PERFORMANCE ASSESSMENT:

A VEHICLE FOR IMPROVING THE UTILITY AND VALIDITY OF LOCAL AND STATE ASSESSMENT SYSTEMS

Education Commonwealth Project Spring 2023

Performance assessment: A vehicle for improving the utility and validity of local and state assessment systems

Education Commonwealth Project

University of Massachusetts Lowell
Coburn Hall 222, 850 Broadway St.
Lowell, MA 01854
EdCommonwealth.org

Authors

Sanford R. Student, Research Associate, sandy@lyonsassessment.com
Susan Lyons, Technical Adviser, susan@lyonsassessment.com
Dan French, Director of Performance Assessment, danvfrench@gmail.com

About the Education Commonwealth Project

The Education Commonwealth Project (ECP) works to support assessment of student learning and school progress that is valid, democratic, and equitable. Pushing back against the overreliance on standardized testing, ECP offers free and open-source resources that all schools can use. And thanks to support from the Massachusetts State Legislature, ECP offers additional support for public schools and districts in Massachusetts.

Introduction

The Education Commonwealth Project (ECP) works with Massachusetts school districts to support the use of performance assessment as part of their local assessment systems. But what is performance assessment? Performance assessment is a broad category of assessments of student learning centered around students' production of original work in the context of an open-ended task. Performance assessment is often discussed in contrast to standardized tests that emphasize multiple-choice items in which students *identify* correct responses. Performance assessment, on the other hand, consists of a smaller number of extended tasks in which students *generate* solutions to problems. The use of performance assessment has been studied in both classroom and large contexts. In this brief, we review the research on performance assessment to outline the ways in which performance assessment can benefit all students and may be especially beneficial for addressing persistent inequities in assessment for marginalized and/or minoritized students–Black, Latinx, low-income, and English learner students as well as students with disabilities. Of course, as with any meaningful disruption to standard practice, the design and implementation of performance assessment does not come without its challenges, which are also discussed.

What is a performance assessment?

Performance assessment has been defined in multiple ways. Drawing on the literature, we offer the following definition.[1,2] Performance assessment consists of:

1. An extended *task* in which students have opportunities for *sensemaking* and *problem-solving* and/or *original thinking* in the context of a *phenomenon* or *unresolved question*.
2. A method of *capturing student work* that is *open-ended* and *generative*, designed to represent not only a solution, but also the student thinking that underlies that solution.
3. *Evaluation criteria* that describe how different aspects of students' work can be connected to substantive conclusions about what they know and can do.

In the classroom, the primary benefit of performance assessment is providing meaningful opportunities for students to demonstrate their higher-order problem-solving skills. These demonstrations make the evidence of student learning more visible than assessments using primarily multiple-choice items, and therefore more instructionally useful. In a large-scale assessment context, where policymakers are more interested in group outcomes than the scores of individual students, performance assessment represents a path to improving the

validity of claims made about what students know and can do, particularly as content standards are increasingly emphasizing complex thinking over rote content knowledge.[3,4] At its core, performance assessment is about providing students with opportunities to make sense of problems and phenomena in ways that are authentic to the discipline.

Benefits

Research indicates that performance-based approaches to assessment can be beneficial to students and educators in a classroom context. There are several mechanisms by which performance assessment can support high quality learning for all students.

First, performance assessment can be a mechanism for students to meaningfully engage with content in the context of solving complex, real-world problems. This process of applying what one has learned to address complex problems is at the core of essentially every academic subject; for example, recent standards in social studies and science have emphasized critical inquiry and synthesis of new ideas over rote context knowledge.[4,5,6] Across these disciplines, performance assessment can play a key role, both in signaling to students what kinds of thinking are valued in the discipline and in giving students the opportunity to engage in authentic, higher-order thinking that supports their learning.

When teachers feel pressure to teach to heavily multiple-choice tests, the breadth and rigor of instruction typically suffers.[7] In contrast, instruction designed to support student performance on performance assessment involves emphasizing transferable skills and complex thinking. While the pressures of current test-based accountability systems may appear to disincentivize this type of teaching, there is evidence from New Hampshire that students in schools focused on performance assessment do at least as well on year-end standardized tests as their peers, if not better.[8] Looking to long-term outcomes, performance-based approaches to assessment in New York were associated with improved outcomes beyond standardized testing including higher graduation rates and success and persistence in college.[9]

By providing students with more opportunities to make their thinking visible, performance assessment in the classroom is likely to be particularly beneficial for historically marginalized students: students of color, English language learners (ELLs), low-income students, and students with disabilities, among others. For example, Noble et al. found that for low-income and/or linguistically minoritized students, multiple-choice test items tended to systematically underrepresent the capabilities they were able to display in an open-ended context

addressing the same content.[10] Similarly, a team found that it was difficult to elicit ELLs' full capabilities in science using multiple-choice items, despite several linguistic changes intended to support them.[11] In contrast, Llosa provides a case study of performance assessment designed around the strengths of ELL students which provided them more equitable opportunities to demonstrate their learning, while Fine & Furtak note that open-ended design is a key component of equitable classroom assessment of ELL students.[12,13] Evans found preliminary evidence of association between the use of performance assessment and improved assessment outcomes for students with disabilities.[8] There is little research in the literature on performance assessment that explicitly disaggregates effects for students on the basis of race/ethnicity. Still, there is reason to believe that performance assessment may hold similar promise for equitably assessing students of color–students whose learning may be systematically undervalued by multiple choice items, especially when those items are devoid of authentic and/or cultural context.[14,15]

At the state year-end testing level, the use of authentic, curriculum-embedded performance assessment has not been systematically investigated since the 2001 passage of No Child Left Behind effectively ended the burgeoning use of performance assessment in primary and secondary large scale assessment. Yet, evidence from the 1990s indicates that large-scale performance assessment systems do hold appeal for educators: Koretz found that when Vermont piloted a performance-based state assessment system built around student portfolios, educators responded positively, to the point that several schools expanded portfolios beyond grades where they were required.[16]

In contrast, the primarily multiple-choice standardized tests that states introduced in the wake of NCLB have been associated with pervasive teaching to the test, including narrower curriculum and more didactic pedagogy.[7] Large-scale assessment signals to educators what is valued in instruction, especially when accountability systems reward or label schools or teachers based on test scores. Concerns persist that large-scale standardized tests have the potential to derail innovations in curriculum and instruction.[17,18] Research has shown that low-quality curricula and teaching to the test are particularly pervasive in schools that serve predominantly students of color, the schools that are most likely to be labeled as underperforming by the state and even subject to state takeover.[19] To counteract these perverse incentive structures, investment in the development and piloting of large-scale testing systems that leverage performance assessment as the primary means of evaluating learning would provide meaningful signals for schools and districts to follow suit.

Challenges

Although performance assessment holds a great deal of promise, there are challenges associated with the construction and use of authentic tasks that require sense-making and problem-solving and/or original thinking in the context of a phenomenon or unresolved question.

First, performance task design is complex and time-consuming. Early work on performance assessment identified time to develop tasks as a key challenge. Teachers also need professional support to develop high-quality classroom tasks.[20] Of course, traditional standardized tests involve significant costs with less potential educational benefit. Because task development from scratch can be costly in a performance assessment context, the Education Commonwealth Project supports the dissemination of the MCIEA Performance Task Bank that offers freely available, high-quality performance tasks that are ready for schools and educators to use or adapt. Computer-based assessment technology also has the potential to lower the barrier to entry to introducing performance assessments in locales where, for example, funds may not be available to purchase lab equipment and/or physical space is in high demand.

Second, not all performance tasks are equal in their educational value and quality: open-ended tasks that focus too narrowly on a specific component of the discipline, such as a science task that is open-ended but requires only rote content knowledge to complete, may not be successful in eliciting authentic thinking.[21]

Third, performance assessment alone is no guarantee of more equitable assessment practice.[22] Assessment designers must consider the sociocultural context when designing performance tasks. Any assessment, performance or not, that neglects context may reify negative stereotypes, reinforce existing power dynamics, or misalign with the lived experiences and assets of students from non-dominant cultural backgrounds. A lack of attention to students' linguistic needs can have the same effect; fortunately, performance assessments can be uniquely flexible across languages compared to typical monolingual standardized tests. Thus, support in the form of professional development and communities of learning are necessary in any context where educators are working towards equitable design and implementation of performance assessments.[20]

Finally, it is difficult to generalize from a single performance task to a student's knowledge, skills and abilities across an entire domain, and meaningful score differences can occur across different raters of the same work.[23,24,25,26] To address this, professional learning communities of teachers might score the same pieces of student work together and discuss scoring

differences to attain consistent ratings across classrooms if scores are part of students' course grades. In contexts where teachers want to survey a student's knowledge, skills and abilities in a domain broadly, one might use multiple tasks broken up over time and draw upon evidence of learning beyond just formal assessments.

At the state level, implications are a bit different, and updated empirical evidence on the potential use of performance assessments in large-scale assessment is urgently needed. Accountability systems are often concerned with aggregate scores, and it may be the case that lower individual-level reliability (the extent to which any single score is representative of how the student would perform on other tasks targeting the same material) is a reasonable tradeoff in exchange for richer evidence of learning, as high individual reliability is not necessarily required to achieve high aggregate reliability.[27,28] It is also possible that tasks that use computer-based administration or are based on common templates and rubrics may result in more reliable scores.

Conclusion

Performance assessments represent a potentially powerful alternative to standardized assessment formats. In both small- and large-scale assessment, performance tasks are likely to provide richer and more authentic evidence about what students know and can do, yielding more accurate inferences and providing more meaningful information to support instruction and learning for students across cultural and linguistic contexts.

Notes

[1] Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, *33*(10), 1045–1063. https://doi.org/10.1002/(SICI)1098-2736(199612)33:10<1045::AID-TEA1>3.0.CO;2-S

[2] Solano-Flores, G., & Shavelson, R. J. (2005). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice*, *16*(3), 16–24. https://doi.org/10.1111/j.1745-3992.1997.tb00596.x

[3] Student, S. R., & Gong, B. (2022). Supporting the interpretive validity of student-level claims in science assessment with tiered claim structures. *Educational Measurement: Issues and Practice*. Advance online publication. https://doi.org/10.1111/emip.12523

[4] Wertheim, J., Osborne, J., Quinn, H., Pecheone, R., Schultz, S., Holthuis, N., & Martin, P. (2016). *An analysis of existing science assessments and the implications for developing assessment tasks for the NGSS*. https://scienceeducation.stanford.edu/sites/g/files/sbiybj25191/files/media/file/snap_landscape_analysis_of_assessments_for_ngss_1.pdf

[5]Saye, J. W., Stoddard, J., Gerwin, D. M., Libresco, A. S., & Maddox, L. E. (2018). Authentic pedagogy: Examining intellectual challenge in social studies classrooms. *Journal of Curriculum Studies*, *50*(6), 865–884. https://doi.org/10.1080/00220272.2018.1473496

[6] Tekkumru-Kisa, M., Stein, M. K., & Schunn, C. (2015). A framework for analyzing cognitive demand and content-practices integration: Task analysis guide in science. *Journal of Research in Science Teaching*, *52*(5), 659–685. https://doi.org/10.1002/tea.21208

[7] Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

[8] Evans, C. M. (2018). *Can schools be reformed by reforming assessment?: The effects of an innovative assessment and accountability system on student achievement outcomes*. University of New Hampshire.

[9] Foote, M. (2007). Keeping Accountability Systems Accountable. *Phi Delta Kappan*, *88*(5), 359–363. https://doi.org/10.1177/003172170708800506

[10] Noble, T., Suarez, C., Rosebery, A. S., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*(6), 778–803. https://doi.org/10.1002/tea.21026

[11] Noble, T., Sireci, S. G., Wells, C. S., Kachchaf, R. R., Rosebery, A. S., & Wang, Y. C. (2020). Targeted linguistic simplification of science test items for English learners. *American Educational Research Journal*, *57*(5), 2175–2209. https://doi.org/10.3102/0002831220905562

[12] Llosa, L. (2021). *Expanding the evidence of learning to promote equity through formative classroom assessment*. NCME Classroom Assessment Conference, Conference held remotely.

[13] Fine, C. G. McC., & Furtak, E. M. (2020). A framework for science classroom assessment task design for emergent bilingual learners. *Science Education*, *104*(3), 393–420. https://doi.org/10.1002/sce.21565

[14] Randall, J. (2021). "Color-neutral" Is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, emip.12429. https://doi.org/10.1111/emip.12429

[15] Randall, J., Poe, M., & Slomp, D. (2021). *Ain't* Oughta Be in the Dictionary: Getting to Justice by Dismantling Anti-Black Literacy Assessment Practices. *Journal of Adolescent & Adult Literacy*, *64*(5), 594–599. https://doi.org/10.1002/jaal.1142

[16] Koretz, D. (1992). *The Vermont Portfolio Assessment Program: Interim report on implementation and impact, 1991-92 school year. Project 3.2: Collaborative development of statewide systems. Report of year 1 Vermont study.* CRESST. https://files.eric.ed.gov/fulltext/ED351345.pdf

[17] Alonzo, A. C., & Ke, L. (2016). Taking stock: Existing resources for assessing a new vision of science learning. *Measurement: Interdisciplinary Research and Perspectives*, *14*(4), 119–152. https://doi.org/10.1080/15366367.2016.1251279

[18] Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, *37*(1), 21–34. https://doi.org/10.1111/emip.12189

[19] Davis, J., & Martin, D. B. (2008). Racism, assessment, and instructional practices: Implications for mathematics teachers of African American students. *Journal of Urban Mathematics Education, 1*(1), 10–34.

[20] Kang, H., & Furtak, E. M. (2021). Learning theory, classroom assessment, and equity. *Educational Measurement: Issues and Practice*, *40*(3), 73–82. https://doi.org/10.1111/emip.12423

[21] Tekkumru-Kisa, M., Stein, M. K., & Doyle, W. (2020). Theory and research on tasks revisited: Task as a context for students' thinking in the era of ambitious reforms in mathematics and science. *Educational Researcher*, *49*(8), 606–617. https://doi.org/10.3102/0013189X20932480

[22] Delain, M. T. (1995). Equity and performance-based assessment: An insider's view. *The Reading Teacher*, *48*(5), 440–442.
[23] Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, Edward H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*(3), 373–399. https://doi.org/10.1177/0013164497057003001
[24] Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education*, *7*(4), 323–342. https://doi.org/10.1207/s15324818ame0704_4
[25] Penuel, W. R., Turner, M. L., Jacobs, J. K., Van Horne, K., & Sumner, T. (2019). Developing tasks to assess phenomenon-based science learning: Challenges and lessons learned from building proximal transfer tasks. *Science Education*, *103*(6), 1367–1395. https://doi.org/10.1002/sce.21544
[26] Koretz, D., McCaffrey, D. F., Klein, S. P., Bell, R., & Stecher, B. M. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program*. CRESST.
[27] Hill, R., & DePascale, C. A. (2002). *Determining the reliability of school scores*.
[28] Marion, S. F., & Buckley, K. (2016). Design and implementation considerations of performance-based and authentic assessments for use in accountability systems. In H. Braun (Ed.), *Meeting the Challenges to Measurement in an Era of Accountability* (pp. 59–86). Routledge. https://doi.org/10.4324/9780203781302-10

Education Commonwealth Project
University of Massachusetts Lowell
Coburn Hall 222, 850 Broadway St.
Lowell, MA 01854
EdCommonwealth.org