Performance Assessments: Promises and Pitfalls

By learning from the past, state boards can add depth and relevance to their assessment systems.

Marianne Perie

Statewide performance assessments have gone in and out of vogue for over 30 years. While embraced as authentic representations of student work, they often prove burdensome for standardized assessment purposes. They can be a strong addition to a balanced assessment system—when implemented well for the right purposes. State policymakers looking to bring performance assessments to their schools need to think deeply about their purpose and intended use.

A performance assessment may comprise one or more tasks. A performance task is any activity that asks students to do something to demonstrate their knowledge, understanding, and proficiency. Performance tasks yield a tangible product and/or performance that serve as evidence of learning.

Performance assessments let students show what they know in an applied fashion. They are commonplace in the arts, where students are judged on

painting, sculpture, musical performance, or dance. Even in academic subjects, performance tasks may be more representative of work done in the classroom, be it running science experiments, writing essays, or recreating an event in history.

But because they are difficult to standardize and expensive to score, states have often abandoned efforts to administer performance assessments. This article will describe their benefits and challenges.

Learning from Past Attempts

The amount of subjectivity in assigning and scoring the tasks led to low levels of reliability and incentives to game the system.

Before Congress passed No Child Left Behind (NCLB) in 2001, three states were using performance assessments as part of their accountability systems: Kentucky, Vermont, and Maryland. Kentucky began requiring writing portfolios at grades 4, 8, and 12 in response to the 1990 Kentucky Education Reform Act. A portfolio consists of multiple samples of student work taken throughout the school year to demonstrate progress.

The policymakers' goal was to improve their students' writing and use the work to hold schools accountable for providing better instruction that increased the progress of all students. The initiative required professional development for all teachers, not just English teachers, to create and score writing tasks. However, the program ultimately failed due to the high-stakes nature of the consequences for teachers. Teachers were not incentivized to challenge students, leading to a wide degree of variability in the tasks assigned. Additionally, there were concerns with the reliability of the results, as teachers scored their own students.

Vermont implemented a similar program in 1991 for writing and mathematics. Schools were encouraged to use results for individual instruction and intervention, but aggregated scores were used at the state level for school accountability. They, too, struggled with standardizing tasks and implementing comparable processes for providing feedback and allowing revisions. As in Kentucky, the amount of subjectivity in assigning and scoring the tasks led to low levels of reliability and incentives to game the system.

The Maryland School Performance Assessment Program (MSPAP) was a performance-based assessment implemented in 1991 that covered reading, writing, language usage, mathematics, science, and social studies. It was implemented in grades 3, 5, and 8, and the assessment tasks were unique in that they involved group work. This format led to results that could be generalized at the classroom, school, and district levels but did not provide individual scores. This program appeared to have some early success, with large majorities of teachers and principals reporting that they saw improved instruction. Test prep activities became much more broadly focused on application of knowledge. However, the large focus on writing across all subjects interfered with the interpretation of scores in subjects like math and science.[1] Ultimately, the MSPAP had to be terminated with the passage of the NCLB, which required student-level scores in grades 3–8.

Current State Efforts

Currently, most performance assessments consist of writing tasks—either for ELA alone or combined with history/government/social studies—and computer simulation models for science, technology, engineering, and mathematics (STEM) subjects. New Hampshire has gained notoriety for being the first state to implement a performance assessment under the innovative assessment flexibility offered under the Every Student Succeeds Act (ESSA). The Performance Assessment of Competency Education (PACE) project, which began in 2014, includes locally developed, locally administered performance assessments tied to grade and course competencies determined by local school districts that are aligned with the state's challenging academic content standards. In each grade and subject, one common complex performance task, administered by all participating schools and districts, helps with calibration and provides evidence about the comparability of judgments related to student achievement across schools and districts. The state has continued their traditional statewide assessments in select grades and subjects to provide an ongoing audit of the innovative assessment system.[2]

Rhode Island has a performance assessment graduation requirement. However, it is intentionally subjective and not intended for statewide accountability. Since 2005, Rhode Island has required an individual learning plan to be designed at grade 6 and revisited throughout

high school. Students work with their teachers to determine whether they will develop a portfolio, exhibition, or capstone product in one or more subject areas selected by the student and approved by their teachers. The product is intended to demonstrate students' culminated applied learning skills and knowledge. Portfolios include performance-based entries required by the school district and some selected by the students. Exhibitions or capstone products tend to demonstrate deeper understanding of a selected topic. The state provided guidance on developing strong performance tasks, but districts ultimately set the graduation requirements.

Guide to Adaptation

All these state stories provide lessons for state boards of education interested in adopting performance assessments in their states or recommending them to districts. Assuming past is prologue, performance assessments require a level of local flexibility to survive in the long term.

State boards should ask questions about the purpose, content area, and criteria for proposed performance assessments. Bringing in resources to enrich student learning is very different from introducing a new assessment in order to evaluate teachers or schools. State leaders should also consider whether a performance assessment should be purchased, built from the ground up, or should leverage existing performance banks and standardized rubrics. Likewise, training educators to develop good performance tasks and strong rubrics can have effects beyond a single assessment. Finally, when introducing performance assessment to a balanced system of assessments, educators will need training on interpreting and using the results to inform teaching.

Purpose. The first question policymakers need to ask is what the purpose of the performance task is. If the intent is to use results for school and district accountability under ESSA, there are much stricter guidelines for reliability and comparability. If the goal is to incorporate performance tasks into instruction to help teachers better assess the depth of student understanding or to allow for more authenticity, then the requirements are much more flexible. A third vector is the use of performance tasks for student accountability by incorporating them into grades or graduation requirements, which requires high reliability, but the comparability requirement can be set locally.

Once the purpose is established, it will drive the remaining decisions. For instance, if the goal of adopting formative assessment is primarily for school use, then the focus should be on looking for banks of performance tasks and training on administering and scoring them. If the plan is to use them for accountability and the tasks need to be kept secure, then a vendor with experience in developing and scoring performance tasks would be a better choice. Again, New Hampshire has had success with a hybrid approach of common tasks plus locally developed tasks.

Subject. The second consideration is the subject or content area of interest. The area with the longest history of performance tasks is, of course, writing. Asking students to write an essay, either with or without a reading passage for context, is a task that states have been requiring for years. Within writing, there are several approaches. First, the rubric can be standardized and circulated for teachers to use throughout the school year. Then, only the writing prompt is a secure test question. This approach became more commonplace under NCLB, where the approach to writing was to have students respond to a passage they read on the test. The writing prompt could still be narrative, informative, or persuasive, with a rubric for each genre, but the score was based on organization, clarity, and ability to reference the passage sensibly.

Some groups are more interested in a student's ability to respond to feedback, so the score is based both on the final writing sample and on the growth from the first draft. Others are interested in a student's ability to research a topic and then write about it, so the task may be given over several weeks rather than as an on-demand prompt. This approach is often used in college courses. Again, the rubric is typically provided ahead of time.

One area that has garnered much research is whether students should be allowed to choose the prompt on which they wish to write. Because prior knowledge is an important component in the ability to write effectively, the theory is that if students choose a topic they are interested in, they will produce a better product. This approach should also lead to more

Performance assessments require a level of flexibility to survive in the long term.

equitable products, as students have varying exposure to different topics.

Research on Advanced Placement tests indicated that students who were given a choice of prompts tended to choose the one that looked easiest rather than the topic they were most drawn to. Instead, letting students choose the topic without first seeing the prompt should better match the theory and minimize the effects of unequal prior knowledge. More research is needed in this area, but allowing students some freedom in the topic would ultimately increase the fairness of the task.

Performance tasks in mathematics are most commonly developed as on-demand tasks in which students must perform several steps and show their work. Their score is based on process as well as product. A task with greater scaffolding to help struggling students often has several questions that build on one another to help guide the student to the final question. Other performance tasks involve a single scenario with multiple questions regarding a similar concept at varying degrees of difficulty. This latter approach is useful for classroom instruction to help teachers determine the depth of understanding within a single standard or concept.

With the release of the next-generation science standards, intense work began on developing computer-based simulations to test students process skills in addition to their content knowledge. Students may be given a scenario and allowed to try different approaches before answering a series of questions. For example, a physical science question may be about the relationship between a projectile and its velocity. Students may be given a simulation of a model rocket where they can vary the length of the tube, the shape of the nose cone, or the number of fins on the tail. After running the simulations, they then answer questions about the relationship between those characteristics and the distance it can travel before gravity pushes it into descent. The National Science Teachers Association gathered information on science simulations to provide teachers resources for using them in their classroom.[3]

When teacher evaluation became a part of the NCLB waiver requirements, much work was done in the area of performance tasks for teachers of subjects other than English and mathematics.[4] Teachers were either trained to create performance tasks or were given a bank to select from. Typically, two to three items per year were administered, and the teachers were evaluated on how much each student grew in their abilities throughout the year. Denver Public Schools created such banks for teachers of all subjects and provided professional development for teachers to create and field test their own tasks that could then be added to the bank.[5] Although the efforts in many states were largely abandoned, the exercise taught us that teachers could be trained to both develop and implement such tasks effectively. More work was needed on demonstrating comparability across classrooms and schools, as these efforts focused primarily on within-classroom comparability, which was not fully aligned with the purpose of teacher accountability.

Criteria. Multiple repositories exist for performance tasks and rubrics. Additionally, companies offer professional development to train teachers to develop and field test their own tasks.[6] State policymakers may consider first providing access to a bank of performance tasks and training teachers on implementing and scoring them. Grant Wiggins and Jay McTighe have written extensively on how to evaluate the quality of performance tasks. They focus on the importance of content and process knowledge as well as the student's ability to generalize that knowledge. Any context should be age appropriate, and the rubrics should be standardized. Their guidance for rubrics suggest evaluating students' products based on whether they meet the requirements of the task, the accuracy and thoroughness of the task, the evidence of following effective procedures, and the quality of the organization of the product (figure 1).

Other possible approaches involve hiring a vendor to create customized tasks and rubrics or train teachers to develop the tasks and rubrics, pretest them, revise them, and implement them effectively. As with most processes, evidence that the vendor has successfully completed such work previously is important.

Performance tasks can bring great depth to assessments and often result in tests that actually measure what teachers want to teach. However, they can be subjective and costly, making them difficult to implement at a large scale. Adding performance tasks to more traditional standardized assessments or building a performance assessment for specific grades or subjects may be the best approach to

Marianne Perie is the president of Measurement in Practice, LLC.

Figure 1. Four Types of Criteria with Sample Questions

Source: Grant Wiggins and Jay McTighe, Understanding by Design: Guide to Advanced Concepts in Creating and Reviewing Units (Alexandria, VA: ASCD, 2012).

creating a balanced assessment system that also includes professional development for educators to design tasks for use within instruction. Several states doing this work are combining some degree of standardization with room for customization. By doing so, they can meet the demands of comparability while allowing teachers and students to focus on areas of interest and produce their best work. ■

[1]Daniel Koretz et al., "Final Report: Perceived Effects of the Maryland School Performance Assessment Program," CSE Report 409 (Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing, 1996).

[2]New Hampshire Department of Education, "Performance Assessment of Competency Education," web page, https://www.education.nh.gov/who-we-are/division-of-learner-support/bureau-of-instructional-support/performance-assessment-for-competency-education.

[3]See, e.g., Argenta Price, Carl Wieman, and Katherine Perkins, "Teachers Use Simulations for Student Motivation, Content Learning, and Engagement in Science Practices," *Science Teacher* 86, no. 7 (March 2019): 46–52.

[4]See, e.g., Lisa Lachlan-Haché, Ellen Cushing, and Lauren Bivona, "Student Learning Objectives as Measures of Educator Effectiveness" (Washington, DC: American Institutes for Research, 2012).

[5]The Commons, "Student Learning Objectives (SLOs) for SSPs," web page http://thecommons.dpsk12.org/Page/417.

[6]Grant Wiggins and Jay McTighe, *Understanding by Design: Guide to Advanced Concepts in Creating and Reviewing Units* (Alexandria, VA: ASCD, 2012).