. The MBE measures an applicant's ability to analyze legal issues arising from certain facts. More recently, NCBE added other components to the bar exam process for states to use if they so desire, including a performance assessment component called the Multistate Performance Test (MPT). The MPT evaluates an examinee's ability to handle tasks typically required of a lawyer. In 2005, thirty-one jurisdictions administered the MPT as one component of the bar exam for their state or jurisdiction. A few states that do not use the MPT have created their own performance tests that are similar to the MPT. California is one state that has both done so and documented its results. Slightly more than 8,000 applicants took the 2005 California Bar Exam which includes this performance assessment.

### *Design*

One criterion required to practice law in nearly every state is passing the state Bar Exam. The exam is completed over multiple days and usually includes multiple components. While one component, the Multistate Bar Examination (MBE), contains only multiple choice items, the other two components are performance-based assessments: an essay exam and a case study assessment. The essay portion of the exam, which is comprised of six essay questions, is intended to measure the same skills as the MBE component.

The third component developed by the NCBE, the Multistate Performance Test (MPT), is a set of performance tasks to test the applicant's skills in legal analysis, fact gathering, fact analysis, tactics and problem solving, ethics, and communication. For each task, the applicant is provided with all of the material to complete the task, including the

instructions, the client file, and law library materials relevant to the case. The task requires legal analysis, ethical considerations, and the creation of a product such as a trial brief or case plan that resolves the client's problem.

The relative weight of each component is left to the state. In California, scores on three components are combined, with each individual component carrying different weights in the final determination of pass/fail: MBE at 35%, essay exam at 39%, and the performance test at 26%.

### ***Professional Development***

Each state must identify graders and provide appropriate training in the scoring process for its own Bar Exam. The NCBE provides technical assistance to states for this process. In California, to identify graders for the Bar Exam, the Committee of Bar Examiners solicits applications from attorneys who are in good standing with the Bar and passed the California Bar Exam on the first or second attempt. Once graders are selected, they attend an orientation session and three calibration meetings, where scoring is discussed. Compensation is provided to the graders for both the training as well as the actual scoring of the performance assessments.

### ***Technical Evidence***

While the NCBE provides grading guidelines, each jurisdiction is responsible for scoring the bar exam. Thus, conducting reliability and validity studies would fall under the control of each individual state's bar association. Evidence of the technical quality of each state's Bar Exam was not publicly available.

### **Conclusions from the Use of Performance Assessments for Licensure in Other Fields**

Before synthesizing the lessons learned of using performance assessments in these fields, it is important to acknowledge several differences between K-12 public education and teaching, medicine, and law. In the cases of law, medicine, and teaching, the candidates taking the exams are a smaller group of people than students in public education. Therefore, the number of assessments administered each year is substantially less than would typically be administered by a state education agency. In addition, the candidates represent a more homogenous group because they have already attained high levels of education through graduate and professional schools. So, while these fields have grappled with the logistics of balancing the costs (financial as well as human capital) with the rich information gained from performance assessments, in public education, these costs would be magnified.

Another difference related to the cost, and therefore sustainability, of scaling up performance assessments is the source of funding for administering the assessment system. The candidates who are completing the teaching, medicine, and law assessments pay a substantial fee to take them.

Finally, performance assessments in law and medicine are high stakes for individuals. These assessment systems began with the development of a low stakes system while conducting research over time to ensure the technical quality of a higher stakes system. However, these assessment systems were never designed for accountability purposes for groups of individuals, such as schools. In K-12 public education, most of the efforts reviewed in this paper describe the use of performance assessments in high stakes annual accountability reporting for students as well as schools. The scale of such an effort, and the professional development required, far surpass that of the other fields.

Despite these differences, because of the long history of performance assessment in medicine and law, and more recently teaching, much can be learned from the experiences in these fields about scaling up performance assessments. Swanson, Norman, and Linn (1995) indicate there are several important lessons from the health professions that can shed light on this work in public education. These professions have introduced performance assessments which simulate real world situations and have been able to elicit rich information about large numbers of candidates that is not available through multiple choice exams. However, assessment design and scoring is complex and often problematic. Some of the lessons from other fields applicable to task design and scoring in K-12 education include:

- Evidence of performance assessment validity must be documented and include examination of threats to validity, such as comparison with scores of expert raters or interviews with students about the task requirements.
- Due to the relatively small number of independent performance tasks administered, assessments must change from year to year for security reasons while maintaining equal difficulty
- A sufficient number of tasks that appropriately sample the breadth of content that one desires to generalize about must be included in the assessment
- Holding everything else constant, increasing the number of tasks a student is given is more critical than increasing the number of raters per task
- To ensure appropriate levels of inter-rater reliability, the quantity and quality of the professional development provided to raters is crucial
- Scoring is based on judgment and therefore the assessments and scoring must be free from bias.

## Discussion

In this synthesis of scale-up efforts, a number of cross-cutting findings emerged. A discussion of notable successes, challenges, and policy and practice implications follows.

### Successes in Efforts at Scaling Up Performance Assessments

The benefits to teachers of using performance assessments in the studied locales were made clear in interviews and published articles. Not only did teachers' knowledge and understanding of assessment improve through the use of performance assessments in their classrooms, but they also noted how this work led to improvements in their instruction and curriculum. For example, teachers changed the types of questions they asked in class, and having assessment tasks helped teachers plan backwards and change their curriculum. In addition, teachers reported improved collegiality in their buildings due to the conversations and sharing encouraged by the use of performance assessments.

In many locales, systems for ensuring reliability were built into the performance assessment initiatives. These systems typically included scoring of student work by educators who were external to teachers' classrooms, and often to their schools. Even local scoring was done in many cases externally, by other teachers in the building or by teachers in informal and formal school networks. In some cases, external scoring was conducted for reliability checks, on a sample of student work rather than on every piece.

Finally, most of the scale-up efforts showed improvement in technical quality over time. In other words, when reliability of scoring was low, or the number of tasks was low, or teacher buy-in or assessment literacy was low, over time, these challenges abated. These initiatives showed that technical quality can improve in the course of a few years, and that once teachers begin to understand and use performance assessments, their enthusiasm for them increases.

### Challenges in Efforts at Scaling Up Performance Assessments

#### Technical Quality

While there were challenges in a number of areas related to the implementation of performance assessment, most of the challenges were directly related to the technical quality of the performance assessments. In only two locales was evidence of the reliability of scoring available for review. In a few others, reliability evidence was cited but not shown. The literature discussed the difficulty in building consistency of scoring across teachers, both within schools and across schools. There was evidence that students' own teachers tended to score student work higher than external reviewers.

In several locales, evidence of the validity and quality of the performance assessment tasks lacked rigor. In only one locale, Los Angeles, was the gathering of reliability and

validity data part of the design of the initiative. Despite the publication of Standards for Educational and Psychological Testing (American Educational Research Association, et al., 1999), there is little consensus on how to conduct reliability and validity studies for performance assessments. Although the evidence of technical quality from the US efforts is weak, other countries besides Queensland, Australia provide important examples technical quality from which the states in the US may learn (Darling-Hammond & Pecheone, 2009).

### **Resources for Professional Development**

One reason for the discontinuation of several of the initiatives discussed was the resource-intensiveness of implementing performance assessments at scale. Each stage of implementation – from the design of performance assessments to implementation, to professional development, to scoring, to the collection of data that serves as evidence for the assessment’s quality – could be costly in terms of staffing and professional development. These costs must be taken into account in designing sustainable accountability systems with performance assessments.

### **Political Will**

Finally, the lack of or loss of political leadership and will to include performance assessments in accountability systems caused the elimination of several scale-up efforts. Even when evidence of technical quality was strong or improving, Vermont, Kentucky, LAUSD, and Nebraska ended promising initiatives due to lack of political support.

## **Lessons and Implications for States, Districts, Schools, and Intermediary Organizations**

### **Design of Assessment and Accountability Systems**

This review has systematically discussed several elements of the scale up of performance assessment initiatives. All efforts attempted to use performance assessments in ELA and in math; some were more ambitious and tackled additional core academic subjects. Given the complexity and breadth of the work, all of the efforts began with a selection of grade levels rather than all grades. Some started with elementary grades while others have focused on high school exit grades. Interestingly, even in grades where the performance assessments were not required, when teachers learned of these initiatives, they were eager to use them.

Six out of the seven locales in this review included local assessment systems. While there was a range in what assignments those local assessment systems entailed, all but one permitted schools to have autonomy over the system’s individual assignments. Given the diversity of school missions, student composition, and staffing across districts and states, this autonomy in choosing the elements of a local assessment system makes sense. However, districts and schools must be given guidance as to the parameters of that

system. In other words, what are crucial elements and characteristics of a local assessment system? These guidelines should be communicated to schools and districts before any initiative begins, and its results should be reviewed through moderation or external review.

From this investigation, the terms “performance assessment” or “performance task” took on new complexity. We found that while each locale labeled its initiative as one that included performance assessments, they took on different forms. One aspect of complexity is about how the performance task is administered – on a certain day or during an extended period of time. At one extreme, every student in a grade took the same district performance assessment on the same day, with paper and pencil. At the other extreme, students worked towards culminating projects of their choice over the course of their high school careers, and these culminating projects could take multiple forms. We began to term performance assessments as “on demand” and “extended time” to differentiate between the two types of task administration.

Each type of performance assessment has its advantages and disadvantages. For example, on-demand performance assessments usually take less time to implement, are administered under more secure conditions, and can address some 21<sup>st</sup> century skills like problem solving or creative thinking. However, it is difficult to truly embed a district-wide on-demand performance assessment in individual classroom curriculum. A common, on-demand performance assessment is also less likely to address students with different learning and testing styles, since it is less likely to accommodate non-traditional modalities such as visual and oral communication.

On the other hand, extended administration window performance tasks also have their advantages and disadvantages. Due to the extra time usually given for these assessments, they may be curriculum-embedded, interdisciplinary, and use materials and resources beyond paper and pencil. For example, a research paper or an experiment would have all of these characteristics and also allow students who learn and test differently to use other modalities.

Extended time performance tasks at scale, for accountability purposes, presented challenges in all of the locales in which they were tested. Because of the flexibility in when the assessments are given and how long they take, it is difficult to control the administration environment, such as the amount of teacher input and revision, the amount of help from other adults, and the equity of materials and resources available for students. One way to begin to control these variables is to produce detailed administration guides for teachers. Such performance tasks must be different from year to year so that teachers do not start “teaching to the task,” yet they also must be of equal difficulty from year to year. Scoring products of extended time performance tasks is also more complex than on-demand, paper and pencil assessments because they often involve more than writing –

they might involve visuals, oral presentations, and props. Local scoring is necessary due to the expense of centralized scoring, but for local scoring to reach high standards, teachers require professional development on scoring and moderation sessions for checking the accuracy of their scoring.

Clearly, short and extended time performance tasks have their advantages and disadvantages. Short performance tasks have more secure administration conditions while still assessing some skills not measured by fixed response tasks. Extended time performance tasks elicit more information about student learning because they require more creation and analysis than short performance tasks do. Therefore, deliberate decisions about the length of performance tasks for accountability must be made by system designers.

In summary, the following *design implications* for performance assessment scale up efforts emerge:

- Be explicit about the purpose of performance assessments; use the purpose to determine whether common performance assessments are given on-demand or in an extended administration window
- Begin scale-up initiatives with common performance assessments in a couple of content areas and grades rather than all content areas and grades
- Monitor and address the burden of administering local assessment systems and common performance tasks in the same school year, especially when they are first implemented
- Provide teacher administration and scoring guides for common performance assessments
- Support review of assessment systems and tasks through external review by other teachers or expert panels
- Provide autonomy to schools and districts around what their local assessment systems will look like within the context of design guidelines
- Provide quality criteria that schools and districts must reach for each performance assessment, whether it is a part of the local assessment system or is an individual common performance task.

### **Professional Development**

In this review, we found that districts and states had a number of ways to provide professional development about assessment to teachers. Performance assessments require additional expertise that teachers and school leaders, and even district personnel in some cases, may not have acquired in their schooling or on-the-job. External partners provide expertise to teachers in the design of assessment tasks and accompanying administration and scoring guides and the collection of evidence of meeting quality criteria. Professional

development partners included universities, state departments of education, state service-providing organizations, and non-profits.

Not only are pre-service and externally provided professional development important, so are the ongoing, day-to-day discussions that happen in a school about assessment.

Therefore, schools using performance assessments must have schedules that permit groups of teachers and staff members (such as grade level teams and subject area teams) to meet, plan, and discuss teacher assignments and student work.

Local assessment systems and common performance tasks may be developed *with* teachers rather than *for* teachers. Several locales in this review involved teachers in the development of their assessment systems. This type of involvement not only benefited the districts and states, but also the teachers and their colleagues, as there was more ownership for the performance assessments when it was a teacher-driven initiative.

When assessment drives curriculum and instruction, and teachers are invested in the design of the school's assessment system, a natural positive consequence can be that the school culture becomes a collaborative learning community over time. Strategies for improving the technical quality of performance assessments involve teachers collaboratively and iteratively discussing their task assignments and student work products, resulting in improved curriculum, instruction, and assessment. The resources of time, energy, and facilitation and training to initiate these processes are significant. Most schools and districts that undertake performance assessments require intense professional development support in assessment literacy from district assessment coordinators or university partners (Knight and Gray, personal communication). Teachers consistently share practice, look at teacher assignments and student work together, and reflect and collaborate in an ongoing way (Isernhagen & Mills, 2008). Over time, the culture of schools and districts changes so that teacher planning time is focused on assessment, which interacts with curriculum and instruction. As these practices become embedded in the daily lives of teachers, they become more routine, meaningful, and less burdensome. In this way, they become sustainable.

While some argue that performance assessments can have excessive resource costs in professional development and scoring, the benefits are substantial - teachers are more skilled, engaged, and effective (Darling-Hammond & Pecheone, 2009). Thus, the long-term benefits to teachers and students can offset the up-front professional development costs.

The lessons from this review suggest that using performance assessments in schools and districts requires intentional pre-service and in-service *professional development*:

- Use external partnerships to provide and facilitate professional development
- Assure that all new teachers become assessment literate

- Review the successes and challenges of scaled up performance assessment initiatives
- Include teachers in the design of and professional development for performance assessments
- Structure school days so that teachers have time to plan and debrief assignments and discuss student work, scoring, and performance assessment revision.

### Technical Quality

In order for performance assessment results to be used in accountability systems, evidence about their technical quality must be measured and documented. Technical quality involves a number of attributes, the most prominent of which are termed reliability and validity. These technical qualities have guided the standardized testing industry and have been documented the most explicitly. Tests of reliability, or consistency of scoring, are chosen depending upon the nature of the assessment. In general, primary evidence of reliability was not accessible for this review. In several locales, such evidence was referenced in papers that were publicly accessible. When we were able to talk with individuals in some locales, they shared samples of unpublished data that was submitted as evidence of reliability. In several sites, collection of reliability evidence is underway and has not yet been published. Given the importance of reliability to the use of performance assessments in accountability, the lack of accessible reliability evidence was problematic.

Similarly, validity evidence for performance assessments may include a broad range of data and procedures for collecting and analyzing data. In the locales studied in this review, primary validity evidence was only accessible for one locale. In four other locales, publicly accessible documents referenced validity data gathered. Even when validity evidence was gathered, there was no clear consensus on the types of validity evidence necessary to meet quality criteria. In two cases, it was content validity; in two cases, it was about the relationship of the performance assessments to other assessments; and in one case, it was about the relationship of the performance assessments to future outcomes. In a couple of sites, the lack of good reliability evidence hindered the ability of reviewers to judge validity.

Despite the enthusiasm of teachers and other educators for performance assessments *for learning*, few rigorous quantitative studies have been conducted on the impact of performance assessments on instructional improvement and student outcomes. When researchers have studied the technical merits of performance assessments in accountability systems, the assessments have fallen short in their early years on evidence of reliability and validity (Koretz, 1998). In addition, there has been no consensus in the field on how it should be collected, what types of evidence would suffice, or where and how it should be documented or reported. The evidence has also shown that with time

and professional development, reliability can improve and validity can be achieved. Accountability programs must make gathering technical quality evidence a part of their development.

In summary, much more guidance for schools, districts, and states is needed in the area of technical evidence. While knowledge and expertise about technical quality of performance assessments exists, it has not been made accessible to practitioners. A body of peer-reviewed evidence showing high quality scaled up performance assessment initiatives will help policy-makers and the public take note of the possibilities in using performance assessments in state level accountability. Therefore, advocates of performance assessments must invest in collecting and documenting evidence of technical quality in their initiatives. The following are *technical quality implications* for these advocates and practitioners:

- Policy incentives from states or districts should encourage focus and attention to technical quality of performance assessments.
- Develop technical quality guidelines for scale-up efforts with performance assessments. For each type of performance assessment, what must sites do for evidence of technical quality?
- Implement local assessment systems and common performance tasks in a way which integrates the collection of data for research and inquiry about their technical qualities.
- Reliability evidence must be collected, analyzed, and reported regularly in order to understand whether or not there is consistency in judgments about student work on performance tasks.
- Validity evidence must be collected, analyzed, and reported regularly in order to understand whether or not the performance tasks elicit the information they were meant to assess.
- State departments of education and the US Department of Education should build in incentives for states to develop local assessment systems that meet quality guidelines

Overall, in developing an assessment system that includes performance assessments, designers should use a multiple measures approach because neither traditional testing nor performance-based assessments are a panacea. All testing formats have drawbacks; using multiple measures allows for a more complete picture of a student's skills and abilities to form. Above all, selection of the assessment method(s) should depend on the skills to be assessed and the purpose of the assessment. Some skills are more appropriate to be assessed using a performance task while others can be easily and more efficiently assessed with a pencil and paper format. In addition, designers must be cognizant of the nature of the stakes attached to an assessment. When high stakes decisions about

individuals or schools will be made based on the results of the assessment, one must pay very close attention to the technical adequacy of the assessment components. This focus could mean limiting or precluding some assessments, or tailoring them to the high stakes situation – or it could mean changing how high-stakes decisions are made, bringing them in line with the Joint Organizational Statement on No Child Left Behind (NCLB) Act ([http://www.edaccountability.org/Joint\\_Statement.html](http://www.edaccountability.org/Joint_Statement.html)). Furthermore, while studies of the intended and unintended benefits and consequences of administering a high stakes accountability system are seldom conducted, studies of this type are greatly needed to better understand the impact of these high stakes decisions on the larger system as a whole.

The benefits of performance assessments as low stakes tools informing teachers about individual student knowledge and skills are undisputed. However, bringing them to scale for use in accountability systems, which are high stakes for district, schools, and students, is complex and perceived as costly in human and other resources. This review found seven locales that have boldly introduced performance assessments, both to inform instruction and to report on student outcomes. In each locale, there were clear benefits to students, teachers, and administrators. The main challenges the educators faced were the technical quality of performance assessments, the cost of necessary professional development for sustainability, and political support. Through their collective experiences summarized in this review, it is clear that we have the knowledge, experience, and capacity to tackle each of the challenges. These cases provide a wealth of information about the possibilities and challenges of performance assessment systems that have already been implemented. We should apply these experiences and lessons towards the next generation of accountability systems.

## Glossary<sup>3</sup>

21<sup>st</sup> century skills: skills that take into account the global economy, technology, and changing workforce requirements. These skills include complex thinking, analytical skills, computer skills, creativity, media literacy, and cross-cultural skills.

Accountability: A concept encompassing the responsibility of an educational system to the public, which includes students.

Administration window: The span of time or point in time when an assessment task is given to students.

Anchor paper: A sample of student work demonstrating a given level of proficiency, by which scorers may judge student work.

Assessment: Any systematic method of obtaining information about knowledge and skills and using it to draw inferences about people or programs.

Assessment literacy: The ability to communicate, express ideas and opinions, and problem solve about the use of assessments.

Bias: A systematic flaw in an assessment that differentially affects the performance of certain groups of students.\*

Central scoring: Assessments are judged by trained raters at one location, usually not within the school or district where the assessments were administered, nor by teachers who administered the assessments.

Cognitive demand: The level of difficulty of the mental activities, such as processing, organizing, and retaining information, needed to complete an assessment.

Common assessment task: The same task is administered to all students in a grade and subject across multiple classrooms or schools or districts.

Concurrent validity: Evidence that an assessment gives similar results to another assessment given at the same time.

Content validity: Evidence of the degree to which an assessment tests the content it is designed to test.

---

<sup>3</sup> Definitions with a \* are adapted from Standards for Educational and Psychological Testing (American Educational Research Association, et al., 1999).

Equating: Ensuring that two or more assessments cover the same content and outcomes share a common scale.

Exhibition: A type of performance assessment that requires students to show their work in a public forum, describe their process and product, and answer questions from viewers.

Extended time administration window: The span of time when an assessment task is given to students is a period of weeks or months, depending what best fits the teacher's or school's curriculum.

Extended time assessment task: Those assessments that can not be completed in one sitting and that require more time, group work, or the creation of a substantial product.

External scoring: Assessments are judged by trained raters, who may be teachers or others, from outside the school or district in which the assessment was administered.

Fairness: The principle that every student should be assessed in an equitable way.\*

Fixed response: An assessment task that requires a student to choose an answer rather than construct one. Multiple choice items are a type of fixed response task.

Formative assessment: Generates information that the teacher feeds back to the student for reflection, refinement, and improvement.

Generalizability: The usefulness of an assessment outcome to forming conclusions or making predictions about student knowledge and skills.

High stakes: Used to provide results that have important, direct consequences for individuals, programs, or institutions (American Educational Research Association, et al., 1999)

Inter-rater reliability: The consistency with which two or more judges rate the work or performance of assessment takers.\*

Large scale accountability: The system that reports upon how well the public educational system is meeting its responsibility to educate all students

Local: At the individual, classroom, or school level

Local assessment system: The set of assessments that a school uses to collect information about what its students know and can do.

Local scoring: Assessments are judged by trained raters who are teachers from the same school or district in which the assessment was administered.

Low stakes: Used to provide results that have only minor or indirect consequences for individuals, programs, or institutions\*

Moderation studies: A process by which teams of teachers from across schools and districts review and judge student work in order to improve and document evidence of reliability.

Multiple measures: More than one way of assessing outcomes.

On-demand administration window: There is one point in time when an assessment task is given to students in controlled situations (typically classrooms) on the same day throughout a school or district. Students must complete the assessments in one sitting.

Open-ended response: The answer for an assessment task requires students to construct a short response, write an essay, or create a product based on a project or experiment.

Performance assessment: Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied (American Educational Research Association, et al., 1999)

Predictive validity: Evidence of how accurately an assessment can predict future outcomes.

Reliability: The degree to which assessment outcomes for a group of students are consistent over repeated administrations of the assessment.\*

Scale up: Application across multiple individuals, classrooms, schools, or districts of a particular task, system, or program.

Scorer drift: The way in which scores on a particular assessment change over time. Likely reasons for scorer drift include changes in the educational programming, teacher knowledge, or student familiarity with format.

Short assessment tasks: Assessments which take less than 30 minutes to complete.

Sufficiency: Assessment contains enough items to measure the standards it is intended to measure.

Summative assessment: Generates information about the efficacy of curriculum and instruction.

Sustainability: Ability to hold up over time

Technical quality: The aggregate evidence that an assessment gives accurate and useful information about what students know and can do.

Transferability: The process of applying the results of an assessment to a different situation.

Validity: The degree to which accumulated evidence and theory support specific interpretations of assessment scores entailed by the proposed uses of the assessment.\*

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing* (5th ed.). Washington, DC: Author.
- Carraccio, C., & Englander, R. (2000). The objective structured clinical examination: A step in the direction of competency-based evaluation. *Archives of Pediatric Adolescent Medicine*, *154*, 736-741.
- Darling-Hammond, L., & Pecheone, R. L. (2009). Reframing Accountability: Using Performance Assessments to Focus Learning on Higher-Order Skills *Meaningful Measurement: The Role of Assessments in Improving High School Education in the Twenty-First Century* Washington, D.C.
- Darling-Hammond, L., & Wood, G. H. (2008). Refocusing Accountability: Using Performance Assessments to Enhance Teaching and Learning for Higher Order Skills. Retrieved from [www.forumforeducation.org](http://www.forumforeducation.org)
- Duncan, A. (2009). American Recovery and Reinvestment Act Letter to Governors. Washington, DC: Education Department.
- Dupras, D. M., & Li, J. T. (1995). Use of an OSCE to determine clinical competence. *Academic Medicine*, *70*, 1029-1034.
- FairTest (2008). President-Elect Barack Obama on No Child Left Behind Retrieved 12/21/09, 2009, from <http://www.fairtest.org/presidentelect-barack-obama-no-child-left-behind>
- Fontana, J. (1995). Portfolio Assessment: Its Beginnings In Vermont and Kentucky. *NASSP Bulletin*, *79*(573), 25-30.
- Foote, M. M. (2007). Keeping Accountability Systems Accountable. *Phi Delta Kappan*, *88*(5), 359-363.
- Forum on Educational Accountability (2007). Assessment and accountability for improving schools and learning: Principles and recommendations for federal law and state and local systems (pp. 55).
- Freebody, P. (2005). *Background, rationale and specifications: Queensland Curriculum, Assessment and Reporting Framework*: Queensland Department of Education.
- Goldschmidt, P., Martinez, J. F., Niemi, D., & Baker, E. L. (2007). Relationships Among Measures as Empirical Evidence of Validity: Incorporating Multiple Indicators of Achievement and School Context. *Educational Assessment*, *12*(3), 239-266.
- Gong, B., & Reidy, E. F. (1996). Assessment and Accountability in Kentucky's School Reform. In J. B. Baron & D. P. Wolf (Eds.), *Performance-Based Student Assessment: Challenges and Possibilities* (Vol. I, pp. 215-233). Chicago, IL: The National Society for the Study of Education.
- Grade-by-Grade State Testing Policies (2006). *Quality Counts at 10: A Decade of Standard-Based Education*, from <http://www.edweek.org/media/ew/qc/2006/17sos.h25.ggtp.pdf>
- Hubbard, J. (1971). *Measuring Medical Education*. Philadelphia: Lea and Febiger.
- Isernhagen, J. C., & Mills, S. J. (2008). *Charting STARS Year 7 Report: Engaging Conversations* (Final Evaluation Report). Lincoln, NE: Nebraska Department of Education.
- Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing Performance: Designing, Scoring, and Validating Performane Tasks*. New York: Guilford Press.
- Koretz, D. (1998). Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education: Principles, Policy & Practice*, *5*(3), 309.

- Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994a). *The Evolution of a Portfolio Program: The Impact and Quality of the Vermont Program in Its Second Year (1992-1993)* (Technical Report No. 385). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994b). The Vermont Portfolio Assessment Program: Findings and Implications. *Educational Measurement: Issues and Practice*, 13(3), 5-16.
- Linn, R. L., & Baker, E. L. (1996). Can Performance-Based Student Assessments Be Psychometrically Sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-Based Student Assessment: Challenges and Possibilities* (Vol. Part I). Chicago, IL: University of Chicago Press.
- Martinez, J. F., Goldschmidt, P., Niemi, D., Baker, E. L., & Sylvester, R. M. (2007). Language Arts Performance Assignments: Generalizability Studies of Local and Central Ratings. *Educational Assessment*, 12(3), 267-282.
- Mills, R. P. (1996). Statewide Portfolio Assessment: The Vermont Experience. In J. B. Baron & D. P. Wolf (Eds.), *Performance-Based Student Assessment: Challenges and Possibilities* (Vol. I, pp. 192-214). Chicago, IL: The National Society for the Study of Education.
- National Board for Professional Teaching Standards (2009). *Guide to National Board Certification*.
- Nebraska Department of Education (2007). *Guidelines & Requirements for Documenting Assessment Quality for STARS*. from <http://www.nde.state.ne.us/assessment/documents/GuidelinestoDocumentAQMASTER2.pdf>.
- Niemi, D., Baker, E. L., & Sylvester, R. M. (Writer) (2007). Scaling Up, Scaling Down: Seven Years of Performance Assessment Development in the Nation's Second Largest School District [Article], *Educational Assessment*: Lawrence Erlbaum Associates.
- Papadakis, M. A. (2004). The Step 2 Clinical-Skills Examination. *The New England Journal of Medicine*, 350(17), 1703-1705.
- Pearson, P. D., Calfee, R., Webb, P. L. W., & Fleischer, S. (2002). *The Role of Performance-Based Assessments in Large-Scale Accountability Systems: Lessons Learned from the Inside. Technical Guidelines for Performance Assessment*.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing What Students Know: The science and design of educational assessments*. Washington, DC: National Academy Press.
- Queensland Department of Education (2005). Queensland Curriculum, Assessment and Reporting Framework.
- Queensland Studies Authority (2009). P - 12 Assessment Policy. In Q. S. Authority (Ed.) (pp. 4). Brisbane: State of Queensland.
- Sanders, W. L., Ashton, J. J., & Wright, S. P. (2005). *Comparison of the Effects of NBPTS Certified Teachers with Other Teachers on the Rate of Student Academic Progress*.
- Silva, E. (2008). Measuring Skills for the 21st Century. *Education Sector Reports*, 11. Retrieved from [http://www.educationsector.org/research/research\\_show.htm?doc\\_id=716323](http://www.educationsector.org/research/research_show.htm?doc_id=716323)
- Stecher, B. (1998). The Local Benefits and Burdens of Large-scale Portfolio Assessment. *Assessment in Education: Principles, Policy & Practice*, 5(3), 335 - 351.
- Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., Jarvin, L., & Sharpes, K. (2009). Using the theory of successful intelligence as a framework for developing assessments in AP physics. *Contemporary Educational Psychology, In Press, Corrected Proof*.

- Stiggins, R. (2008). Assessment Manifesto: A Call for the Development of Balanced Assessment Systems. In E. A. T. Institute (Ed.) (pp. 1-12). Portland, OR: Educational Testing Service.
- Swanson, D. B., Clauser, B. E., & Case, S. (1999). Clinical Skills Assessment with Standardized Patients in High-Stakes Tests: A Framework for Thinking about Score Precision, Equating, and Security. *Advances in Health Sciences Education, 4*, 67-106.
- Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National Board Certified Teachers and Their Students' Achievement. *Education Policy Analysis Archives, 12*(46), 1-117.
- Viadero, D. (2009). Student Testing: Using Assessments to Improve Learning and Student Progress. *Education Week*. from <http://www.edweek.org/ew/articles/2009/05/20/32report-4.h28.html>.
- Wood, G., Darling-Hammond, L., Neill, M., & Roschewski, P. (2007). *Refocusing Accountability: Using Local Performance Assessments to Enhance Teaching and Learning for Higher Order Skills*. from <http://www.fairtest.org/files/PerformanceAssessments.pdf>.
- Zane, T. W. (2009). Performance Assessment Design Principles Gleaned from Constructivist Learning Theory (Part 1). *TechTrends: Linking Research & Practice to Improve Learning, 53*(1), 81-90.

The mission of the Center for Collaborative Education (CCE) is to transform schools to ensure that all students succeed. We believe that schools should prepare every student to achieve academically and make a positive contribution to a democratic society. CCE partners with public schools and districts to create and sustain effective and equitable schools.

#### ACKNOWLEDGEMENTS

This study was made possible with generous support from the Nellie Mae Education Foundation.

We thank the following people for their time, thoughtfulness, and resources as we conducted the research for this paper: Eva Baker, Rosemary Burns, Jennifer Chidsey Pizzo, Doug Christensen, David Conley, Ann Cook, Marcia Cross, Linda Darling-Hammond, Janina Drazek, Martha Foote, Brian Gong, Cindy Gray, Kyle Hartung, Joan Herman, Karin Hess, Stuart Kahl, Neal Kingston, Mary Knight, Sharon Lee, Scott Marion, Charis McGaughy, Monty Neill, Ray Pecheone, David Ruff, Kristin Russo, Roy Seitsinger, Robert Sternberg, Rick Stiggins, David Swanson, Phyllis Tashlik, Art Thacker, and Kit Viator.

#### CITATION

Tung, R. and Stazesky, P. (2010). (2010). *Including Performance Assessments in Accountability Systems: A Review of Scale-Up Efforts*. Boston, MA: Center for Collaborative Education.



33 Harrison Street  
Boston, MA 02111  
617.421.0134 phone  
617.421.9016 fax  
[www.cce.org](http://www.cce.org)



1250 Hancock Street, Suite 205N  
Quincy, MA 02169  
781.348.4200 phone  
781.348.4299 fax  
[www.nmefdn.org](http://www.nmefdn.org)