Center for Collaborative Education

Nellie Mae Education Foundation
Opening Doors to Tomorrow

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

January 2010

Executive Summary

The purpose of this literature and field review is to understand previous efforts at scaling up the use of performance assessments across districts and states.[1] Through systematic description and comparison of seven large-scale initiatives, as well as analogous efforts from teacher certification, medicine, and law, the paper identifies the strengths and vulnerabilities in each initiative. In addition, this paper is the first to evaluate the reliability and validity evidence gathered in each effort. The review concludes with implications and recommendations for districts and states as to some standards and procedures that will support the success of using performance assessments in accountability systems.

This executive summary and the full report may be downloaded at www.cce.org.

In K-12 education, "performance assessments" set forth expectations for students and require them to:

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

In looking at each effort to implement performance assessment, we considered these guiding questions:

- How were performance assessments included in state and district accountability systems?
- What was the nature of the professional development provided to support assessment literacy and to successfully introduce performance assessments in state accountability systems?
- How was the technical quality of each accountability system assured?
- Taken together, what lessons can we learn from the reviewed scale-up efforts?

The seven scale-up efforts examined in this review were located in:

- Vermont
- Kentucky
- New York
- Los Angeles
- Nebraska
- Rhode Island
- Queensland, Australia

Findings and Implications

The benefits of performance assessments as low stakes tools informing teachers about individual student knowledge and skills are undisputed. 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. However, bringing such assessments to scale for use in accountability systems, which are high stakes[2] for district, schools, and students, is complex and perceived as costly. Thus, the use of performance assessments in accountability systems is still relatively rare in the U.S. For such assessments to scale up to the district or state level, this study found a number of conditions in the areas of design, professional development, technical quality, and political support that maximize chances for success. The seven locales in this study provide valuable information and lessons for the development of future district or state level accountability systems that include performance assessments.

Other professional fields that use Performance Assessments for high stakes decisions:

- Law
- Medicine
- Teacher Certification

Design

The Massachusetts Education Reform Act of 1993: 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…

Performance assessments have historically been included in district or state level assessment systems in two major ways: as common assessment tasks across schools and in local (school or district level) assessment systems. In studying the assessment components that each locale developed in its scale-up effort, a recurrent theme was to start implementation in a few grades and subjects rather than all at once. In addition, schools and districts were given autonomy for their school- or district-designed assessment systems as well as strong guidance for the critical components of an assessment system and its associated tasks. The following design implications for performance assessment scale-up efforts emerged:

- Scale-up initiatives should begin with common performance assessments[3] in a couple of content areas and grades, rather than in all content areas and grades.
- Assessment system designers must be explicit about the purpose of performance assessments; they must use the purpose to determine whether common performance assessments are given on-demand[4] or in an extended window of time[5].
- Schools and districts should 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.
- Teachers should have administration and scoring guides for common performance assessments in order to develop shared understanding and increase consistency in scoring.
- Schools and districts need to have autonomy around what their local assessment systems will look like within the context of explicit design guidelines.
- States and districts should develop quality criteria for each performance assessment, whether it is a part of the local assessment system or is a common performance task.

Professional Development

When assessment drives curriculum and instruction, and teachers are invested in the design of the school's assessment system, a natural consequence is that the school's culture becomes a collaborative, professional learning community. Professional development strategies in both the study locales as well as in other performance assessment initiatives included providing time for teachers to collaboratively discuss assignments and student work and to document and improve the technical quality of performance assessments, resulting in improved curriculum, instruction, and assessment. As these practices were embedded in the daily lives of teachers, they became routine, meaningful, and less burdensome. Teachers across the study locales became more skilled, engaged, and effective with performance assessments. In this way, using performance assessments school-wide became sustainable over time.

Districts and states had a number of ways to provide professional development about assessment to teachers. External partners such as universities and non-profit service organizations provided expertise to teachers in the design of assessment tasks, in developing administration and scoring guides, and in the collection of evidence for meeting quality criteria.

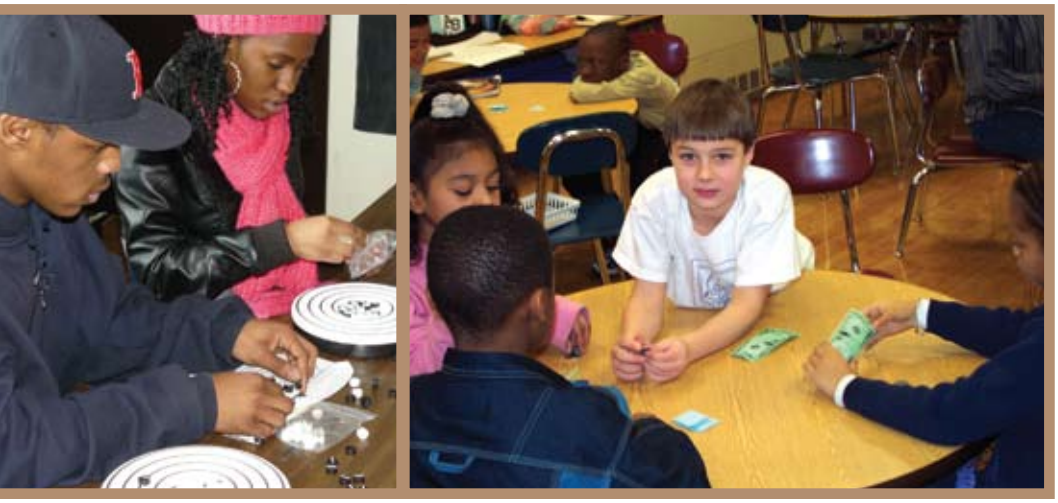
The lessons from this review suggest that using performance assessments in schools and districts benefits from pre-service and in-service professional development:

- External partners, such as universities and technical assistance organizations, should provide and facilitate professional development to build assessment literacy, construct rubrics for assessments, conduct scoring sessions, and document technical quality.
- All new teachers should be provided professional development in assessment literacy.
- School staff should learn about other initiatives that have used performance assessments for accountability purposes.
- Teachers should participate in the design of professional development for performance assessments.
- School days should be structured so that teachers have time to plan and debrief assignments and discuss student work, scoring, and performance assessment revision.

9th grade math performance assessment: 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.

Technical Quality

In order for performance assessment results to be used in accountability systems, their technical quality must be documented and measured. Technical quality involves a number of attributes, the most prominent of which are termed reliability[6] and validity[7]. Clear evidence of reliability and validity will enhance the acceptance of including performance assessments for high stakes accountability. In most locales, scoring consistency improved significantly over time with appropriate guidance, professional development, and practice among teachers. Evidence of validity was not well documented in all study locales. One challenge to educators using performance assessments is that there is little consensus on the types of validity evidence necessary to meet quality criteria.



7th grade science performance assessment: 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.

In short, there must be a body of peer-reviewed evidence showing high technical quality in scaled up performance assessment initiatives. This evidence will support educators and policymakers in efforts to include multiple measures in state level accountability systems. 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 focus attention to technical quality of performance assessments.
- Quality guidelines must be developed for scale-up efforts with performance assessments. For each type of performance assessment, these guidelines would specify what sites must do for evidence of technical quality.
- Data on the technical quality of local and common performance assessment systems must be collected and analyzed by practitioners in an ongoing way.
- 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 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.

Political Support

Inadequate political support for performance assessments in accountability systems undermined several strong scale-up efforts, even when evidence of technical quality was strong or improving. Vermont, Kentucky, Los Angeles, and Nebraska ended promising initiatives due to lack of political support or active political opposition. In view of the need for political support, we recommend paying close attention to the following conditions:

- Performance assessments in accountability systems should begin on a smaller scale than state-wide, to ensure that technical quality is achieved and documented.
- Experts in assessment and assessment literacy should be a part of the staff at state departments of education, district central offices, and within schools.
- Each state's political leaders and education policy experts should support research that documents the technical quality of its accountability system. In particular, the medium-term benefits to students (through college) and to teachers of using performance assessments at scale must be understood and documented.
- Political and education leaders must consistently articulate the value of using multiple measures in assessing what students know and can do through publicizing and seeking recognition for successful implementation of performance assessment.

No Child Left Behind Act of 2001: The NCLB Act called for "multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding…"

In summary, 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 that 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.

Reference

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). Standards for Educational and Psychological Testing (5th ed.). Washington, DC: American Educational Research Association.

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.

ENDNOTES

[1] This executive summary and the full report may be downloaded at http://www.cce.org.

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

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

[4] 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.

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

[6] Reliability: The degree to which assessment outcomes for a group of students are consistent over repeated administrations of the assessment (American Educational Research Association, et al., 1999).

[7] Validity: The degree to which accumulated evidence and theory support specific interpretations of assessment scores entailed by the proposed uses of the assessment (American Educational Research Association, et al., 1999).

CITATION

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

Center for Collaborative Education

33 Harrison Street
Boston, MA 02111
617.421.0134 phone
617.421.9016 fax
www.cce.org

Nellie Mae Education Foundation
Opening Doors to Tomorrow

1250 Hancock Street, Suite 205N
Quincy, MA 02169
781.348.4200 phone
781.348.4299 fax
www.nmefdn.org