LEARNING POLICY INSTITUTE

CCSSO

Developing and Measuring Higher Order Skills: Models for State Performance Assessment Systems

RESEARCH BRIEF MAY 2017

Linda Darling-Hammond

Abstract

After passage of the Every Student Succeeds Act (ESSA) in 2015, states assumed greater responsibility for designing their own accountability and assessment systems. ESSA requires states to measure "higher order thinking skills and understanding" and encourages the use of open-ended performance assessments, which are essential for measuring these skills. The report from which this brief is drawn, which was commissioned by the Council of Chief State School Officers, reviews four models for large-scale assessment systems that include performance-based components and references research in the U.S. and abroad showing how states can design and score these assessments with high levels of comparability and reliability.

The full report is available online at https://learningpolicyinstitute.org/product/models-state-report.

External Reviewers

This report benefited from the insights and expertise of two external reviewers: Paul Leather, Deputy Commissioner at the New Hampshire Department of Education, and Gretchen Morgan, Fellow at the Center for Innovation in Education, University of Kentucky. We thank them for the care and attention they gave the report. Any shortcomings are our own.

The S. D. Bechtel, Jr. Foundation, the Hewlett Foundation, and the Sandler Foundation have provided operating support for the Learning Policy Institute's work in this area.

Introduction

The Every Student Succeeds Act (ESSA) of 2015 replaced No Child Left Behind and opened up new possibilities for defining and supporting student success in American public education. Under its provisions, states play a larger role in setting academic standards that will guide instruction for all students. ESSA also gives states more opportunities to innovate with respect to student testing. It requires states to assess proficiency in English language arts, mathematics, and science, but it allows them to build their own assessment and accountability systems.

Unlike its precursor, ESSA requires states to measure "higher order thinking skills and understanding." It also permits the use of multiple assessments, including "portfolios, projects, or extended-performance tasks." Under No Child Left Behind, tests focused on reading and mathematics, but there was little emphasis on applying those skills to complex, real-world situations. Higher order thinking skills—including the ability to find, evaluate, synthesize, and use information to solve problems—are increasingly necessary for academic and vocational success, yet studies show that they are inadequately represented among first-year college students and the current workforce.

To evaluate these higher order thinking skills, more open-ended performance assessments are needed, and state policymakers and educators have a range of options for including them. This report reviews four models for large-scale assessment systems that include performance options—and what the research reveals about each one:

1. Performance items or tasks as part of traditional "sit-down" tests
2. Curriculum-embedded tasks carried out in the classroom during the school year
3. Portfolios or collections of evidence that display a broad set of competencies
4. Comprehensive assessment systems that include traditional sit-down tests, curriculum-embedded tasks, and portfolios and exhibitions leading to a student defense

All four models have been used successfully by states, other nations, and networks of schools. The models are not mutually exclusive, and the report considers ways they might be blended.

Performance Assessments

Performance assessments require students to construct an answer, produce a product, or perform an activity rather than simply identify a predetermined answer. They include, for example, science experiments that students design, perform, analyze, and write up; computer programs that students create and test; and written or oral presentations about a research topic. Because these assessments typically require students to integrate knowledge, analysis, and action, they are better than multiple choice tests at measuring higher order thinking skills. They are also better predictors of academic and vocational success.

1. Performance Items or Tasks Within Tests

Basic performance assessments can be conducted within traditional sit-down tests and scored by teachers or other trained raters; in some cases, computers can also be used to assess student performance.

Some performance tasks require students to draw on multiple sources of textual, graphic, and quantitative evidence to evaluate a real-world situation, come to a conclusion, and explain their solution or rationale for a course of action. Many countries in Europe, Asia, Africa, and the Caribbean use essays, open-ended problems, oral examinations, and inquiry tasks almost exclusively in their examinations. Some states—such as Kentucky, New York, Massachusetts, and other New England states that jointly created the New England Common Assessment Program (NECAP) tests—have long included constructed response items along with open-ended essays and problem solutions. These items and tasks typically account for a substantial part of the overall score. On Kentucky's Core Content Tests, for example, open-ended items and tasks accounted for 50%.

Several new tests—such as those from the Smarter Balanced Assessment Consortium (SBAC), the Partnership for Assessment of Readiness for College and Careers (PARCC), and the College and Work Readiness Assessment (CWRA)—also include open-ended items and performance tasks that require students to engage in more complex research, problem solving, and analysis. Examples of these performance tasks include:

Essays used to evaluate writing, either as part of an English language arts test or as a stand-alone writing assessment, responding to a question or interpreting literature. In New York State, for example, students are asked to write an essay about a controlling idea in two literary texts as well as the authors' use of literary elements and techniques.

Document-based questions (DBQ) used to examine students' knowledge, reasoning, and use of evidence in a content area. Both the Advanced Placement history tests and the New York State Regents history tests provide multiple documents that must be evaluated in answering a complex question. The College and Work Readiness Assessments provide an in-basket of

documents that require students to undertake qualitative and quantitative analysis to evaluate a problem and propose a solution or course of action.

Mathematical or scientific problem solutions that require calculating and explaining the reasoning that leads to a solution—and how that solution would differ with changes in the conditions or variables concerned.

Computer-based simulations in which students pursue interactive inquiries to solve questions or problems. The National Assessment of Educational Progress (NAEP), for example, tests students' abilities to design experiments, display and interpret results, and search the internet effectively.

Research tasks that engage students in investigating questions and evaluating evidence to reach a conclusion or explanation. In the Smarter Balanced English language arts assessments, for example, students conduct online research on a question, weigh and balance evidence, and come to a well-defended conclusion.

2. Curriculum-Embedded Performance Assessments

Other assessments evaluate students on tasks that are embedded within units in the curriculum and may extend over days or weeks. These tasks might include, for example, researching and designing a software solution to meet a specific need, testing that solution with users, and offering improvements. To demonstrate speaking and listening skills for the General Certificate of Secondary Education (GSCE), students might be asked to perform a drama-focused activity, a group activity, and an individual extended contribution. To demonstrate reading comprehension, they might be required to show an understanding of three texts in their social, cultural, and historical context.

Curriculum-embedded performance assessments can be standardized in their design and still permit student choice—for example, on the topic to be addressed or product to be designed. The tasks are usually scored using common rubrics. Many countries—and the International Baccalaureate program, which operates in 125 countries—rely on this assessment method. They often combine papers or projects in the classroom (which are completed to certain specifications) with an end-of-year test to produce a summative score. The tasks, which are scored by trained teachers, typically account for 30–60% of the total scores.

Curriculum-embedded assessments offer several advantages over other approaches. The tasks can be performed over longer periods, thereby allowing students to undertake more challenging work and to demonstrate a broader range of skills. Also, students and teachers do not experience these tasks as formal tests. Although they are more carefully constructed and scored, and teachers have guidance about how to support the work, the tasks resemble normal school assignments. Finally, these assessments create greater curriculum equity by ensuring that all students, and not only those with proactive teachers, have the opportunity to investigate, analyze, write about, and revise their work.

States can add one or more curriculum-embedded tasks to assessments in any subject area. States that use curriculum-embedded performance tasks often create a statewide bank of tasks that can be shared across classrooms. Many draw on the nationally available Performance Assessment Resource Bank (PARB) for high-quality tasks with rubrics and instructional guidance that have been vetted and field tested. PARB also includes protocols for developing tasks and scoring them consistently.

3. Portfolios or Collections of Evidence

Portfolios, which collect evidence of student learning over time, are typically organized around a set of standards or competencies students must demonstrate they have met. They are often collections of performance tasks, although evidence from traditional sit-down tests or internships is sometimes included. Often students must present and defend their work before a jury of teachers and outside judges.

The portfolio approach is meant to develop self-directed learners who can evaluate and improve their own work. Although the parameters for tasks are specified, students often choose their own topics and revise their submissions to meet the relevant standards. They can see their own progress over time and reflect on how they have improved. They also receive specific and detailed feedback to guide that improvement. When students receive such feedback from different sources, they can identify patterns of strength and weakness that go well beyond correct or incorrect answers to specific questions. When they defend their work, they must show that they deeply understand the concepts and issues associated with the areas they have studied. They also internalize rigorous standards and develop the ability to plan, persevere, use feedback productively, and communicate effectively.

Kentucky and Vermont have used single-subject portfolio systems for writing and mathematics, with positive effects on instruction. Studies of these reforms found that teachers changed their classroom practices to support problem solving and communication, and both states experienced student achievement gains on the NAEP. Portfolios are also used in the Advanced Placement (AP) program for Art, Technology, and the new AP Research and AP Seminar courses that together compose the Capstone program for which students complete a digital portfolio of work, scored partly by their own teachers and partly by other AP teachers, all of whom are trained for reliable scoring.

Portfolios covering multiple disciplines are often used at the high school level. Rhode Island has long used portfolios for graduation purposes, Oregon permits them as an option for graduation, and New Hampshire will implement graduation capstone projects next year. Some districts and many school networks require portfolios for graduation, and schools participating in the New York Performance Standards Consortium may use them instead of most Regents Examinations. The National Academies Foundation also has developed a portfolio model that is scored with common standards across hundreds of schools.

4. Comprehensive Assessment Systems

Comprehensive assessment systems strategically combine several of these models to provide reliable information about student learning, often with less traditional testing. They typically include classroom-embedded performance assessments and standardized statewide measures to validate local results. They may also include portfolios in some areas.

New Hampshire's Performance Assessment of Competency Education (PACE) system is one such comprehensive model. The PACE system uses curriculum-embedded assessments across all subject areas and grades. Traditional exams are used less frequently to validate the results of the performance tasks. For federal purposes, PACE supplements a standardized test in English language arts and mathematics at one grade level within each grade span with common performance tasks in the other years. The system will soon include graduation capstone projects with exhibitions and defenses before juries of educators and peers.

During the 1990s, Connecticut, Kentucky, Maine, and Vermont also employed comprehensive assessment systems. They combined periodic tests, which included performance items, with curriculum-embedded performance tasks and, sometimes, portfolios of tasks. Studies indicate that this mix of assessments encouraged instructional strategies that fostered reasoning, problem solving, and communication, as well as a focus on research and writing.

Comparability, Task Design, and Scoring

Many questions about performance assessments concern comparability and reliability across tasks, settings, and scorers. Research over many years has demonstrated how comparability can best be achieved. In successful systems, tasks and rubrics are guided by learning standards and focus on clearly specified knowledge and skills. They may be designed within common templates to specifications that help create comparable tasks. They are also reviewed carefully and field-tested.

To produce consistent and reliable scoring, teachers are trained in settings where they review and discuss model answers and their own scores until their judgments are consistent. They may use benchmark examples of student work at different levels, along with a rubric or set of scoring criteria, to calibrate their judgments. As they learn to look for the key features of the work expressed in the criteria, teachers become more aware of the elements of strong student performance. Scores may be audited and the results used to retrain scorers and to calibrate scores for consistency, with improved reliability as a result.

New Hampshire includes an expert review of tasks and rubrics along with training for scorers. It also conducts comparability analyses that measure agreement within and across districts. Finally, it compares performance assessments to standardized tests to evaluate comparability. These analyses have found strong and increasing agreement among raters and acceptable levels of comparability across assessments.

Similar strategies have been used in Kentucky, New York, Vermont, England, Singapore, and Queensland, Australia. Some developers of performance assessments have achieved inter-rater reliabilities of 90% or more, matching the level achieved in the Advanced Placement system.

Taken together, studies point to five factors for comparability across tasks and consistent and reliable scoring:

- Designing tasks with a clear idea of what is being measured and what constitutes acceptable performance;
- Developing clear and specific scoring guides;
- Selecting qualified raters;
- Providing sufficient training; and
- Monitoring the scoring process through moderation and auditing.

Using Technology and Teachers for Scoring

Computer-based scoring has been used successfully in certain contexts. Essays are often scored by computers with high levels of reliability. In one NAEP study that used physics simulations, the agreement between human raters and their computer counterparts was 96%. In a more complex assessment—designed by the Collegiate Learning Assessment to elicit student reasoning, use of quantitative and qualitative evidence, and writing—correlations between human and computer ratings were nearly as high, at 86%.

More frequently, technology is used to support the human scoring process. In the International Baccalaureate program, teachers receive papers via computer and calibrate their scoring to common benchmarks through an online training program. The teachers also upload their scored papers to be further evaluated or audited, as needed, and to record scores. Similarly, in Hong Kong, most delivery and scoring of open-ended assessments is becoming computer-based, as it is in 20 other Chinese provinces. There, as in many other places, double scoring is used to ensure reliability, and a third scorer may be called in to resolve discrepancies. In the United States, educators score portfolios for teacher licensure online and receive training and calibration through a computer-based program.

Although technology is a powerful assessment tool, human scoring is important for several reasons. Studies show that involving teachers in scoring improves classroom instruction by helping teachers link it more firmly to state standards. Where school systems invest in teacher scoring of classroom-based performance assessments, teachers develop shared expertise about what high-quality instruction, assessment, and student work look like, and how they can better support such work.

These benefits were illustrated in the Building Educator Assessment Literacy (BEAL) project that allowed teachers from California, New Hampshire, and Oregon to score and discuss the performance tasks from the Smarter Balanced tests. Across all three states, 97% of respondents said that the training "deepened my understanding of the assessment system"; 96% said it

"helped me think about ways to enact curriculum-embedded performance assessment with my students"; and 88% said that the scoring process "deepened my understanding of the State Standards." This professional development training continues to be offered by WestEd and the Stanford Center for Assessment, Learning, and Equity (SCALE). Meanwhile, individual states offer similar training for scoring their state-developed assessments.

Conclusion

Performance assessments are an important way to evaluate higher order thinking skills as required by ESSA. States have a number of options for implementing such assessments in their public school systems. Each of the models discussed here has been developed, studied, and refined both in the United States and abroad.

When designed well, these assessments create and evaluate worthwhile tasks that link classroom instruction to state standards and encourage stronger teaching of ambitious skills. When performance assessments are conducted in classrooms, students deepen their understanding of content. At the same time, they develop a range of cognitive and co-cognitive skills as they work intensively on their tasks, revise their work to meet standards, and display their learning to parents, peers, teachers, and employers. Teachers' engagement in using and scoring performance assessments has been found to improve instruction and student learning.

At the same time, systems can be designed so that policymakers, parents, and educators can track progress and trends as these scores are reported, aggregated, and analyzed, thus providing an engine for ongoing improvement.