Bulletin
May 2011 | Issue 20
www.pearsonassessments.com

Performance-based Assessment: Some New Thoughts on an Old Idea

Emily R. Lai

Our field has devised many terms to describe assessments in which examinees demonstrate some type of performance or create some type of product (e.g., performance, performance-based, "authentic," constructed response, open-ended). Whatever you call them, performance-based assessments (PBAs) have a long history in educational measurement with cycles of ups and downs. And once again, PBAs are currently in vogue. Why? To address the federal government's requirements for assessment systems that represent "the full performance continuum," the two consortia formed in response to *Race to the Top* funding have both publicized assessment plans that involve a heavy dose of performance-based tasks (PARCC, 2010; SBAC, 2010). Thus, PBAs are relevant to any discussion about the future of testing in America.

The purpose of this bulletin is to review arguments in favor and arguments against the use of PBAs to assess student achievement in light of proposed test score uses.[1] In addition, the paper will make recommendations for moving forward in the design of "next generation" assessment systems that incorporate performance-based tasks. Current plans dictate that such systems must track student growth, measure students' readiness for college and the workforce, and provide a wide variety of information for teachers and policy-makers: formative information for making instructional adjustments and summative information for supporting inferences about individual student proficiency and teacher effectiveness.

There are many old arguments in favor of PBAs, which educators, policy-makers and parents continue to find compelling. Proponents claim these types of tests are more motivating to students (Hancock, 2007). They provide a model for what teachers should be teaching and students should be learning (Baron, 1991). They serve as professional development opportunities for teachers involved in developing and scoring them (Borko et al., 1997). They constitute complex, extended performances (Baron, 1991) that allow for evaluation of both process and product (Messick, 1994). Moreover, performance-based tasks provide more direct measures of student abilities than multiple-choice items. They are able to assess students' knowledge and skills at deeper levels than traditional assessment approaches and are better suited to measuring certain skill types, such as writing and critical thinking (Frederiksen, 1984). They are more meaningful because they are closer to criterion performances, constituting representations of "criterion activities valued in their own right" (Linn, 1993, p. 9)

> Moreover, there are ways of making PBAs, even those used for classroom purposes, more reliable and comparable.

Despite their recent renaissance, PBAs have well-known limitations: lower reliability and generalizability than selected-response items, primarily because of differences in efficiency between the two task types (one hour of testing time buys you many fewer performance-based tasks than multiple-choice items). But these limitations also arise because PBAs are frequently scored by humans—a process that introduces a certain amount of rater error (Dunbar et al., 1991). In exchange for greater depth of content coverage, PBAs compromise breadth of coverage (Messick,

[1] For a comprehensive review of the issues associated with performance-based assessments, see Lane & Stone, 2006.

1994). Generalizability studies of PBAs have found that significant proportions of measurement error are attributable to task sampling, manifested in both person-by-task interactions and person-by-task-by-occasion interactions in designs that explicitly model the occasion facet (Shavelson et al., 1999). Again, this is largely because there are many fewer performance-based tasks on any given test.

PBAs are used in a variety of contexts, including summative, high-stakes contexts, such as certification and licensure, as well as employment and educational selection. PBAs are also used for formative or instructional purposes. When well-designed PBAs are administered and scored in the classroom, they can provide valuable information for evaluating and improving instruction when tasks have high fidelity to important criterion performances and when they are designed to align with instruction (Lane et al., 2002).

In high-stakes contexts, strict standardization of task development, administration, and scoring is critical for promoting comparability, reliability, and generalizability (Haertel & Linn, 1996). In classroom assessment contexts, such rigid standardization may be relaxed. Clearly, what makes a particular PBA useful for one context will make it less so for the other. For example, strict standardization of task development, administration, and scoring (which is impractical in classroom settings anyway) makes assessment less amenable to organic adjustment by the teacher to support the learning and motivation of students with different needs. In turn, the unstandardized procedures typically favored in classroom settings—extended administration time, student choice of tasks, student collaboration—can introduce construct-irrelevant variance and diminish the comparability of tasks that is necessary for supporting high-stakes inferences about student proficiency or growth (Linn, Betebenner & Wheeler, 1998; Webb, 1993).

PBAs are here for the foreseeable future. If past experience is any lesson, individual PBAs will almost certainly prove less reliable than traditional assessment approaches. However, supporters would argue that this compromise in reliability means an upgrade in terms of greater construct validity for skills not easily assessed using traditional approaches. Furthermore, “next generation” assessment approaches that distribute assessment opportunities throughout the school-year (e.g., through-course assessments) may help to partially offset low reliability of a single PBA by taking a composite over multiple assessment tasks and occasions (Wise, 2011).

Moreover, there are ways of making PBAs, even those used for classroom purposes, more reliable and comparable. For example, thoughtful reflection on the construct to be assessed (Messick, 1994), coupled with carefully-crafted test specifications (Haertel & Linn, 1996, can go a long way in creating comparable tasks. Although the measurement field has traditionally avoided classroom assessment, certain groups have begun participating in collaborative initiatives to create curricula with psychometrically-sound, embedded PBAs (Furtak et al., 2008).

Doing this well requires new assessment development models that incorporate close collaboration between curriculum designers and assessment developers to ensure tight alignment and seamless integration of assessment and instruction. Such models also require closer collaboration between the content specialists who write the tasks and the psychometricians charged with collecting evidence to support overall assessment quality. Finally, such embedded assessments will need to be piloted along several dimensions: to investigate task and rubric performance, to examine the cognitive processes students use to complete the tasks, and to collect student responses for anchoring performance scoring rubrics. In addition, we will also need to obtain feedback from teachers regarding assessment functionality and usefulness.

It’s a brave, new world of assessment. To truly advance and sustain these developments, we need to start thinking in brave, new ways. Such an approach will help ensure that the current wave of performance assessment has more staying power than the last.

A previous version of this manuscript was published as a TrueScores blog entry: http://www.truescores.com/2011/03/performance-based-assessments-brave-new.html.

References

Ayala, C. C., Shavelson, R. J., Araceli Ruiz-Primo, M., Brandon, P. R., Yin, Y., Furtak, E. M., Young, D. B. & Tomita, M. K. (2008). From formal embedded assessments to reflective lessons: The development of formative assessment studies. *Applied Measurement in Education, 21*(4), 315-334.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4*(4), 305-318.

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education, 13*(3), 259-278.

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

Frederiksen, F. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(3), 193-202.

Haertel, E. H. & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 59-78). Washington, D.C.: U.S. Department of Education.

Hancock, D. R. (2007). Effects of performance assessment on the achievement and motivation of graduate students. *Active Learning in Higher Education, 8*(3), 219-231.

Lane, S., Parke, C. S., & Stone, C. A. (2003). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and student performance. *Educational Assessment, 8*(4), 279-315.

Lane, S. & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th Ed.) (pp. 387-424). Westport, CT: Praeger.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1-16.

Linn, R. L., Betebenner, D. W., & Wheeler, K. S. (1998). *Problem choice by test takers: Implications for comparability and construct validity*. (CSE Tech. Rep. No. 482). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Partnership for Assessment of Readiness for College and Careers. (2010). Application for the Race to the Top Comprehensive Assessment Systems Competition. Retrieved from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf

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

SMARTER Balanced Assessment Consortium. (2010). Race to the Top assessment program application for new grants. Retrieved from http://www.k12.wa.us/SMARTER/RTTTApplication.aspx

Webb, N. M. (1993). *Collaborative group versus individual assessment in mathematics: Group processes and outcomes*. (CSE Tech. Rep. No. 352). Los Angeles: University of California, Center for the Study of Evaluation.

Wise, L. L. (2011, February). Picking up the pieces: Aggregating results from through-course assessments. Paper presented at the Invitational Research Symposium on Through-Course Summative Assessment. Atlanta, GA.

About Pearson

Pearson, the global leader in education and education technology, provides innovative print and digital education materials for preK through college, student information systems and learning management systems, teacher licensure testing, teacher professional development, career certification programs, and testing and assessment products that set the standard for the industry. Pearson's other primary businesses include the Financial Times Group and the Penguin Group. For more information about the Assessment & Information group of Pearson, visit http://www.pearsonassessments.com/.

About Pearson's Research Bulletins

Pearson's research bulletins help clarify specific assessment-related topics for educators, parents, students, researchers and policy makers through brief and non-technical discussions. Pearson's publications in .pdf format may be obtained at: http://www.pearsonassessments.com/research.