

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



THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Melody Schopp (South Dakota), President

Chris Minnich, Executive Director

CONTENTS

Executive Summary.....	1
Introduction.....	2
What is Performance Assessment? Why is it Important?	4
A Continuum of Assessment Options.....	6
Models of Performance Assessment	7
I. Tests that include Performance Items or Tasks	8
Essays and Inquiry Tasks	9
Computer-Based Simulation Tasks	14
II. Curriculum-Embedded Performance Assessments.....	16
Curriculum-Embedded Performance Tasks	16
Performance Assessment Task Banks	18
Science Assessments	20
Assessments in Social Studies, the Arts, and other Areas.....	25
Culminating Projects and Exhibitions	26
III. Portfolios / Collections of Evidence	28
Single-Subject Portfolios.....	29
Multiple Subject Portfolios.....	33
IV. Comprehensive Assessment Systems.....	38
Comparability, Task Design, and Scoring	44
Task Design	44
Scoring	46
Uses of Technology in Scoring	48
Teacher Involvement in Scoring	49
Conclusion.....	52
Appendix A: New York Performance Standards Consortium Science Rubric	53
Endnotes	54

Acknowledgements: The author would like to gratefully acknowledge the helpful insights of 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.

EXECUTIVE SUMMARY

The Every Student Succeeds Act (ESSA) opened up new possibilities for how student and school success are defined and supported in American public education. States have greater responsibility for designing and building their assessment and accountability systems. The law also broadens the concept of student learning, requiring that assessments measure “higher-order thinking skills and understanding.” It explicitly allows the use of multiple assessments including “portfolios, projects, or extended-performance tasks” as part of state systems. States are also invited to apply for an innovative assessment pilot to develop new approaches to assessment and gradually scale them up statewide.

These new opportunities to develop performance assessments are critically important to provide incentives for teaching the more complex skills students increasingly need to succeed in the rapidly evolving U.S. society and economy. The modern workplace requires students to demonstrate well-developed thinking skills, problem solving abilities, design strategies, and communication capabilities that cannot be assessed by most currently used tests.

This paper discusses four models for integrating performance-based components into assessment systems, all of which have been used successfully at scale in states and

nations around the world. It also discusses what is needed to assure validity, reliability, and comparability in the use of such assessments. These models --which can also be combined in various ways -- include:

- I. **Performance items** or tasks as part of traditional ‘sit-down’ tests.
- II. **Curriculum-embedded tasks** that are implemented in the classroom during the school year, assessing more complex sets of skills. These may be common or locally developed and may stand alone or be combined with test results to produce a summative score.
- III. **Portfolios or collections of evidence** that aggregate multiple tasks to display a broad set of competencies in multiple domains or genres.
- IV. **Comprehensive assessment systems** that include traditional sit-down tests, curriculum-embedded tasks, and portfolios and exhibitions leading to a student defense, each serving distinctive complementary purposes.

In each case, the paper describes what states and some nations have done and are doing to develop and implement sound assessments in terms of design, implementation, and scoring. It also outlines what research has found in terms of productive practices in developing performance assessment practices that produce strong outcomes for teaching and learning.

INTRODUCTION

In December 2015, passage of the Every Student Succeeds Act (ESSA) opened up new possibilities for how student and school success are defined and supported in American public education. One of the most notable shifts in the law is that states have greater responsibility for designing and building their state assessment and accountability systems. The concept of student learning is also much broader than it was under NCLB.

States are expected to adopt challenging academic standards that will serve to guide curriculum and instruction for all students. Furthermore, states must implement assessments that measure “higher-order thinking skills and understanding.” Because traditional multiple-choice tests are insufficient for these goals, the law explicitly allows the use of “portfolios, projects, or extended-performance tasks” as part of state systems.¹

To measure academic achievement in mathematics, reading/language arts, and science, states may use a single summative assessment or “multiple statewide interim assessments during the course of the academic year that result in a single summative score that provides valid, reliable, and transparent information on student achievement or growth.”² This strategy might allow schools to better integrate assessment into curriculum and teaching and provide timely information to inform instruction.

States are also invited to apply for an innovative assessment pilot³ that will allow up to seven states initially to develop and pilot new approaches to assessment, refine the assessments, and gradually scale them up across the state.

These new opportunities are critically important because current tests in the U.S. are focused almost exclusively on low-level skills of recall and recognition.⁴ Consequently, they do not provide incentives for teaching the more complex skills students increasingly need to succeed in the rapidly evolving U.S. society and economy. The modern workplace increasingly requires students to demonstrate well-developed thinking skills, problem solving abilities, design strategies, and communication capabilities.

To succeed, people need to be able to find, evaluate, synthesize, and use knowledge in new contexts, frame and solve non-routine problems, and produce research findings and solutions – skills employers find inadequately represented in the current workforce.⁵ Additionally, college faculty have identified critical thinking and problem solving as areas in which first-year college students are lacking when they enroll.⁶

As important as these skills are, the educational policy system and the larger political system are not functioning effectively to foster their development and implementation in U.S. schools. More than a decade of test-based accountability targeted narrowly on reading and mathematics focused schools on the importance of these subjects, but ignored the application of these skills to complex, real-world situations. New systems of curriculum, assessment, and accountability will be needed to ensure that students are given the opportunities to learn what they need to be truly ready to succeed in college and careers.

Given these expectations, states are examining how they can create systems that include more robust assessments that encourage and measure higher-order thinking and performance skills. Many states created systems in the 1990s that included performance tasks and portfolios, and learned to manage these so that they produced reliable results at scale. Most of these were abandoned during the No Child Left Behind Act of 2001 (NCLB) era, but some survived, and a number of states are re-establishing performance-oriented systems today. Many countries also routinely use performance tasks to measure higher-order thinking skills as part of their examination systems.

In this paper, I discuss four models for integrating performance-based components into assessment systems, all of which have been used successfully at scale in states and nations around the world. I also discuss what is needed to assure validity, reliability, and comparability in the use of such assessments. The models below can be combined in various ways:

- I. **Performance items** or tasks as part of traditional 'sit-down' tests.
- II. **Curriculum-embedded tasks** that are implemented in the classroom during the school year, assessing more complex sets of skills. These may be common or locally developed and may stand alone or be combined with test results to produce a summative score.
- III. **Portfolios or collections of evidence** that aggregate multiple tasks to display a broad set of competencies in multiple domains or genres.
- IV. **A comprehensive assessment system** that includes traditional sit-down tests, curriculum-embedded tasks, and a portfolio leading to a student defense, each serving distinctive complementary purposes.

Before I describe these models at length, I discuss what we mean by performance assessment and why it is essential for measuring higher-order skills and abilities to apply knowledge.

WHAT IS PERFORMANCE ASSESSMENT? WHY IS IT IMPORTANT?

For many people, performance assessment is most easily defined by what it is *not* — specifically, it is not multiple-choice testing. In a performance assessment, rather than choosing among pre-determined options, students must construct an answer, produce a product, or perform an activity.⁷ From this perspective, performance assessment encompasses a very wide range of activities from writing a few sentences (short response), to developing a thorough analysis (essay), to conducting and analyzing a laboratory investigation (hands-on).

The goal of performance assessment is to more closely reflect the genuine performance of interest to “emulate the context or conditions in which the intended knowledge or skills are actually applied,”⁸ so that they are better predictors of what students can do in the real world. Because such assessments allow students to construct or perform an original response rather than just recognize a potentially right answer out of a list provided, performance assessments can measure students’ cognitive thinking and reasoning skills and their ability to apply knowledge to solve realistic, meaningful problems.

Almost every adult in the United States has experienced at least one performance assessment — the driving test that places new drivers into an automobile with a DMV official for a spin around the block and a demonstration of a set of driving maneuvers, including, in some parts of the country, the dreaded parallel parking technique. Few of us would be comfortable handing out licenses to people who have only passed the multiple-choice written test also required by the DMV. We understand the value of this performance assessment as a real-world test of whether a person can actually handle a car on the road. Not only does the test tell us some important things about potential drivers’ skills, we also know that preparing for the test helps improve those skills as potential drivers practice to get better. (What parent doesn’t remember the hair-raising outings with a 16-year-old wanting to practice taking the car out over and over again?) The test sets a standard toward which everyone must work. Without it, we’d have little assurance about what people can actually *do* with what they know about cars and road rules, and little leverage to improve actual driving abilities.

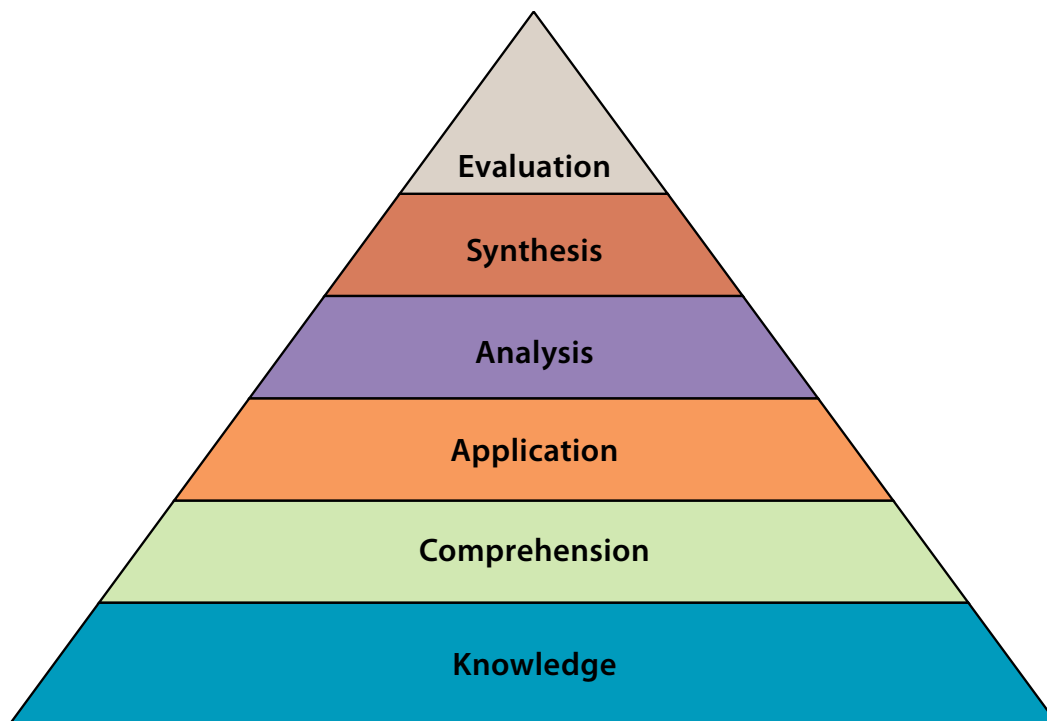
What makes the driver’s performance assessment valid is that it directly exhibits the actual skills needed, as they are used in the real world. The assessment does not need to be secret in order to be a useful test, since the driver must work to acquire and display the necessary skills in order to pass. Rather than relying on secrecy around what facts must be memorized, a robust performance assessment evaluates the way knowledge and skills are mastered, combined, and used in practice.

Performance assessments in education are very similar. They gather information about what students can actually do with what they are learning — science experiments that students design, carry out, analyze, and write up; computer programs that students create and test out; research inquiries that they pursue; evidence they have assembled about a question that they present in written and oral form. Whether the skill or standard being measured is writing, speaking, scientific, or mathematical literacy, or knowledge of history and social science research, students perform tasks in which they directly apply the relevant knowledge and skills. As with the driver’s test, even if the task is known, the student must work to acquire and display the necessary skills in order to pass.

Performance assessments are essential to measuring higher order skills — those shown at the top of Bloom’s taxonomy:⁹ applications of knowledge, analysis, synthesis, and evaluation. (See Figure 1.) These assessments can take different forms, including questions that can be answered by what are called “constructed-response” items — those that require students to create a response — within a relatively short time in a traditional “on-demand” test that students sit down to take. They can also include more extended tasks that require time in class. These classroom-based performance tasks allow students to engage in more challenging activities that demonstrate a broader array of skills, including problem framing and planning, inquiry, and production of more extended written or oral responses.

Figure 1:

Bloom’s Taxonomy of the Cognitive Domain



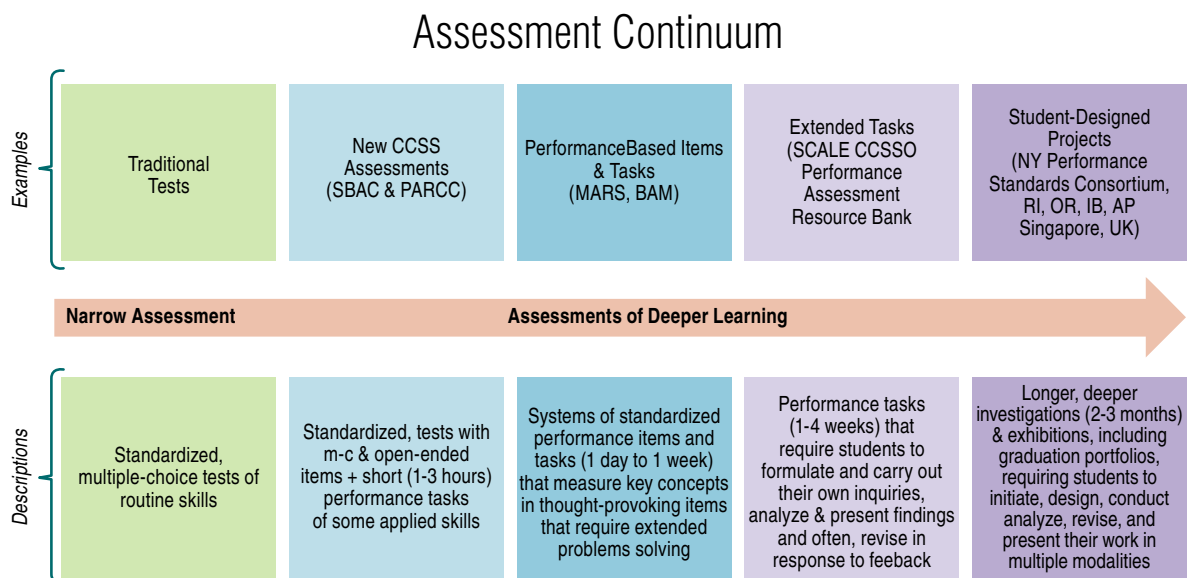
A CONTINUUM OF ASSESSMENT OPTIONS

Performance tasks may be highly standardized in their content or they may offer students some choices, for example, in the topic they research or write about, the way they conduct an inquiry, or in the way they display their results. In any event, the tasks are scored based upon a set of pre-determined criteria, usually codified in a rubric. Scoring may be conducted by the student’s classroom teacher if the purpose is to inform classroom instruction, or by another rater (usually another trained teacher) or even a jury of assessors, if the purpose is for comparable reporting or accountability. When comparability is needed, scorers are trained to rate the work consistently, often in a “moderated” process that assures reliability, and sometimes with an external audit of scores.

Assessment strategies can be thought of as existing along a continuum.¹⁰ At one end are the multiple-choice and close-ended items found in today’s traditional tests. These items measure recall and recognition, but cannot measure higher level thinking skills or the ability to apply them. When the RAND Corporation evaluated the depth of knowledge represented in state tests under NCLB, for example, they found that only 2 percent of math items and only about 20 percent of English language arts items represented higher-order thinking skills, and that the limitations imposed by multiple choice testing were a major reason for this ‘dumbing down’ of test content.¹¹

At the other end are assessments that require substantial student initiation of designs, ideas, and performances, tapping the planning and work management skills especially needed for college and careers. As shown in Figure 2, in between, at each step along the continuum, tasks become more complex, measuring progressively larger and more integrated sets of knowledge and skill, more cognitively complex aspects of learning, and more robust applications of knowledge to new problems and situations.

Figure 2



Along this continuum, the role of the student also changes from passively receiving and responding to external questions at one end of the continuum, to taking increasing initiative for finding and making sense of information, as well as determining questions, methods, and strategies for investigation at the other end. At the right hand end of the continuum, where students are conducting substantial research, presenting and defending their work, and revising it in response to feedback, they are also developing and demonstrating a range of communication skills, meta-cognitive and “learning-to-learn” skills, resilience that accompanies a growth mindset with regard to academic pursuits, and – in some cases – skills of collaboration, as well.

These deeper learning skills are demonstrated in the context of robust performance tasks, portfolios, and exhibitions of work that more authentically represent how work is developed and evaluated outside of school. Interestingly, a growing number of countries include these kinds of assessments in their examination systems as they seek to move their systems toward 21st century skills.

Rather than trying to have one test address all needs, different methods can be combined in a *system of assessments* that strategically uses different types of information for different purposes, as our fourth model illustrates. Performance assessments can be designed to provide formative and/or summative information, to gauge student growth on learning progressions, to support proficiency determinations, or to be combined in a student profile or portfolio.

MODELS OF PERFORMANCE ASSESSMENT

Along a continuum of assessment options, schools, districts, and states can encourage and evaluate the development of a range of knowledge, skills, and dispositions – collecting evidence for a range of different purposes and supporting instruction that is focused both on deep understanding of content and its use in complex applications. States can mix and match these approaches as they develop their overall assessment models, depending on their theory of action and the kind of educational improvements they are seeking to support.

Under ESSA, states must assess students annually to make a determination about each student’s degree of proficiency in ELA and math in grades 3-8 and once in high school, and at least once in each grade span in science.¹² They can do this with a single test or with a set of assessments that also includes classroom-based projects or performance tasks. They may also combine multiple student pieces of student work into portfolios that are scored. Considerable work has been done over the last 25 years to develop and implement systems that allow for comparability in tasks and scoring, as well as feasibility in implementation.

This report is meant to inform state agency leaders, other state and district policymakers, and educators about the options that are available, where and how they have been used, and the considerations decision makers and users should keep in mind as they evaluate what is most appropriate for their own contexts. It reviews possibilities and their potential utility for various purposes within each of the three categories of assessment models: 1) tests that include performance items or tasks; 2) curriculum-embedded performance tasks; and 3) portfolios. The report then discusses how task design and scoring can be structured to support both comparability and teacher learning.

I. TESTS THAT INCLUDE PERFORMANCE ITEMS OR TASKS

The most basic form of performance tasks may require a student to write an essay that analyzes a piece of text or other evidence; solve a multi-part problem and explain his or her solution; or conduct a brief inquiry and analyze the resulting data to answer a question or solve a problem. These tasks assess knowledge and skills that cannot be gauged well with multiple-choice items. They are used in traditional testing contexts, where students are taking a sit-down test in which they respond to specific prompts in a standardized fashion.

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, Massachusetts, other New England states who jointly created the New England Common Assessment Program (NECAP) tests, and New York have long included constructed response items, along with open-ended essays and problem solutions in their tests, accounting for a substantial part of the score. (On Kentucky's Core Content Tests (KCCT), for example, open-ended items and tasks accounted for 50 percent of the total score.)

New tests that evaluate more challenging standards, such as the Smarter Balanced and PARCC assessments and the College and Work Ready Assessment (CWRA) include open-ended items and performance tasks that require students to engage in more complex research, problem solving, and analysis. Tests like the National Assessment of Educational Progress (NAEP) science test include computer-assisted simulations that evaluate inquiry, and new science assessments under development may adopt these strategies.

In the context of large-scale assessment systems, examples of these kinds of 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.

- **Document-based questions** (DBQ) used to examine students' knowledge, reasoning, and use of evidence in a content area – as in the essays that are part of the Advanced Placement history tests or the New York State Regents history tests, which provide multiple documents that must be evaluated in answering a complex question.
- **Problem solutions** that require showing the work and explaining the reasoning that leads to a solution – for example to a mathematics or physics problem.
- **Computer-based simulations** in which students pursue interactive inquiries to solve questions or problems.
- **Research tasks** that engage students in investigating questions and evaluating evidence to reach a conclusion or explanation.

Essays and Inquiry Tasks

States can choose to develop or select assessments that incorporate performance tasks to better measure higher order thinking skills and to encourage teachers to attend to these skills in their teaching. The rationale for such tasks is based on what the learning sciences reveal about transferable knowledge — that true understanding is best developed and revealed by students' abilities to apply what they know in the context of new questions or situations where they must apply, analyze, evaluate, and communicate their ideas. Furthermore, assessing knowledge in ways that require these cognitive moves is more likely to encourage the teaching that develops such skills.

New York Regents Tests. Since 1865, for example, New York State has had a history of state-level assessment that includes performance-based testing. The Regents examinations, emulating the British tradition, began as open-ended essays and tasks. The Regents Science Examination still includes expectations for laboratory performance tasks, along with a written test with a number of open-ended questions. In English, students write responses to both spoken and written texts. In addition, they are asked to write an essay discussing a controlling idea within two literary texts and the authors' use of literary elements and techniques, and, in a separate essay, "to interpret a statement provided to them about some aspect of literature and write an essay using two works they have read to support their interpretation of the statement."¹³

In history and social studies, students complete essays that are document-based questions requiring analysis of a set of documents and artifacts to weigh and balance the answers to a question. Teachers are trained to score all extended writing tasks using benchmark performances and rubrics.¹⁴ They do

so on professional development days set aside at the end of the school year. A certain proportion of tests are annually audited by the state education agency to assure consistent standards.

New York Regents U.S. History Document-Based Question

After the Civil War, the United States became a much more industrialized society. Between 1865 and 1920, industrialization improved American life in many ways. However, industrialization also created problems for American society.

Using information from at least four of the documents provided and your knowledge of United States history, write an essay in which you discuss the advantages and disadvantages of industrialization to American society between 1865 and 1920. In your essay, include a discussion of how industrialization affected different groups in American society.

The Partnership for Assessing Readiness for College and Careers and Smarter Balanced Assessment Consortium Tests. The Partnership for Assessing Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) assessments, launched in 2014-15, were designed to measure higher order skills more fully, and analyses of the tests have found they do so.¹⁵ The increased use of constructed response items and performance tasks provides opportunities for students to analyze information; collect, evaluate, and use evidence to solve problems; and to communicate their results and reasoning. The sample tasks released by the two consortia include performance tasks that encourage instruction aimed at helping students acquire and use knowledge in more complex ways. (See Figures 3 and 4 below.)

Figure 3

Mathematics Performance Tasks

SBAC 6th Grade Task: Planning a Field Trip

Classroom Activity: The teacher introduces the topic and activates students' prior knowledge of planning field trips by:

- Leading students in a whole class discussion about where they have previously been on field trips or other outings, with their school, youth group, or family.
- Creating a chart showing the class's preferences by having students' first list and then vote on the places they would most like to go on a field trip, followed by whole class discussion on the top choices.

Student Task: Individual students:

- Recommend where their class should go on a field trip, based on their analysis of the class vote.
- Determine the per-student cost of going on a field trip to three different locations, based on a chart showing the distance and entrance fees for each option, plus formula for bus charges.
- Use information from the cost chart to evaluate a hypothetical student's recommendation about going to the zoo.
- Write a note to their teacher recommending and justifying which field trip the class should take, based on an analysis of all available information.

PARCC High School Task: Golf Balls in Water

Part A: Students analyze data from an experiment involving the effect on the water level of adding golf balls to a glass of water in which they:

- Explore approximately linear relationships by identifying the average rate of change.
- Use a symbolic representation to model the relationship.

Part B: Students suggest modifications to the experiment to increase the rate of change.

Part C: Students interpret linear functions using both parameters by examining how results change when a glass with a smaller radius is used by:

- Explaining how the y-intercepts of two graphs will be different.
- Explaining how the rate of change differs between two experiments.
- Using a table, equation, or other representation to justify how many golf balls should be used.

Source: Herman & Linn (2013).¹⁶

English Language Arts Performance Tasks:

PARCC 7th Grade Task: Evaluating Amelia Earhart's Life

Summary Essay: Using textual evidence from the Biography of Amelia Earhart, students write an essay to summarize and explain the challenges Amelia Earhart faced throughout her life.

Reading/Pre-Writing: After reading *Earhart's Final Resting Place Believed Found*, students:

- Use textual evidence to determine which of three given claims about Earhart and her navigator, Noonan, is the most relevant to the reading.
- Select two facts from the text to support the claim selected.

Analytical Essay: Students:

- Read a third text called *Amelia Earhart's Life and Disappearance*.
- Analyze the evidence presented in all three texts concerning Amelia Earhart's bravery.
- Write an essay, using textual evidence, analyzing the strength of the arguments presented about Amelia Earhart's bravery in at least two of the texts.

SBAC 11th Grade Task: Nuclear Power - Friend or Foe?

Classroom Activity: Using stimuli such as a chart and photos, the teacher prepares students for Part 1 of the assessment by leading students in a discussion of the use of nuclear power. Through discussion:

- Students share prior knowledge about nuclear power.
- Students discuss the use and controversies involving nuclear power.

Part 1: Students complete reading and pre-writing activities in which they:

- Read and take notes on a series of Internet sources about the pros and cons of nuclear power.
- Respond to two constructed-response questions that ask students to analyze and evaluate the credibility of the arguments in favor and in opposition to nuclear power.

Part 2: Students individually compose a full-length, argumentative report for their congressperson in which they use textual evidence to justify the position they take pro or con on whether a nuclear power plant should be built in their state.

Source: Herman & Linn (2013).

These tasks are scored by teachers or other trained raters. As described in the later section on scoring, some states like California, New Hampshire, and New York have required that practicing teachers must be the primary scorers of the performance tasks in statewide assessments. Evidence shows that this

involvement strengthens teachers' understanding of the standards and the assessments and informs classroom instruction.¹⁷

Collegiate Learning Assessments. The tasks young people encounter in college and in modern careers increasingly require them to analyze and synthesize diverse kinds of information, weighing and balancing evidence to solve complex problems. The Council for Aid to Education has developed assessments for high school and college students that represent this kind of learning. The Collegiate Learning Assessment (CLA) used at the college level, and the College and Work Ready Assessment (CWRA, used at the high school level, both use an in-basket approach. Students draw on multiple sources of textual, graphic, and quantitative evidence to evaluate a real-world situation, come to a conclusion, and explain their solution to a problem or their rationale for a course of action.

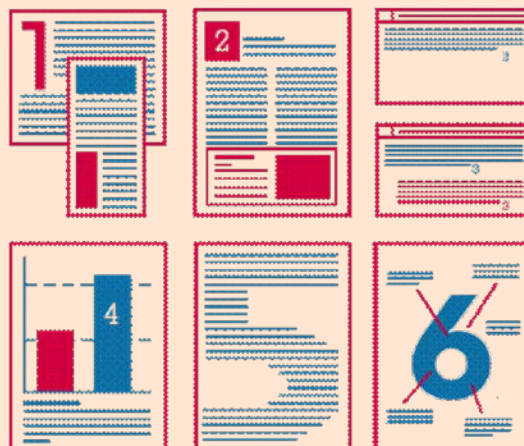
Research shows a strong relationship between performance on these assessments and success in college.¹⁸ While measuring complex skills, the responses can be scored reliably by computer, as well as by human scorers.

Figure 5:

Collegiate Learning Assessment Sample Performance Task

You are the assistant to Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigational equipment. Sally Evans, a member of DynaTech's sales force, recommended that DynaTech buy a small private plane (a SwiftAir 235) that she and other members of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. You are provided with the following documentation:

- 1: Newspaper articles about the accident
- 2: Federal Accident Report on in-flight breakups in single engine planes
- 3: Pat's e-mail to you & Sally's e-mail to Pat
- 4: Charts on SwiftAir's performance characteristics
- 5: Amateur Pilot article comparing SwiftAir 235 to similar planes
- 6: Pictures and description of SwiftAir Models 180 and 235



Please prepare a memo that addresses several questions, including what data support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups, what other factors might have contributed to the accident and should be taken into account, and your overall recommendation about whether or not DynaTech should purchase the plane.

Computer-Based Simulation Tasks

The advancements of computer technology have made it possible to use performance-based simulations, which assess problem-solving and reasoning skills in large-scale assessment programs. The most prominent large-scale assessments that use computer-based simulations are licensure examinations in medicine, architecture, and accountancy. As an example, computer-based case simulations have been designed to measure physicians' patient-management skills, providing a dynamic interaction simulation of the patient-care environment.¹⁹ The examinee is first presented with a description of the patient and then must manage the case by selecting history and physical examination options or making entries into the patient's chart to request tests, treatments, and/or consultations. The patient's condition changes in real time based on the disease and the examinee's course of action. The computer-based system generates a report that displays each action taken and when it was ordered. The examinee's performance is then scored by a computerized scoring system for the appropriateness of the sequence of actions. The intent of this examination is to capture essential and relevant problem-solving, judgment, and decision-making skills required of physicians.

Some designers of new K-12 science assessments are seeking to build in such simulations, as has the National Assessment of Educational Progress (NAEP) in items that test students' abilities to design experiments, display and interpret results, and search the internet effectively. One 8th grade NAEP simulation task, for example, required students to investigate why scientists use helium gas balloons to explore outer space and the atmosphere. Below is an example of an item within this task that requires students to conduct an internet search:

Figure 6:

NAEP Science Inquiry and Simulation Tasks

Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from a spacecraft near other planets.

Using the web, investigate the answer to this question: Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons. Base your answer on more than one web page or site. Be sure to write your answer in your own words.²⁰



This task assesses students' online research skills. A related scientific inquiry task required students to evaluate their work, form conclusions, and provide rationales after designing and conducting a scientific investigation to answer this question:²¹

How do different amounts of helium affect the altitude of a helium balloon? Support your answer with what you saw when you experimented.

These simulation tasks assess problem-solving, reasoning, and evaluation skills valued within the scientific discipline, providing new possibilities for evaluating student cognition and learning. They, too, can use computer-based scoring as well as human scoring.

Using Performance Items and Tasks in Tests: A Summary of Implications for States

Features

Open-ended performance items and tasks can be used to evaluate students' abilities to solve problems, conduct research, communicate, and explain their thinking. In addition to individual state tests, such tasks are part of the SBAC, PARCC, and College and Work Readiness Assessments (CWRA). Among others, tasks can include

- Essay responses or problem solutions in response to a prompt
- Online research to answer a question
- Interactive simulations of experiments or strategies
- Designs (such as laying out a garden or designing a structure using mathematical considerations)

Benefits

Including performance items and tasks in summative tests allows states to

- More completely assess college and career-ready standards, including communication, research, and inquiry
- Evaluate higher order skills, such as analysis, synthesis, evaluation, and application of knowledge to complex problems
- Better reflect how learning is applied in real world settings (and thus strengthen validity)
- Incentivize good practice in classrooms and broaden the focus of curriculum to include the skills that are tested
- Provide opportunities to teachers to see and analyze student work and, when they are involved in scoring open-ended tasks, to deepen their understanding of the standards, curriculum, and assessment.

Considerations

Scoring of open-ended tasks requires strong task design and careful training. (See also the section on scoring below.)

- Performance items or tasks can sometimes be evaluated using computer-based AI scoring. This is true for many essays and for tests like the CWRA, as well as some simulations.
- Often these tasks must be human-scored, which adds modest costs. SBAC and PARCC developed systems for reliably scoring tasks for a few dollars per item per student.
- Reliable scoring can be achieved through training, moderation processes, and auditing.
- Teachers learn significantly and can improve their practice from the scoring process. One way to enhance teacher learning and reduce costs is to allocate professional development days for scoring, or to include teacher scoring as part of the test administration contract.

II. CURRICULUM-EMBEDDED PERFORMANCE ASSESSMENTS

Curriculum-Embedded Performance Tasks

Moving rightward along the continuum in Figure 2 toward student-directed inquiry, *curriculum-embedded performance tasks* extending over many days or weeks can test more challenging intellectual skills that come even closer to the expectations for performance found in colleges and careers. These tasks are conducted during the school year and are typically scored using common rubrics. They can be highly standardized in their design or they can allow elements of student choice (for example, choice of topic or product design) with standardized rubrics. (For an example of such a rubric for a science investigation, see Appendix A.) Several curriculum-embedded tasks can be combined into a summative score or determination, or one or more performance tasks can be combined with a traditional test (sometimes an end-of-year test) to produce a summative score.

There are several reasons to choose these kinds of assessments. First, because the tasks are embedded in classroom units that can be conducted over an extended period of time, they allow students to undertake more challenging work and demonstrate a broader range of skills that more closely resemble what they will need to do in real-life situations. Second, high-quality tasks can strengthen classroom instruction, helping teachers learn how to teach the higher-order skills the tasks embody and providing greater curriculum equity for students who experience common opportunities to do research, write about, and present their findings. This enables them to develop a deeper understanding of content and college- and career-ready skills they need.

Third, students and teachers do not experience these tasks as formal tests, as they are embedded into instruction like any assignment would be. They are simply more carefully constructed and scored, and more commonly used than an individual classroom project might be. For this reason, these tasks should not be thought of as part of “testing time.” They are more appropriately considered part of teaching and learning time, although states or districts need to put aside professional development time for scoring the tasks.

Many countries and the [International Baccalaureate \(IB\) program](#) use a combination of externally designed tasks (papers or projects) that are conducted in the classroom and scored by trained teachers in systems that are “moderated” or audited as part of their assessment system. These are often coupled with the results of an end-of-year test in producing a

summative score. The tasks typically comprise 30-60 percent of the total score. For example, the General Certificate of Secondary Education (GCSE) exams in England, like the exams in many Australian states and in Singapore, include performance tasks during the year coupled with an end-of-the-year test, usually comprised of essays and problem solutions.

The General Certificate of Secondary Education. In the General Certificate of Secondary Education (GCSE) English exam, there are a number of what might be called “through course assessments,” designed to evaluate different genres and demonstrations of reading, writing, speaking, and listening. These are either designed by a centralized exam board and marked by teachers or designed by teachers and marked by the exam board. Either way teachers determine the timing of the assessments. Together, they count for 60 percent of the total score; the remainder is from a written exam which asks students to write responses to specific prompts.

Example of Tasks: GCSE English	
Unit and Assessment	Tasks
<i>Reading Literacy Texts</i> Classroom assessment 40 marks	Responses to three texts from choice of tasks and texts. Candidates must show an understanding of texts in their social, cultural, and historical context.
<i>Imaginative Writing</i> Classroom assessment 40 marks	Two linked continuous writing responses from a choice of Text Development or Media.
<i>Speaking and Listening</i> Classroom assessment 40 marks	Three activities: a drama-focused activity, a group activity, an individual extended contribution. One activity must be a real-life context in and beyond the classroom.
<i>Information and Ideas</i> Written exam 80 marks (40 per section)	Non-Fiction and Media: Responses to unseen passages. Writing information and Ideas: One continuous writing response – choice from two options.

In GCSE Interactive Computer Technology Task, the performance assessment is a single task that combines into one major project many of the major skills taught in the class and used in the real world: researching and designing a software solution to meet a specific need, testing it with users, and figuring out improvements.

GCSE Controlled Assessment Task in Interactive Computer Technology (ICT)

Litchfield Promotions works with over 40 bands and artists to promote their music and put on performances in England. The number of bands they have on their books is gradually expanding. Litchfield Promotions needs to be sure that each performance will make enough money to cover all the staffing costs and overheads as well as make a profit. Many people need to be paid: the bands; sound engineers; and lighting technicians. There is also the cost of hiring the venue. Litchfield Promotions needs to create an ICT solution to ensure that they have all necessary information and that it is kept up to date. Their solution will show income, outgoings, and profit.

Candidates will need to: 1) Work with others to plan and carry out research to investigate how similar companies have produced a solution. The company does not necessarily have to work with bands and artists or be a promotions company. 2) Clearly record and display your findings. 3) Recommend a solution that will address the requirements of the task. 4) Produce a design brief, incorporating timescales, purpose and target audience.

Produce a solution, ensuring that the following are addressed: 1) It can be modified to be used in a variety of situations. 2) It has a friendly user interface. 3) It is suitable for the target audience. 4) It has been fully tested. You will need to: 1) incorporate a range of software features, macros, modeling, and validation checks - used appropriately. 2) Obtain user feedback. 3) Identify areas that require improvement, recommending improvement, with justification. 4) Present information as an integrated document. 5) Evaluate your own and others' work.

States could add one or more curriculum-embedded tasks as components of the state assessment in any subject area, to contribute to the overall assessment score, with proper management of the task selection and scoring. Alternatively, they could create a system, as New Hampshire has, that uses curriculum-embedded assessments as the bulk of the system, with traditional standardized tests as periodic information to validate the results of the performance tasks. (See Section IV on Comprehensive Assessment Systems.) Finally, states can offer high-quality tasks to districts for their own instructional and formative assessment use – for example in subjects and graduate levels that are not otherwise tested.

Performance Assessment Task Banks

States that are using curriculum-embedded performance tasks often create a statewide bank of tasks from among those developed by teachers that have been reviewed and validated so that they can be shared across classrooms. Some of these can be selected as common tasks used for comparisons across districts and schools. Educators in these and other states can also contribute to and draw from a task bank available nationwide to schools, districts, and states — the Performance Assessment Resource Bank²² — developed by the Council for Chief State School

Officers (CCSSO) in collaboration with the Stanford Center for Assessment, Learning, and Equity (SCALE) and the Stanford Center for Opportunity Policy in Education (SCOPE). Other states can use performance tasks from this bank that have been reviewed for quality by a team of assessment experts and, frequently, piloted and revised. These tasks are presented with the units within which they are embedded, along with rubrics and scored samples of student work. The resource bank includes tools and protocols for training educators to develop, review, revise, and score tasks with consistency.

The resource bank includes tasks which apply concepts to real world contexts. For instance, in the mathematics task below, students are asked to research the rising costs of a college education in several kinds of colleges. They are encouraged to choose schools that they may be interested in. They need to collect and analyze data, develop equations and graphs that represent the different trajectories of increases, and ultimately interpret what they have found in a new article on the subject.

Rising Cost of a College Education

STUDENT INSTRUCTIONS

A. Task context:

You are a reporter for the *US News and World Report* magazine. (They are the ones who rank colleges). You have been tasked with writing an article about the rising cost of obtaining a college education. In order to be able to write the article you first need to collect and analyze data on the cost of a college education. You will be creating equations and graphs showing the rising cost of education at different types of colleges including an in-state college, a community college, and out-of-state college, and an Ivy League college. You will provide a short (500 - 750 words max) article on the rising cost of college education. It is recommended that you choose schools that are relevant to you. Are there schools that you might consider attending in the future that you might consider researching?

These tasks require students to tackle a substantial, multi-part problem and use a range of analytic skills while producing a solution and a product that illustrates and explains their thinking.

New Hampshire and Colorado are drawing on the Performance Assessment Resource Bank while developing their own task banks. Kentucky is developing a performance task bank for science, initially, which it expects to expand to other content areas.

Science Assessments

Science is an area where curriculum-embedded assessments are widely used around the world. In the 1990s, Connecticut, Maryland, New York, and Vermont included common science inquiry tasks conducted by students in the classroom as part of their science assessments, in some cases paired with a traditional “sit-down” test at year’s end. Kentucky is developing a new science assessment that will include curriculum-embedded inquiry tasks along with a test that includes performance components in its system.

An example of one of Connecticut’s tasks can be seen in Figure 7. This kind of standardized classroom-embedded task, which all students complete, is scored by teachers using common rubrics. Before NCLB, this assessment was factored into the score on the end-of-year science test to produce a summative score used in state-level and federal reporting, as is done in many countries’ examination systems.

Figure 7:

Connecticut 9th / 10th Grade Science Assessment Acid Rain Task

Acid rain is a major environmental issue throughout Connecticut and much of the United States. Acid rain occurs when pollutants, such as sulfur dioxide from coal burning power plants and nitrogen oxides from car exhaust, combine with the moisture in the atmosphere to create sulfuric and nitric acids. Precipitation with a pH of 5.5 or lower is considered acid rain. Acid rain not only affects wildlife in rivers and lakes but also does tremendous damage to buildings and monuments made of stone. Millions of dollars are spent annually on cleaning and renovating these structures because of acid rain.

Your Task

Your town council is commissioning a new statue to be displayed downtown. You and your lab partner will conduct an experiment to investigate the effect of acid rain on various building materials in order to make a recommendation to the town council as to the best material to use for the statue. In your experiment, vinegar will simulate acid rain.

You have been provided with the following materials and equipment. It may not be necessary to use all of the equipment that has been provided.

Suggested materials:

- containers with lids
- graduated cylinder
- vinegar (simulates acid rain)
- pH paper/meter
- safety goggles

Proposed building materials:

- limestone chips
- marble chips
- red sandstone chips
- pea stone

Designing and Conducting your Experiment

- 1. In your words, state the problem you are going to investigate.** Write a hypothesis using an “If ... then ... because ...” statement that describes what you expect to find and why. Include a clear identification of the independent and dependent variables that will be studied.
- 2. Design an experiment to solve the problem.** Your experimental design should match the statement of the problem and should be clearly described so that someone else could easily replicate your experiment. Include a control if appropriate and state which variables need to be held constant.
- 3. Review your design with your teacher before you begin your experiment.**
- 4. Conduct your experiment.** While conducting your experiment, take notes and organize your data into tables.

Communicating your Findings

Working on your own, summarize your investigation in a laboratory report that includes the following:

- **A statement of the problem you investigated. A hypothesis (“If ... then ... because ...” statement) that described what you expected to find and why.** Include a clear identification of the independent and dependent variables.
- **A description of the experiment you carried out.** Your description should be clear and complete enough so that someone could easily replicate your experiment.
- **Data from your experiment.** Your data should be organized into tables, charts and/or graphs as appropriate.
- **Your conclusions from the experiment.** Your conclusions should be fully supported by your data and address your hypothesis.

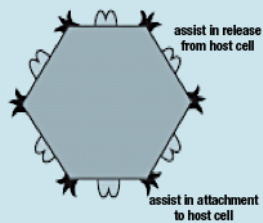
Discuss the reliability of your data and any factors that contribute to a lack of validity of your conclusions. Also, include ways that your experiment could be improved if you were to do it again.

The curriculum-embedded inquiry tasks can also be connected conceptually to the end-of-the year test as Connecticut did. Having designed and conducted their own experiments, which they wrote up during the year, students would also demonstrate their understanding of scientific inquiry in a variety of ways on the end-of-year test. For example, students might receive a sample of a report from an experiment, which

they would have to analyze in terms of the appropriateness of its methods and the validity of its results, drawing on the experiences they have had in the classroom conducting experiments. Thus, the scientific inquiry skills developed through more extensive performance tasks can also be validated on the sit-down test.

Similarly, in Victoria, Australia, students engage in a set of activities that essentially serve as “through-course assessments” that allow them to learn hands-on investigation skills while also preparing them for questions in the end of the year test. Figure 8 shows one example from a high school biology course in which students complete a set of “practical tasks” during the year. These tasks are graded according to criteria set out in the syllabus and count toward the examination score. The quality of the tasks assigned by teachers, the work done by students, and the appropriateness of the grades and feedback given to students are audited through an inspection system which provides schools feedback on all of these components.

Figure 8:

Victoria Australia Biology Course Assessment	
Classroom-based assessments – 50 percent of score (conducted during the year)	End of the Year Test – 50 percent of score Sample Question (open-ended)
<p>A set of practical tasks during the school year cover specific outcomes in the syllabus and prepare students for the end of year test. In combination, these count for 50% of the final exam score. They include:</p> <ol style="list-style-type: none"> Using a microscope to study plant and animal cells by preparing slides of cells, staining them, and comparing them in a variety of ways, resulting in a written product with visual elements. Conducting lab experiments on enzymes and membranes, and on the maintenance of stable internal environments for animals and plants. Conducting and presenting a research report on characteristics of pathogenic organisms and mechanisms by which organisms can defend against disease. 	<p>A. Scientists aim to develop a drug against a particular virus that infects humans. The virus has a protein coat and different parts of the coat play different roles in the infective cycle. Some sites assist in the attachment of the virus to a host cell; others are important in the release from a host cell. The structure is represented in the following diagram:</p>  <p>The virus reproduces by attaching itself to the surface of a host cell, injecting its DNA into the host cell. The viral DNA then uses the components of the host cell to reproduce its parts and hundreds of new viruses bud off from the host cell. Ultimately the host cell dies.</p> <p>B. Design a drug that will be effective against this virus. In your answer outline the important aspects you would need to consider. Outline how your drug would prevent continuation of the cycle of reproduction of the virus particle. Use diagrams in your answer. Space for diagrams is provided on the next page.</p> <p>C. Before a drug is used on humans, it is usually tested on animals. In this case, the virus under investigation also infects mice. Design an experiment, using mice, to test the effectiveness of the drug you have designed.</p>

Including the curriculum-embedded component offers at least four benefits:

- 1) It incentivizes and helps teachers learn to teach scientific inquiry.
- 2) It supports students in learning to design and conduct such investigations so that they begin to deeply understand the process.
- 3) It also expands curriculum equity by ensuring that all students, not just the advantaged, experience high-quality science instruction and tasks, so that performance is more equitably improved both in the classroom and on the tests.
- 4) By involving teachers, supported by assessment experts, in scoring tasks, their understanding of the standards and assessments and their shared sense of what constitutes high-quality performance are increased.

All of these things strengthen instruction and learning, as well as the quality of testing.

The practice of requiring curriculum-embedded assessments in science is widespread across the world, because learning scientific inquiry is intrinsically performance-based. The example below from Queensland, Australia, is very similar to the assessments in Great Britain, Hong Kong, Singapore, and other nations. It is a step beyond the examples from Connecticut and Victoria, because it requires students to identify and define their own, more extensive investigation. Students who have had the experience of investigations in more structured tasks will be learning how to take this next step, which might occur as a capstone assessment in which they design and conduct their own investigation in the 11th or 12th grade. (See Figure 9.)

Figure 9:

Queensland, Australia
Extended Experimental Investigation at the Senior Level (Grade 11-12)
Over four or more weeks, students must develop and conduct an extended experimental investigation to investigate a hypothesis or to answer a practical research question. Experiments may be laboratory or field based. The outcome of the investigation is a written scientific report of 1500 to 2000 words.
The student must:
<ul style="list-style-type: none">• develop a planned course of action• clearly articulate the research question and provide a statement of purpose for the investigation• provide descriptions of the experiment• show evidence of student design• provide evidence of primary and secondary data collection and selection• execute the experiment(s)• analyze data• discuss the outcomes of the experiment• evaluate and justify conclusion(s)

Kentucky is currently creating a science assessment system that will combine performance tasks that engage students in science investigations during the school year with an end-of-year test that includes open-ended tasks along with selected-response items. Teachers are helping to develop the assessments and will be involved in scoring them. The tests will meet federal requirements for a science assessment once in each grade span. In addition, a bank of performance tasks drawing on the tasks teachers have developed will make it possible for educators statewide to select and use curriculum-embedded investigations at every grade level, building a science inquiry culture throughout the state.

A sample science assessment plan that follows a similar model is shown below in Figure 10. Once in each grade span, a federally-required summative assessment would be offered, with scores combining the results of an innovative test (including constructed-response items, web-based research, and simulations that tap inquiry skills) at perhaps 50-70 percent of the score and a common investigation, scored by teachers with statewide training and moderation, comprising the other 30-50 percent of the score. (Teachers would not score their own students' work for this purpose.) In other years, teachers could use the tasks and related curriculum units pegged to the standards in their grade levels individually or on a school-wide basis, scoring the tasks themselves. Schools or districts that want to develop strong understanding and curriculum planning among teachers could sponsor joint scoring and curriculum discussions on professional development days. This approach would develop a culture of science inquiry across a state and give teachers and students regular experiences of well-designed tasks.

Figure 10:

Sample Science Assessment Plan		
Grades K-2		Locally-selected/designed performance tasks
Grade 3		Locally-selected/designed performance tasks
Grade 4	Innovative Science Test	Common curriculum-embedded science inquiry
Grade 5		Locally-selected/designed performance tasks
Grade 6		Locally-selected/designed performance tasks
Grade 7	Innovative Science Test	Common curriculum-embedded science inquiry
Grade 8		Locally-selected/designed performance tasks
Grade 9		Locally-selected/designed performance tasks
Grade 10	Innovative Science Test	Common curriculum-embedded science inquiry
Grades 11-12		Capstone science investigation (local)

Assessments in Social Studies, the Arts, and other Areas

Approaches to document-based questions that are part of the Regents exams and the AP exams in U.S. History were discussed earlier. More extensive curriculum-embedded assessments can also be used in a wide range of subjects. For example, Washington state uses state-developed classroom-based assessments (CBA), including performance assessments, to gauge student understanding of the Essential Academic Learning Requirements (EALR) learning standards in social studies, the arts, and health/fitness. Districts must report to the state that they are implementing the assessments/strategies in those content areas, but individual student scores are not reported for state accountability purposes. Below is a civics example that asks students to study a constitutional issue that balances the public good against individual preferences or freedoms, examine case law or legislation on that topic, and represent both sides of the issue in proposing a resolution. (See Figure 9.)

Figure 11:

Washington State Classroom-Based Assessment in Civics

High School
Recommended
for 11th Grade

Constitutional Issues CBA

Citizens in a democracy have the right and responsibility to make informed decisions. You will make an informed decision on a public issue after researching and discussing different perspectives on this issue.

Directions to students¹

In a cohesive paper or presentation², you will:

- State a position on the issue that considers the interaction between individual rights and the common good AND includes an analysis of how to advocate for your position.
- Provide reason(s) for your position that include:
 - An analysis of how the Constitution promotes one specific ideal or principle logically connected to your position on the issue.
 - An evaluation of how well the Constitution was upheld by a court case OR a government policy related to your position on the issue.
 - A fair interpretation of a position on the issue that contrasts with your own.
- Make explicit references within the paper or presentation to three or more credible sources that provide relevant information AND cite sources within the paper, presentation, or bibliography.

Culminating Projects and Exhibitions

Further along the continuum are longer duration projects that require several weeks or even months as students demonstrate a comprehensive set of skills within or across fields. Often, it is the student who defines the focus of the project and who is responsible for organizing the task and locating all the necessary information to complete it. The science investigation task from Queensland is an example. The student may be expected to follow a particular outline or to address a particular problem or range of requirements in the process of completing the project. The project may be judged by the teacher alone, or may be scored by one or more other teachers in a moderated process that allows teachers to calibrate their scores to a benchmark standard.

Finally, a culminating project can be designed to gauge student knowledge and skill cumulatively, including the ability to apply disciplinary standards of practice and modes of inquiry in a subject-specific or interdisciplinary way. These are competency-based assessments that evaluate deep understanding of an area of study, much like a dissertation does for PhD students. Students may study one topic for a semester or even an entire year, applying what they are learning in their academic classes to help them work on the project. In Singapore, the project must also be collaborative, integrating another key skill. The culminating project generally includes a terminal paper and accompanying product and documentation, reflecting overall cognitive development and a range of academic skills. The results may be presented to a panel that includes teachers, experts from the community, and/or fellow students.

This method of juried exhibitions is used in some examination systems abroad (for example, in the Project Work task required as part of the International Baccalaureate and the A-level exams in Singapore) and by a number of school networks in the United States.²³ Students communicate their ideas in writing, orally, and in other formats (e.g., with the use of multi-media technology or through products they have created), while they demonstrate the depth of their understanding as they respond to questions from others, rather like a dissertation defense.

Using Curriculum-Embedded Assessments

Summary of Implications for States

Features

States can include curriculum-embedded performance tasks in their systems of assessment to deepen learning and provide greater curriculum equity. These can occur over several days or weeks to evaluate more challenging intellectual skills that come even closer to the expectations for performance found in colleges and careers.

- Tasks can be highly standardized in their design or they can allow elements of student choice (e.g., choice of topic or product design) with standardized rubrics.
- Common tasks, embedded in curriculum units, can, properly scored, provide comparable results across schools and districts.
- Several of these can be combined into a summative score or determination, or one or more performance tasks can be combined with a traditional test to produce a summative score.
- When tasks and tests are combined, they can be designed together to reinforce knowledge and skills, supporting applied learning and conceptual understanding.
- A system of assessments can be constructed to use a strategic combination of tests, common performance tasks, and locally-developed or selected tasks to support validation, deeper learning, and formative information for teachers and students.

Benefits

Including curriculum-embedded tasks as part of the system of summative assessments allows states to

- More completely assess college and career-ready standards, including independent and collaborative student-initiated research and inquiry; ability to take and use feedback productively; and oral, written, and multimedia communication.
- Evaluate higher order skills, such as analysis, synthesis, evaluation, and application of knowledge to complex problems.
- Better reflect how learning is applied in real world settings (and thus strengthen validity)
- Create greater curriculum equity for students by using assessments to create strong units and instructional practices across classrooms, rather than having only some students experience instruction for deeper learning.
- Increase teachers' understanding of the standards and of high-quality teaching and assessment by involving them in developing, reviewing, and scoring tasks.

Considerations States that want to use curriculum-embedded assessments will need systems to develop and acquire high-quality tasks and engage in reliable scoring. (See also section IV on task design, comparability, and scoring.)

- As one source, states can draw from the CCSSO/SCALE/SCOPE Performance Assessment Resource Bank²⁴ which includes high-quality tasks mapped to standards, grade levels, and learning progressions, along with rubrics, scored samples of student work, and protocols for developing, reviewing, and scoring tasks. The bank can be used for common tasks (which can be kept secure as needed) and for tasks selected for use at the classroom, school, or district level.
- States can also contribute to the bank in order to have tasks developed by their teachers reviewed and revised to meet task quality standards.
- Where common tasks are used, required materials should be readily available in the schools, in homes, or online so that all students and schools can readily and fairly engage in the necessary activities.
- States may want to establish a technical advisory committee or assessment review panel to evaluate and approve performance tasks, and to oversee scoring plans and audits.
- States generally create guidelines for what kind of assistance and feedback are allowable in the classroom as tasks are conducted.
- To support reliable scoring, states will need to create plans for training and calibration. Teachers may come together for training and scoring sessions or they may engage in distributed online scoring that embeds a training and calibration process.
- It will be useful to integrate time for teacher scoring into the annual school schedule, and perhaps to link it to professional development time in order to experience the benefits of both scoring and related reflections on curriculum, instruction, and assessment.
- Finally, as curriculum-embedded tasks are part of the instructional process, they should not be thought of as part of “testing time.” They are more appropriately considered part of teaching and learning time.

III. PORTFOLIOS / COLLECTIONS OF EVIDENCE

Portfolios are collections of evidence about students’ learning, organized around a set of standards or competencies to be demonstrated in a single content area or across multiple content areas. They are often collections of performance tasks,

although other evidence, for example, from traditional sit-down tests or out-of-school internships, can also be included.

Single-subject portfolio systems have been used by states including Kentucky and Vermont, both of which have writing and mathematics portfolios, and by the Advanced Placement (AP) program for course assessments in Art, Technology, AP Research, and AP Seminar. In addition, portfolios covering multiple disciplines are increasingly common at the high school level. Rhode Island has long used portfolios for graduation. Oregon now allows a portfolio as one of several options for graduation. New Hampshire's system envisions a graduation capstone project or portfolio. Some districts (e.g., Pasadena, CA), and many networks of schools (Envision, New Tech High, Asia Society, Big Picture Learning, the Internationals Network) require portfolios for graduation. Schools participating in the New York Performance Standards Consortium are authorized by New York State to use these assessments in lieu of state Regents examinations.

Single-Subject Portfolios

Vermont was an early pioneer in using embedded classroom assessments for accountability and to guide curriculum development. Vermont was the first state to develop portfolios in ELA and math during the 1990s, and the state's experience produced considerable learning about how to use this assessment approach effectively.

Initially, teachers and students jointly selected student work to include in each student's mathematics and writing portfolios, but there was little consistency across students in what kind of work was included. This variation made the first round of portfolios difficult to score reliably. However, the state soon created more standardized portfolios featuring common task expectations and analytic rubrics, which could be scored with much greater consistency.²⁵ Teachers came together in the summers to score the portfolios, engaging in a moderated process designed to produce consistency across raters in how they judged the work.

Although NCLB ended the use of Vermont's portfolios for state accountability, most districts in the state continue to use these strategies locally. Currently, each school's Local Comprehensive Assessment System must assess students in the required standards not covered by the state assessment.²⁶ With the goal of placing "classroom assessment at the core of the assessment system,"²⁷ the state furnishes a variety of assessment tools that schools may use in developing their systems. For example, in the content areas of mathematics and writing, the state offers benchmarks, rubrics, calibration materials, and data analysis tools to effectively use mathematics and writing portfolios as local classroom assessments.

Additionally, the Department of Education reviews district-based assessment systems and gives specific guidance to teachers and other educators

responsible for scoring common assessments.²⁸ For example, districts “need to use common, agreed upon criteria for student expectations, [use either] scoring scales or rubrics, and benchmark performances in order to make consistent judgments about the quality of student work.”²⁹

Kentucky’s writing and math portfolios were begun as part of the Kentucky Instructional Results Information System (KIRIS), a performance-based assessment system introduced in 1992. Eventually the mathematics portfolio was replaced by performance tasks, while the writing portfolio continued for two decades. The Writing Portfolio was used in grades 4, 7, and 12 and an On-Demand Writing Assessment was used in grades 5, 8, and 12.

Figure 12:

Kentucky’s Writing Portfolio

Kentucky’s writing portfolio was designed to ensure that students would write in several major genres, toward a common set of criteria. A 3-piece portfolio was required in grades 4 and 7, and a 4-piece portfolio was required in grade 12. In addition to a letter to the reviewer, the work samples included

- **Personal expressive writing** in the form of a Personal Narrative focusing on one event in the life of the writer; a Memoir, focusing on a person and the student’s relationship with the person; a Vignette which captures a moment in time in the life of the writer and focuses on painting a picture with words, or a Personal Essay, which focuses on a central idea supported by a variety of incidents in the writer’s life.
- **Imaginative writing** in the form of a short story, poem, script, or play
- **Transactive writing** which presents/supports a position, defends a conclusion, tells about a problem, explains a process or concept, or informs. (These selections may include forms such as letters, brochures, and articles, among other appropriate forms.)
- In grade 12, **transactive writing with an analytical or technical focus**.

The writing samples were scored by teachers using common rubrics, supported by scored benchmark portfolio samples, evaluating common criteria:

Purpose/Audience – Students demonstrate a clear sense of the reason(s) for producing a piece of writing. They meet the needs of the audience by focusing on the reason for the piece.

Idea Development/Support – Students decide which idea(s) to develop and make the idea(s) clear to the reader. Students support the idea(s) by elaborating on them with relevant details.

Clear Organization – Students arrange ideas in a clear and logical manner. They join ideas in a smooth way that guides the reader through the piece of writing.

Sentence Level Meaning – Students compose sentences that are grammatically correct, as well as varied in length and structure.

Use of Language – Students use wording and language that demonstrate standard usage. They choose correct and effective words with growing precision and sophistication.

Correctness/Conventions – Students spell correctly, use correct punctuation, and capitalize letters according to standard rules.

The state provided training to teachers, who scored their own students' portfolios. Kentucky used an audit procedure by which samples of portfolios were scored centrally and audit results reported back to schools with additional scorer training provided to teachers as needed. Over time, the scores became highly reliable. By 2008, the agreement rate (exact or adjacent scoring) for independent readers involved in auditing school-level scores was over 90 percent.³⁰

The benefits of a portfolio process include the fact that common standards and high-quality tasks can guide classroom practice throughout the school year; students experience similar kinds of high-quality instruction across classrooms and schools; and students learn how to revise work toward high standards. Teachers' involvement in orchestrating and scoring the assignments that are part of the portfolio helps them learn about the curriculum standards and about how to support learning toward the standards, as well as how to develop curriculum and performance assessments for the classroom.

These portfolios had a noticeably positive effect on instruction. Researchers studying the Vermont and Kentucky reforms found considerable evidence that teachers were changing their classroom practices to support problem solving and communicating in mathematics and writing. Furthermore, Kentucky teachers were more likely to report that open-response items and portfolios had an effect on practice than multiple choice items, adding credence to the idea that performance assessments could help create "tests worth teaching to." Both states experienced increases in their students' achievement on NAEP during these years.

Other single subject portfolios have been used by the College Board for Advanced Placement courses. The College Board has long used an Art portfolio and has recently developed three courses — the [AP Computer Science Principles](#) (CSP), [AP Research](#), and [AP Seminar](#) — in which students complete performance tasks during the academic year with components submitted using the [AP Digital Portfolio](#).

Two new AP courses — **AP Seminar** and **AP Research** — are of particular interest for evaluating college and career readiness. The courses together comprise the AP Capstone, a College Board program that “equips students with the independent research, collaborative teamwork, and communication skills that are increasingly valued by colleges. It cultivates curious, independent, and collaborative scholars and prepares them to make logical, evidence-based decisions.”³¹ AP Capstone was developed in response to feedback from higher education about what students really need to be able to do to be college ready.

The two AP Capstone courses, with their associated performance tasks, assessments, and application of research methodology, require students to

- Analyze topics through multiple lenses to construct meaning or gain understanding
- Plan and conduct a study or investigation
- Propose solutions to real-world problems
- Plan and produce communication in various forms
- Collaborate to solve a problem
- Integrate, synthesize, and make cross-curricular connections

In AP Research, students are assessed on an academic paper of 4,000 to 5,000 words based on an original research question, along with a presentation and oral defense of research to a panel of at least three members, including their AP teacher.

In the AP Seminar, five different work samples are collected and assessed,³² then combined with an end-of-course exam to create the final summative score. These include a team research project and multimedia presentation (20 percent altogether), along with an individual research-based essay, multimedia presentation, and oral defense (35 percent altogether). All of these are scored by the classroom teacher with the written products’ scores validated by the College Board. The end-of-course exam (45 percent altogether) consists of 3 short-answer questions associated with analyzing an argument and a longer essay that produces an evidence-based argument. This is scored by other College Board teachers who teach the course and participate in the annual AP scoring process.

Multiple Subject Portfolios

A growing number of school networks and districts use collections of evidence or portfolios for graduation, as do some states ([Rhode Island](#), for all students; Oregon, as an option for demonstrating graduation competencies; and New York, for the [New York Performance Standards Consortium](#) schools, which operate on a waiver from traditional Regents exams). These are designed to demonstrate that students have met defined standards or competencies within and across subject areas. These, too, are scored with common rubrics, often with teacher training and moderation to support comparability.

Similarly, the National Academies Foundation has developed a portfolio model used in its career academies and scored with common standards across hundreds of schools nationally. Both colleges and employers can use the portfolio to evaluate student learning and accomplishments.

The Rhode Island High School Diploma System³³ requires that all students must demonstrate proficiency in applied learning skills — critical thinking, problem solving, research, communication, decision making, interpreting information, analytic reasoning, and personal or social responsibility — across six core content areas. The Diploma System requires local districts to determine, with state guidance and review, how they will certify mastery of content knowledge as well as the ability to apply that knowledge to real world projects and problems through portfolios, exhibitions, or a certificate of mastery. The state’s description notes

For decades, employers and colleges complained that applied skills are sorely lacking in current high school graduates. Merely remembering facts is only a good first step toward a true subject mastery, which involves using facts and formulas to solve problems in widely different contexts. The mechanics of English are only valuable if a student can compose competent, effective business letters to a variety of clients, co-workers or potential employers, for example.... After high school, employers and higher education evaluate their workers or students primarily from evidence of mastery – such as completed and on-time tasks, written work, plans, designs, products, records and so forth.³⁴

Students demonstrate applied learning skills through evidence of mastery from presentations – such as speeches, projects, or performances – or from products – such as essays, collections of short stories, or science journals. In the body of evidence treating the core content areas and Applied Learning standards, students must include one successfully-completed on-demand task, one extended task, and one task reflecting one of their own interests or passions. A goal of the diploma system is that

... it harnesses students' interests in the service of their own learning. Traditional education asked students to 'park' their passions at the door, which invited alienation

among those students who find course work irrelevant to their real concerns. School advisors and content-area teachers help students design exhibition and portfolio projects that satisfy their own natural thirst for information and skills.

As one example, [Central Falls High School's portfolio requirement](#) is designed to reflect the students' best work over a four-year period demonstrating the Applied Learning standards in each of the core content areas. It is compiled over the course of each year, with a written reflection to accompany each of the selected entries. Some of these entries are required by teachers while others are chosen by the student to be a part of their final portfolio. At the end of each school year, students make a presentation to their Advisory class on entries selected for that year. Each entry ultimately placed in the graduation portfolio is scored on a common rubric used for that type of task. A given entry will generally address several of the proficiencies. Students can tap a variety of learning experiences to provide indicators of their Performance-Based Graduation requirements as a Creative problem solver, Effective communicator, Skillful user of technology, Responsible member of the community, and Supporter/performer of the arts.

A final Graduation Portfolio presentation to the Graduation Portfolio Review Committee takes place during their senior year. This committee is comprised of administrators, teachers, support staff, parents, and prominent members of the community, who score the presentation using a common rubric to determine if proficiency is achieved.

Another example of a multi-subject portfolio is that used by the schools in the New York Performance Standards Consortium. All of the schools include at least four entries in their portfolio:

- An analytic essay (often a literary analysis)
- An applied mathematics product (involving mathematical modeling)
- A science investigation
- A research paper (often a social science paper)

Some of the consortium schools also require an arts exhibition, a world language demonstration, and/or a presentation of learning from an internship. Among the assessments, students must provide evidence of competence in

oral and written communication, critical thinking, technology use, and other 21st century skills. They present selected entries to a jury of teachers and external judges from local colleges and businesses in a portfolio defense that includes a formal presentation plus questions and answers about the work, much like a dissertation defense.

Across schools, the portfolio entries and defenses are evaluated using common scoring rubrics that reflect critical skills in each discipline. Teachers are trained to calibrate their scoring within schools and departments, and they periodically engage in cross-school moderation sessions to calibrate the scoring across the consortium as a whole.

This approach is not unlike that taken in Queensland, Australia, where schools use a system of performance assessments with external tests as additional information in alternate years. At the high school level, a student's work is collected into a portfolio that is used as the primary measure of college readiness. Portfolio scoring is moderated by panels that include teachers from other schools and professors from the higher education system. A statewide examination in 12th grade serves as an external validity check, but not as the accountability measure for individual students.³⁵

Assessments can strengthen student learning when

- they are clearly linked to standards that are reflected in the rubrics used for scoring the work;
- these criteria are made available to students as they are developing their work;
- students are given the opportunity to engage in self- and peer review using these tools;
- assessments ask them to exhibit their work in presentations to others, where they must both explain their ideas or solutions and answer questions that probe more deeply; and
- students revise the work to address these further questions and better meet the standards.

Portfolios offer some particular benefits for developing self-directed learners. Portfolio processes assume that students are a primary consumer of the information they produce, as students own their own portfolio and must typically choose and sometimes revise the work samples they will submit to meet the standards. The process develops students' metacognitive skills and gives them opportunities for reflection and revision. As students see their own

progress over time and reflect on how they have improved and can improve further, they develop a growth mindset. Not incidentally, these processes also support student learning by deepening teachers' learning about what constitutes high-quality work and how to support it, both individually and collectively as a staff.

Furthermore, through the use of rubrics and public presentations, students can receive feedback that is specific and detailed, providing them a much better idea of how to improve than would an item analysis from a standardized test or generalized comments from a teacher on a paper such as "nice job" or "good point." When students receive feedback of many different types from different sources, they are able to begin to triangulate among them to identify patterns of strength and weakness beyond just the specific questions they got right or wrong. This more comprehensive, holistic sense of knowledge and skills can empower the learner and build self-awareness and self-efficacy.

When students repeatedly develop and revise projects and exhibitions evaluated according to rigorous standards, they internalize standards of quality and develop college- and career-ready skills of planning, resourcefulness, perseverance, a capacity to use feedback productively, a wide range of communication skills, and a growth mindset for learning — all of which extend beyond the individual assignments themselves in shaping their ability to learn to learn in new contexts.

Using Portfolio Models Summary of Implications for States

Features

States can include portfolios in their systems of assessment for a single subject, such as writing, or across several subject areas.

- Work samples for the portfolio are selected because they demonstrate a set of competencies and represent key subject matter.
- The tasks can be standardized in their design or they can be teacher or student-designed to address the competencies.
- Students often present and defend their work to a jury of educators, peers, and, sometimes, external judges.
- Common rubrics are used to evaluate the individual tasks and the presentation.
- Portfolios can be scored both by task and overall.

Benefits

Including portfolios as part a system of assessments allows states to

- More completely assess college and career-ready standards, including independent and collaborative student-initiated research and inquiry; ability to take and use feedback productively; and oral, written, and multimedia communication.
- Evaluate higher order skills, such as analysis, synthesis, evaluation, and application of knowledge to complex problems.
- Better reflect how learning is applied in real world settings.
- Increase the likelihood that common standards and high-quality tasks will guide classroom practice throughout the school year, and that students will experience similar kinds of high-quality instruction across classrooms and schools.
- Involve students in a process that explicitly develops their metacognitive skills by giving them opportunities for reflection as they choose and revise work to meet standards.
- These processes also deepen teachers' learning about what constitutes high-quality work and how to support it, both individually and collectively as a staff.

Considerations

States that want to incorporate portfolios into their assessments will want to think about how to support classroom work to ensure high-quality portfolio submissions and ensure scorability. (See also section below on scoring.)

- To be scorable with high inter-rater reliability, portfolios must be comprised of tasks that clearly measure the same set of standards with the same or similar genres of tasks (rather than open-ended choices of work samples).
- Teachers will need clear specifications, training, and readily available technical assistance to learn how to select, design, and support student work with guidelines for what kinds of assistance are appropriate.
- States may want to establish a technical advisory committee or assessment review panel to evaluate and approve portfolio specifications, and to oversee scoring plans and audits.
- As with other curriculum-embedded tasks, states will need to create plans for training and calibration. As in Kentucky and the AP program, an audit system can be established to re-score a subset of tasks (10-15 percent is common) to evaluate comparability and to re-train raters as needed.
- Where portfolio defenses or exhibitions are to be presented, schools will need to learn strategies from other experienced schools for adjusting the use of school time to support the process.

IV. COMPREHENSIVE ASSESSMENT SYSTEMS

A comprehensive assessment model is designed to provide the opportunities for high-quality teaching, student learning, and evaluation in a carefully integrated system that artfully blends state and local components to provide reliable information *about* learning while minimizing unnecessary testing and maximizing the benefits of assessment *for* learning. As in many jurisdictions abroad, periodic statewide standardized measures are used to validate local assessment results, while classroom-embedded performance assessments are used to inform instruction, provide feedback to students and teachers, and enable diagnostic decisions, as well as to provide evidence of ambitious student learning. Collections of evidence that allow students to evaluate their own progress and revise and present their work to meet a standard can also play a role in giving students ownership and agency in the process of developing evidence of their readiness for college and careers.

New Hampshire’s PACE system (Performance Assessment for Competency Education), piloted in an expanding number of districts, and eventually to be used statewide, is a comprehensive model that uses a mix of assessments strategically to leverage high-quality learning and teaching. The system includes a standardized test once in each grade span in ELA and math, with common, performance tasks in the other years augmented by locally developed tasks to make determinations about student proficiency. New Hampshire is developing a capstone project/portfolio system at grade 12 through which students will demonstrate graduation competencies with an exhibition and defense before a jury of educators and peers. This component will be implemented in 2017-18. The state hopes to translate its previous NCLB waiver into an innovative assessment pilot under ESSA to continue to develop this model.

Figure 13:

PACE System of Assessments (New Hampshire)

[PBA = Performance-Based Assessment]

Grade	ELA	MATH	SCIENCE
K-2	Local PBA	Local PBA	Local PBA
3	Smarter Balanced	Common PACE PBA	Local PBA
4	Common PACE PBA	Smarter Balanced	Common PACE PBA
5	Common PACE PBA	Common PACE PBA	Local PBA
6	Common PACE PBA	Common PACE PBA	Local PBA
7	Common PACE PBA	Common PACE PBA	Local PBA
8	Smarter Balanced	Smarter Balanced	Common PACE PBA
9	Common PACE PBA	Common PACE PBA	Common PACE PBA
10	Common PACE PBA	Common PACE PBA	Common PACE PBA
11	SAT	SAT	Common PACE PBA
12	Capstone project / Portfolio with Exhibition and Defense		

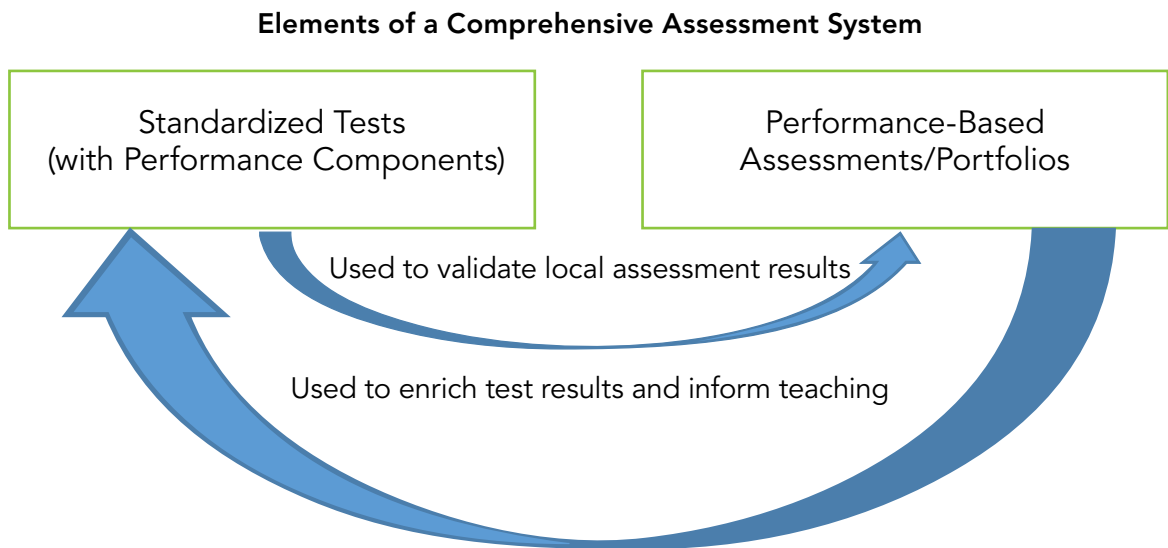
New Hampshire’s system of common tasks plus local performance tasks, validated periodically by standardized tests, is similar to the system in Queensland, Australia. There, national testing occurs at grades 3, 5, 7, and 9, and the state offers a reference exam at grade 12 that is used as a comparison point at the school level for the scores on the graduation portfolios. Most assessment is conducted through common statewide performance tasks that are administered by schools — the centrally developed Queensland Comparable Assessment Tasks — plus a very rich system of local performance assessments that are developed at the school level, but are subject to quality control and moderation of scoring by a state panel. The Queensland Curriculum, Assessment, and Reporting Framework (QCAR) helps provide consistency from school to school based on the state’s content standards, called *Essential Learnings*, which include unit templates, guidance for assessments, and rubrics in each subject. These include extended research projects, analyses, and problem solutions across fields.

Figure 14:

Queensland’s System of Assessments		
	Pre-Secondary Level	Senior Level (Grades 11-12)
External tests	National tests of literacy and numeracy at grades 3, 5, 7, 9 — Centrally scored.	Queensland Core Skills Test, grade 12
Locally administered performance tasks	Queensland Comparable Assessment Tasks (QCAT): Common performance tasks at grades 4, 6, and 9 — Centrally designed and locally scored.	Course assessments, outlined in each syllabus — locally scored / externally moderated
Locally developed assessments	Local performance assessment systems — Locally designed based on the <i>Essential Learnings</i> curriculum framework. Locally scored and externally moderated.	Graduation portfolios — locally scored/externally moderated by a state panel

Like Queensland’s system, New Hampshire has built systems to develop high quality tasks, to train teachers to develop and score these tasks, and to calibrate scoring so that it is consistent across schools and districts. Determinations of student proficiency are made by reviewing the collection of local and common tasks each year. These scores are compared to the outcomes of students on the standardized tests given periodically to validate that the system is working in a consistent fashion. (See Figure 15.)

Figure 15:



New Hampshire's System of Assessments

To ensure its students' preparation for college and careers, New Hampshire has created a system of assessments that is tightly connected to curriculum, instruction, and professional learning. In addition to the Smarter Balanced Assessments in English language arts and mathematics offered at one grade level each in elementary and middle school, this system includes a set of common performance tasks that have high technical quality in the core academic subjects, locally designed assessments with guidelines for ensuring quality, regional scoring sessions, and local district peer review audits to ensure sound accountability systems and interrater reliability, a web-based bank of local and common performance tasks, and a network of practitioner "assessment experts" to support schools.

The state's view is that a well-developed system of performance assessments that augment the traditional tests will drive improvements in teaching and learning, as they "promote the use of authentic, inquiry-based instruction, complex thinking, and application of learning...[and] incentivize the type of instruction and assessment that support student learning of rich knowledge and skills." Because the state's theory of change identifies educator capacity as essential to this goal, the system will also offer a strategic approach for building the expertise of educators across the state, by

organizing professional development around the design, implementation, and scoring of these assessments, which model good instruction and provide insights about teaching and learning.

Assessment information gathered from the local assessment system, including common and locally-developed performance tasks, provides the bulk of the information used for school, educator, and student accountability systems. Meanwhile, the large-scale assessment systems are a means to validate the accountability determinations. The state's approach is to

- Develop a process, tools, and protocols for supporting districts and schools in developing and validating high-quality **local performance tasks**, along with guidance for teachers in how to use these to enhance curriculum and instruction.
- Assemble both the common and locally developed tasks into a **web-based bank** of validated performance tasks to be used for formative as well as summative assessments.
- Organize **professional development institutes** for cohorts of schools to support task design, validation, and reliable scoring, as well as data analysis to track student progress and inform instruction. Build cohorts of *expert teacher leaders* in each content area to support this work.
- Create **regional support networks** led by practitioner assessment experts to help build capacity in schools and to support regional task validation and calibration scoring sessions, with a goal of 80 percent or greater inter-rater reliability on locally-scored tasks.
- Maintain technical quality and consistency through **district peer review audits**, in which districts will submit evidence of their performance assessment systems to peer review teams of external practitioners, who will review the evidence based on common criteria.

A key part of the accountability system, these audits will examine how districts administer common and local tasks, manage a quality assurance process, develop educators' skills, and design policies and practices that support the state performance assessment system.

Several states, such as Connecticut, Kentucky, Maine, and Vermont, built versions of such comprehensive systems of assessment during the 1990s, using a combination of periodic on-demand tests, which included performance items, alongside curriculum-embedded performance tasks and portfolios. Studies of these systems found that the mix of assessments encouraged instructional strategies fostering reasoning, problem solving and communication, as well as a focus on research and writing.³⁶ Furthermore, the regular use of performance assessments measuring complex thinking skills has been found to influence student learning and achievement.³⁷

Systems where performance assessments are regularly embedded in classroom instruction produce stronger learning for students in part by ensuring that students are undertaking intellectually challenging tasks. If teachers use these kinds of assignments consistently, with feedback and opportunities to revise to meet high standards, the level of rigor in the classroom increases. In addition, these assessments can provide information to teachers regarding how students think and try to solve problems. This feedback allows teachers to diagnose students' strengths as well as gaps in understanding.

The clear criteria and rubrics that accompany well-designed performance tasks and portfolio entries also help improve teaching and learning. As rubrics yield multiple scores in different domains of performance, reflecting students' areas of strength and weakness, they help teachers identify what kinds of assistance students need and tailor instruction accordingly.³⁸ They also help students learn how to improve their own work, especially if the criteria carry over across multiple formative and summative assessments over time. For example, if writing is repeatedly evaluated for its use of evidence, accuracy of information, evaluation of competing viewpoints, development of a clear argument, and attention to conventions of writing, students begin to internalize the criteria and guide their own learning more productively.

Gains in student learning increase as students spend more time using such criteria to discuss content, discuss the assignment, and evaluate their products.³⁹ An analysis of hundreds of studies by British researchers Paul Black and Dylan Wiliam found that the regular use of open-ended formative assessments with clear criteria to guide feedback, student revision, and teachers' instructional decisions produces larger learning gains than most instructional interventions that have been studied.⁴⁰

Developing Comprehensive Assessment Systems: Summary of Implications for States

Features

States can create a **comprehensive system of assessments** using both state and local sources of information — periodic standardized tests measuring certain aspects of students' learning that are assessable in a testing context, including performance items that measure analytic skills, augmented by local performance assessments that can support and evaluate more complex abilities. Tests are used periodically to validate the judgments made based on the richer data produced by local assessments, which can include statewide common tasks as well as locally-selected tasks based on the standards.

Benefits

Creating comprehensive systems of assessment can

- Reduce testing time, while more completely assessing college and career-ready standards with classroom-based tasks and providing information throughout the year to improve teaching and learning.
- Create more coherence in instructional efforts, if assessments are orchestrated to allow teachers and students to focus on the same standards across assessment vehicles.
- Evaluate and develop deep understanding of content along with co-cognitive skills, for example, the ability to design and conduct extended investigations; to collaborate; to communicate in multiple forms; to plan and persevere in implementing complex tasks, exhibit resilience, use feedback productively, and learn-to-learn.
- Increase rigor and equity in the classroom by ensuring that students are engaging in challenging work guided by common standards and high-quality tasks across classrooms and schools.
- Improve student achievement through both the quality of the tasks and the quality of feedback by using rubrics that provide more information about strengths and weaknesses that can be addressed through instruction and revision of work.
- Deepen teachers' learning about what constitutes high-quality work and how to support it, both individually and collectively as a staff.

Considerations

States that want to create comprehensive assessment systems will want to design their standardized tests and related performance assessments to complement each other in providing useful, valid assessment decisions.

- Tests and tasks should be designed to measure overlapping constructs in ways that well represent the standards efficiently.
- Systems of task design, scoring, and evaluation of results should be designed to support and evaluate comparability across tasks, venues, and assessment contexts.
- Teachers should receive training and readily available technical assistance to learn how to select, design, support, and score student assessments, as well as how to use the results to improve instruction.
- States may want to establish an assessment quality review panel to set standards for task design, evaluate and approve tasks used for common assessments, and oversee scoring plans and audits.
- States can develop cadres of expert teachers who can lead institutes and teacher networks involved in task design, review, selection, scoring, and improvements in curriculum and instruction.

COMPARABILITY, TASK DESIGN, AND SCORING

Perhaps the most common questions about using performance assessments as part of state accountability systems have to do with the comparability of results across settings and scorers. The key to comparable assessment lies in the design of tasks and rubrics on the one hand, and the implementation of thoughtful scoring systems on the other.

New Hampshire's strategies for establishing comparability in scores on its performance assessments, for example, include guided development with expert review of tasks and rubrics, along with training and calibration of scorers. To evaluate the success of these efforts, the state has regularly conducted comparability analyses, reported as part of its waiver agreement to the U.S. Department of Education, including

- within-district inter-rater agreement and cross-district calibration audits on the common tasks used across schools and districts;
- comparisons of individual student-level annual determinations in grades using performance assessments and those using statewide standardized assessments.⁴¹

These have found strong agreement among raters, improving over time as expected in a new system, and acceptable levels of comparability across assessments.

TASK DESIGN

A well-designed performance assessment begins with clarity about the knowledge and skills to be assessed and the kinds of performances that should be elicited by the assessment. The design should be guided by state standards, as well as the purposes of the assessment, and the intended inferences to be drawn from the assessment results.⁴²

Task models, sometimes called templates or task shells, help ensure the cognitive skills of interest are assessed. Task models can be developed for performance tasks that allow for tasks to be designed that assess the same cognitive processes and skills, and a scoring rubric can then be designed for the tasks that can be generated from a particular task model. The use of task models for task design allows for an explicit delineation of the cognitive skills to be assessed, and can improve the generalizability of the score inferences.

Assessments are stronger when test specifications are clear about what

cognitive skills, subject matter content, and concepts are to be assessed and what criteria define a competent performance.⁴³ Specifications of content, skills, and criteria can guide templates and scoring rubrics that are used with groups of tasks that measure the same sets of skills. Rubrics and templates help ensure that both the content of the assessment and its scoring are comparable across settings, versions, and scorers.⁴⁴

Quality scoring rubrics that support validity and scoring reliability

- Are designed for a family of tasks or a particular task template;
- Include criteria aligned to the processes and skills that are to be measured — for example, in a mathematics task, students' computational fluency, strategic knowledge, and mathematical communication skills;
- Develop criteria for judging the quality of the performance with the involvement of content and teaching experts who know the domain and understand how students of differing levels of proficiency would approach the task;
- Identify score levels that reflect learning progressions as well as each of the important scoring criteria; and
- Are validated through research with a range of students.⁴⁵

More valid and reliably-scored tasks result, in part, from careful review and field testing of items and rubrics to ensure they measure the knowledge and skills intended. This can include interviewing students as they reflect on what they think the task is asking for and how they tried to solve it.⁴⁶ The individual piloting of tasks also provides an opportunity for the examiner to pose questions to students regarding their understanding of task wording and directions, and to evaluate their appropriateness for different subgroups of students, such as students whose first language is not English.

Field testing provides additional information regarding the quality of the tasks, including the psychometric characteristics of items. This includes analyzing student work to ensure that the tasks evoke the knowledge and skills intended, ensuring the directions and wording are clear, and testing different versions of tasks to see which work best across different groups of learners. When these processes are followed, developers have been able to create tasks that are more clearly valid for their intended purposes and are able to be more reliably scored.

SCORING

Perhaps the most frequently asked question surrounding these assessments is how to ensure comparability in scoring across different raters. It is necessary but not sufficient to have well-developed tasks and rubrics. Most of the systems described earlier, both in the United States and abroad, use common scoring guides, or rubrics, and engage teachers who are graders in training, calibration, and moderation processes to ensure consistency.

Much has been learned about how to establish effective processes of training and moderation. In the moderation process, teachers receive training and then score and discuss model answers until their judgments are reliable — that is, that they accurately represent the standards and are consistent with one another. Sometimes these moderation processes occur within schools; at other times, teachers are assembled from across a region. Teachers use benchmark examples of student work at different levels along with a rubric or set of scoring criteria to calibrate their own judgments. As teachers learn to look for the key features of the work expressed in the criteria, they become more aware of the elements of strong student performance. As they continue to score and discuss the work, they fine-tune their capacity to evaluate so that high rates of reliability are achieved.

Developing a shared understanding of student competence among educators relies on discussion regarding specific student performance on specific tasks. Strengthening and expanding this understanding from year to year is facilitated by the creation of professional learning communities that develop shared norms, standards, and practices.

This process drove the strong inter-rater reliability that was achieved in the Kentucky writing portfolio, for example. Moderated scoring processes allowing for these conversations among professionals working together regularly over time was critical to these results, as was the construction of a set of well-specified tasks within particular genres, with well-constructed scoring rubrics, and a strong audit system that provided feedback to schools. Many developers of performance assessments have learned how to manage these processes in ways that achieve inter-rater reliabilities around 90 percent, matching the level achieved in the Advanced Placement system and on other long-standing tests.

A variety of systems for calibration and moderation of teacher scoring exist around the world. In New York State, teacher scoring of Regents examinations has been conducted at the school or regional level following training and is supplemented by a regular audit of scores from the state department of education, which can follow up with both rescoring and retraining of

teachers. In Alberta, Canada, teachers have been convened in centralized scoring sessions that involve training against benchmark papers and repeated calibration of scores until high levels of consistency are achieved. All scoring occurs in these sessions with “table leaders” continually checking and re-checking the scoring for consistency, while it is going on.

In the small state of Vermont, teachers came together in the summer to conduct centralized scoring. Kentucky’s solution (similar to the strategy used in New York for the state Regents examinations) was to have local educators score their students’ work in the writing portfolio, while the state audited the local scoring on a sampling basis and providing additional training as needed. For example, at the end of the second year of assessment, Kentucky audit results showed that the scores submitted by some schools were inappropriately high. These audit results were verified by an audit of the audit. Teachers in schools whose scores were found to be inaccurate were given extra training; they rescored their portfolios with close monitoring for accuracy; and the new scores, which were considerably more comparable, became the scores of record. The following year, the writing portfolio scores in the previously audited schools, where extra training was furnished, were found to be accurate. The audit sample design was such that over a three-year period all schools would have their portfolio scores audited and derive the benefit of additional training, if needed.⁴⁷ Ultimately, Kentucky reached very high levels of inter-rater reliability, with score agreements (exact and adjacent scores) between teachers and auditors of over 90 percent.⁴⁸

In England and Singapore, similar strategies are used, with benchmark papers and student “record files” used to train teachers and calibrate scoring. In addition, moderation processes are used within schools for teachers to calibrate their scores to benchmarks and to each other, while external moderators also examine schools’ scored examinations and initiate additional training where it is needed. At the high school level, examination boards perform these functions of training and calibrating scorers.

In Queensland, Australia, samples of performance tasks from schools are rescored by panels of expert teachers, who guide feedback to schools and potential adjustments in scores. In Victoria, Australia, the quality and appropriateness of the tasks, student work, and grades is audited through an inspection system, and schools are given feedback on all of these elements. In both of these jurisdictions, statistical moderation is used to ensure that the same assessment standards are applied to students across schools. The schools’ results on external exams are used as the basis for this moderation, which adjusts the level and spread of each school’s performance assessments of its students to match the level and spread of the same students’ collective scores on the common external test score.

In sum, it is possible to train qualified raters to score well-constructed, standardized performance tasks with acceptable levels of consistency using thoughtful rating criteria. The keys to achieving consistency among raters on performance tasks include

- 1) selecting raters who have sufficient knowledge of the skills being measured and the rating criteria being applied,
- 2) designing tasks with a clear idea of what is being measured and what constitutes poor and good performance,
- 3) developing scoring guides that are clear and specific about how to apply the criteria to the student work,
- 4) providing sufficient training for teachers to learn how to apply the criteria to real examples of student work, and
- 5) monitoring the scoring process through moderation and auditing to maintain calibration over time.

Uses of Technology in Scoring

In the International Baccalaureate program, which operates in 125 countries, teachers receive papers to score via computer delivery, and they calibrate their scoring to common benchmarks through an online training process that evaluates their ability to score accurately. The teachers upload their scored papers to be further evaluated or audited, as needed, and to have the scores recorded. Similarly, in Hong Kong, most delivery and scoring of open-ended assessments is becoming computer-based, as it is in 20 other provinces of China. There, as in many other places, double scoring is used to ensure reliability, with a third scorer called in if there are discrepancies. In the U.S., teachers and teacher educators who score the edTPA portfolio, used for teacher licensure, receive training and calibration via a computer-based program and do their scoring of portfolios online as well.

More recently, automated scoring procedures have also been developed to score both short and long constructed-response items. Automated scoring has been used successfully in contexts ranging from state end-of courses exams to the Collegiate Learning Assessment⁴⁹ and NAEP — in both the Math Online project that required students to provide explanations of their mathematical reasoning and the NAEP simulation study that required students to use search queries.⁵⁰ In the NAEP study that used physics simulations, the agreement between human raters and computer ratings in a cross-validation study was 96

percent. In the more complex, extended CLA task, correlations of human and computer ratings are nearly as high, at 86 percent.⁵¹

As these innovations have demonstrated, technological advances are beginning to enable highly reliable computer-based scoring of complex student responses. Coupled with appropriate use of human scoring to help produce the data for developing a scoring algorithm, to check on its reliability, and to score outlier responses that cannot be evaluated by machine, this technology can also enhance the feasibility of performance assessments.

Teacher Involvement in Scoring

As noted above, human scoring is needed even when technology can help support some aspects of scoring for performance tasks. Many commercial testing companies send open-ended responses to individuals hired to score who may not be teachers. But some systems in the U.S. and abroad rely on teachers for scoring, which provides additional benefits for instructional quality. Researchers have found that involving teachers in scoring performance assessments is powerful professional development because it connects teacher learning directly to their examination of student learning, and gives them the opportunity to think together about how to improve that learning.⁵² It also sends an important message by signaling that teachers can be active participants in shaping the direction of school change. As this kind of professional development acknowledges the critical role of teachers in supporting students' learning, it put teachers in their rightful place — center stage in the school improvement process.

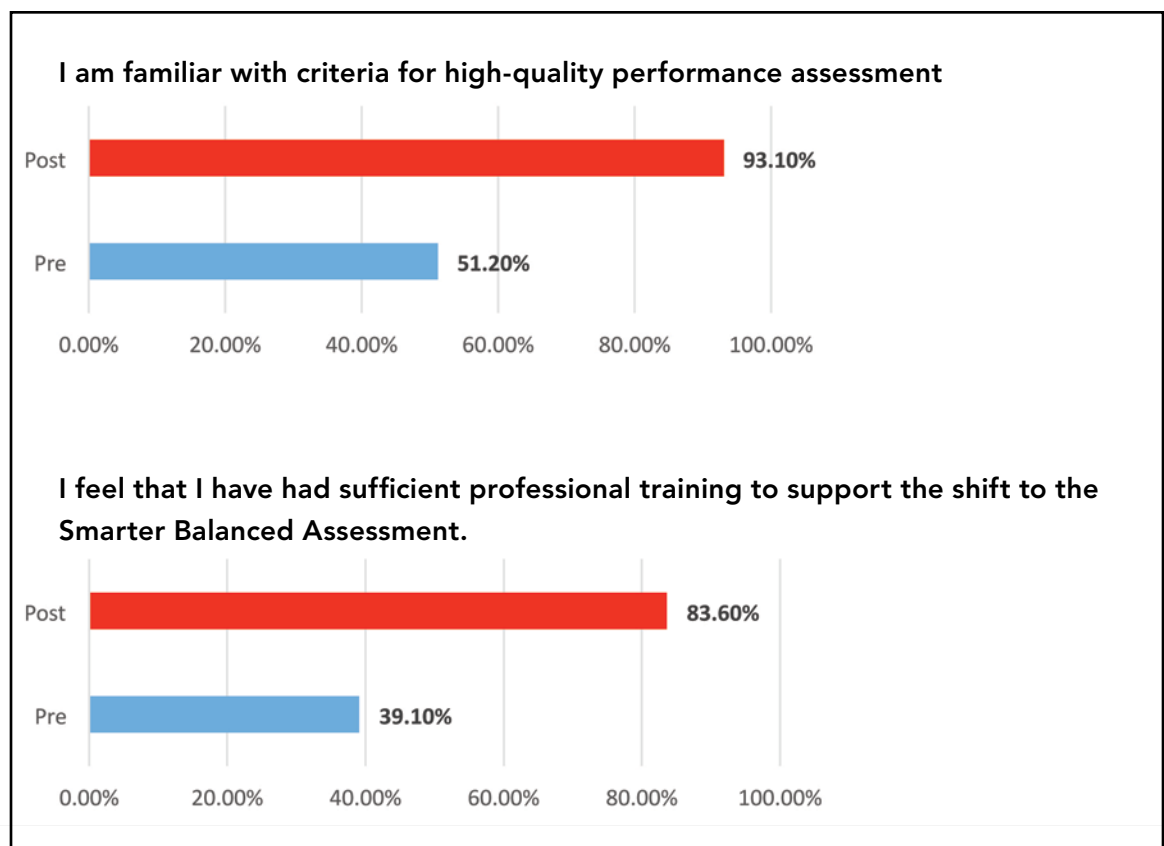
Where school systems have devoted resources to assessment at the classroom level and have invested in classroom-based performance assessors, teachers have developed deep expertise that translates into shared judgments and common mental models of what constitutes acceptable student performance on complex types of learning. Furthermore, when teachers become experienced in developing and evaluating high quality performance assessments, they are more able to design and deliver high quality learning experiences because they have a stronger understanding of what kinds of tasks elicit thoughtful work, how students think as they complete such tasks, and what a quality standard looks like.

These outcomes were recently illustrated in a project launched in 2015 by SCALE and WestEd, which engaged teachers in three states — California, New Hampshire, and Oregon — in scoring the performance tasks from the Smarter Balanced assessments used in those states. The Building Educator Assessment Literacy (BEAL) project, which continues to offer scoring sessions

as a professional development opportunity, seeks to build teacher capacity and knowledge of the new standards and of assessment practices.⁵³ Teachers learn to score student work and reflect on the implications of the tasks, the student work, and the scoring experience for their own instructional practice.

Teachers were emphatic about how valuable this scoring and reflection experience was for their own learning. Across the three states, 97 percent said that the training “deepened my understanding of the assessment system;” 96 percent said it “helped me think about ways to enact curriculum-embedded performance assessment with my students;” and 88 percent said that the scoring process “deepened my understanding of the Common Core State Standards.”

The proportion who agreed they were familiar with criteria for high-quality performance assessment increased from 51 percent to 93 percent, and the proportion who felt they had sufficient training to support the shift to the Smarter Balanced assessment more than doubled, from 39 percent to 87 percent.



Their comments stressed the value of the professional development and its influence on their teaching:

This was probably the most productive professional development I have attended in my 13 years of teaching. I think it would be great to offer it again and involve more districts if possible.

This experience has dramatically impacted my future instruction.

... looking at student work will reveal the gaps and guide the shifts that need to be made in the classroom. Hand scoring a writing task is like opening a student's brain and getting a more intimate perspective on the thinking and learning. There is much to be learned from these comprehensive summative performance tasks.

Many were very specific about the instructional shifts they would make. For example:

This is invaluable to seeing how the rubric criteria translates into a student response, the many different acceptable ways students can respond, and see areas where instruction could be strengthened such as in developing explanations.

Being aware of how items are scored gives me a better idea of the kinds of tasks students will be asked to do and the level of complexity. This will help me to select appropriately rigorous enough tasks. My teaching focus will be primarily on the thinking process and use of information to solve problems.

...teachers could begin to analyze their instruction as it pertains to offering students multiple opportunities to reason, explain their reasoning, and thinking about how assumptions and answers to one part of a question can and does impact other portions. Also, the idea that one needs to consider "what is reasonable" when answering a question and be able to logically defend that decision.

I will be more intentional about classroom discourse and assure my students are doing real problems that push their mathematics to the deeper thinking level.

These comments reflect those of teachers scoring performance assessments in many other contexts. One teacher remarked after a performance assessment scoring session:

We are moving in the right direction as an education system! I am very excited and rejuvenated as an educator after the drill and kill years of NCLB. I can finally teach real skills students will use.

CONCLUSION

Because performance assessments model worthwhile tasks and expectations, embed assessment into the curriculum, and develop teachers' understanding of how to interpret and respond to student learning, their use typically improves instruction. Learning is also strengthened as students are able to work on these assessment tasks intensively, revise them to meet standards, and display their learning to parents, peers, teachers, and even future professors and employers. Both teachers and their students gain insights into how students learn in the specific content area and how, as a team, they can facilitate improvements in this learning. Meanwhile, state and district policymakers are able to track progress and trends as scores from these measures are aggregated, reported, and analyzed. Thus, when states assess performance authentically and engage teachers in the scoring, they generate positive instructional impact as well as leverage on productive accountability.

As described in this report, states can choose among several models for integrating performance assessments into their state systems. Building on models that have been developed, studied, and refined, it is possible to achieve the policy benefits of comparable assessments, reliably scored along with the learning benefits that come from engaging students and teachers in rich tasks that inform the teaching and learning process.

APPENDIX A: NEW YORK PERFORMANCE STANDARDS CONSORTIUM SCIENCE RUBRIC

New York Performance Standards Consortium Student _____

Extended Science Project or Original Experiment Title of Experiment _____

Circle one: Teacher or External Evaluator _____ Date _____

Circle one: Holistic evaluation _____ Signature _____

03/11

Performance Indicator	Outstanding	Good	Competent	Needs Revision
Contextualize	Background research has been thoroughly conducted using at least two original sources. Sources are all appropriately cited. The significance of the problem is clearly stated. The hypotheses/theses are grounded in the background research.	Background research has been thoroughly conducted. Sources are appropriately cited. The significance of the problem is stated. The hypotheses/theses are relevant to the background research.	Background research is included in the introduction. Sources are cited. The significance of the problem is stated. The hypotheses/theses are clearly stated.	Background research is not included in the introduction. Sources are not cited. The significance of the problem is not stated. The hypotheses/theses are not stated.
Critique Experimental Design	Identifies, describes and controls all relevant variables. Thoughtfully evaluates the procedure and/or set up. Clearly describes bias in the design.	Identifies, describes and controls most relevant variables. Evaluates the procedure and/or set up. Clearly describes bias in the design.	Identifies, describes and controls some relevant variables. Evaluates the procedure and/or set up. Attempts to describe bias in the design.	Does not identify, describe or control any variables. Does not evaluate the procedure and/or set up. Does not attempt to describe bias in the design.
Collect, Organize and Present Data	Collects data in a reliable and valid manner. Presents relevant data that is consistent with the problem. Generates appropriate tables, charts and graphs with data and makes appropriate calculations. Conducts thorough mathematical analysis of the data.	Collects data in a reliable and valid manner. Presents relevant data that is consistent with the problem. Generates appropriate tables, charts and graphs with data and/or makes appropriate calculations. Conducts mathematical analysis of the data.	Collects data in a reliable and valid manner. Presents data that is consistent with the problem. Generates tables, charts and graphs with data. Conducts analysis of the data.	Collects data in a non-reliable and/or invalid manner. Does not present data or presents data that is not relevant to the problem. Does not generate tables, charts and graphs. Does not analyze the data.
Analyze and Interpret Results	Draws thoughtful conclusions that are supported by the data. Relates conclusions to original question. Thoroughly describes sources of error and their effects on the data.	Draws conclusions that are supported by the data. Relates conclusions to original question. Describes several sources of error and their effects on the data.	Draws conclusions that are partially supported by the data. Attempts to relate conclusions to original question. Describes sources of error and attempts to describe their effects on the data.	Draws no conclusions or draws conclusions that are not supported by the data. Does not attempt to relate conclusions to original question. Does not describe sources of error or does not attempt to describe their effects on the data.

Performance Indicator	Outstanding	Good	Competent	Needs Revision
Revise Original Design	Proposes effective and relevant revisions for the experimental plan to lessen the effects of bias and sources of error. Poses thoughtful and relevant questions for future research.	Proposes relevant revisions for the experimental plan to lessen the effects of bias and sources of error. Poses relevant questions for future research.	Proposes revisions for the experimental plan to lessen the effects of bias and sources of error. Poses questions for future research.	Does not propose revisions for the experimental plan. Does not pose questions for future research.
Defense (for oral component only)	Thoroughly answers questions relevant to the experiment and related topics.	Adequately answers questions relevant to the experiment and related topics.	Adequately answers questions relevant to the experiment.	Does not adequately answer questions relevant to the experiment.

ENDNOTES

- 1Every Student Succeeds Act, Section 1111(b)(2)(B)(vi) and Section 1111(b)(2)(J)).
- 2Every Student Succeeds Act, Section 1111(b)(2)(B)(viii)(II).
- 3Every Student Succeeds Act, Section 1204.
- 4Yuan, K., & Le, V. (2012). *Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests*. Santa Monica, CA: RAND Corporation.
- 5National Academy of Sciences. (2010.) *Rising above the gathering storm, revisited*. Washington, DC: National Academies Press.
- 6Conley, D.T. (2005). *College knowledge: What it really takes for students to succeed and what we can do to get them ready*. San Francisco: Jossey-Bass.
- 7Madaus, G.F. & O'Dwyer, L.M. (1999). A short history of performance assessment. *Phi Delta Kappan* (May), pp. 688-695.
- 8American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association. p. 137.
- 9Bloom, B. (1956). *Taxonomy of educational objectives. Handbook 1: Cognitive domain*. White Plains, NY: Longman.
- 10The Assessment Continuum was developed by L. Darling-Hammond (2013) for the Stanford Center for Assessment, Learning, and Equity (SCALE), Stanford, CA.
- 11Yuan & Le. (2012).
- 12 Every Student Succeeds Act, Section 1111(b)(2)(B)(v)): "(II) in the case of science, [tests must be] administered not less than one time during—(aa) grades 3 through 5; (bb) grades 6 through 9; and (cc) grades 10 through 12."
- 13 Shyer, C. (August 2009). *Regents examinations and Regents competency tests*. Retrieved February 19, 2017, from www.p12.nysed.gov/assessment/08-09memo/jun-aug-09/724/563-809.pdf.
- 14 Pecheone, R. & Kahl, S. (2014). Where we are now: Lessons learned and emerging directions. In L. Darling-Hammond, L. & F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning*, 53-92. San Francisco: Jossey-Bass.
- 15 Herman, J.L. & Linn, R.L. (2013). *On the road to assessing deeper learning: The status of Smarter Balanced and PARCC assessment consortia*. (CRESST Report 823). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- 16Herman, J.L. & Linn, R.L. (2013). *On the road to assessing deeper learning: The status of Smarter Balanced and PARCC assessment consortia*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- 17 Stanford Center for Assessment, Learning, & Equity & WestEd (forthcoming). *Final Report for the Building Educator Assessment Literacy Project* to the William and Flora Hewlett Foundation.
- 18 Collegiate Learning Assessment. (2010). *CLA: Returning to learning*. Retrieved February 26, 2017, from <http://www.collegiatelearningassessment.org/>; Klein, S., Benjamin, R.,

- Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review*, 31(5), 415–439.
- 19 Clyman, S. G., Clauser, B. E., & Melnick, D. E. (1995). Computer-based case simulations. In E. L. Mancall & P. G. Bashook (Eds.), *Assessing clinical reasoning: The oral examination and alternative methods* (pp. 139-149). Evanston, IL: American Board of Medical Specialties.
- 20 Bennett, R.E., Persky, H., Weiss, A.R., & Jenkins, F. (2007). Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project (NCES 2007-466). Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved February 19, 2017, from <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466>, p. 41.
- 21 Bennett, R. et al. (2007). p. 46.
- 22 <https://www.performanceassessmentresourcebank.org/>
- 23 These include schools working with the Center for Collaborative Education in Boston, the New York Performance Standards Consortium, the Internationals High School Network, New Tech High Schools, Envision Schools, the Met Schools, and others.
- 24 <https://www.performanceassessmentresourcebank.org/>
- 25 Koretz, D., Klein, S.P., McCaffrey, Daniel F. and Stecher, Brian M. (1994). *Interim report, the reliability of Vermont portfolio scores in the 1992-93 school year*. Santa Monica, CA: RAND Corporation. Retrieved February 19, 2017, from <http://www.rand.org/pubs/reprints/RP260>; Koretz, D., Klein, S., McCaffrey, D., & Stecher, B. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, 13(3), 5-10.
- 26 Vermont Department of Education. (n.d.). *Core principles of high-quality local assessment systems*. Retrieved February 19, 2017, from <https://pl.scribd.com/document/72026194/Core-Principles-08>
- 27 Pinckney, E., & Taylor, G. (2006). Standards and assessment memorandum, p.1. Vermont Department of Education. Retrieved February 15, 2017, from http://education.vermont.gov/new/pdfdoc/pgm_curriculum/local_assessment/assessment_guidance_030106.pdf.
- 28 M. Hock, Vermont Department of Education, personal communication, September 17, 2009.
- 29 Vermont Department of Education. (n.d.)
- 30 Measured Progress. (2009). *Commonwealth accountability and testing system: 2007-08 technical report*. Version 1.2, p. 92. Commonwealth of Kentucky Department of Education. Retrieved February 20, 2010 from <http://www.education.ky.gov/KDE/Administrative+Resources/Testing+and+Reporting+/Kentucky+School+Testing+System/Accountability+System/Technical+Manual+2008.htm>.
- 31 <https://advancesinap.collegeboard.org/ap-capstone>
- 32 http://apcentral.collegeboard.com/apc/members/exam/exam_information/226194.html
- 33 RIDE. (2005). The Rhode Island High School Diploma System. Retrieved February 19, 2017, from <http://www.aypf.org/documents/HSDiplomaPDF.pdf>.
- 34 RIDE. (2005). The Rhode Island High School Diploma System, p. 5. Retrieved February 19, 2017, from <http://www.aypf.org/documents/HSDiplomaPDF.pdf>.
- 35 Darling-Hammond, L. & Wentworth, L. (2014). Reaching out: International benchmarks for performance assessment. In L. Darling-Hammond and F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning* (93-130). San Francisco: Jossey-Bass.

- 36 For a summary of this research, see Darling-Hammond, L. (2014). *Next generation assessment: Moving beyond the bubble test to support 21st century learning*. San Francisco: Jossey-Bass, 2014.
- 37 Stein, M.K. & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, 2(1), 50-80; Stone, C.A. & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education*, 16(1), 1-26; Newmann, F.M., Marks, H.M., & Gamoran, A. (1996). Authentic pedagogy and student performance. *American Journal of Education*, 104(8), 280-312. Parke, C.S., Lane, S., & Stone, C.A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation*, 12(3), 239-269; Stone, C.A. & Lane, S. (2003); Linn, R.L., Baker, E.L., & Betebenner, D.W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, 31(6), 3-16.
- 38 Lane, S. (2014). Performance assessment: The state of the art. In L. Darling-Hammond, L. & F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning* (pp. 133-184). San Francisco: Jossey-Bass.
- 39 Barron, B., Schwartz, D.L., Vye, N.J., Moore, A., Petrosino, T., Zech, L., & Bransford, D. (1998). Doing with understanding: Lessons from research on problem and project-based-learning. *Journal of Learning Sciences*, 7(3&4), 271-311.
- 40 Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80, 139-148.
- 41 New Hampshire State Department of Education. (2016, December 5). *New Hampshire Performance Assessment of Competency Education, (PACE): Progress report to the United States Department of Education*. Concord, New Hampshire: Author.
- 42 Lane, S., & Stone, C.A. (2006). Performance assessments. In B. Brennan (Ed.), *Educational measurement*. Westport, CT: American Council on Education and Praeger.
- 43 Baker, E.L. (2007). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. *Educational Assessment*, 12(3&4), 179-194.
- 44 Lane & Stone. (2006).
- 45 Lane, S. (2014).
- 46 Chi, M., Glaser, R., & Farr, M.J. (Eds.). (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum; Ericsson, K.A. & Simon, H.A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.
- 47 Kentucky Department of Education. (1997). *KIRIS accountability cycle 2 technical manual*. Retrieved from contractor files Measured Progress: Technical report. Dover, NH: Author.
- 48 Measured Progress. (2009). *Commonwealth accountability and testing system: 2010-11 technical report*. Version 1.0, p. 74-75. Retrieved February 20, 2010 from education.ky.gov/aa/kts/documents/2010-11%20kcct%20tech%20rep%20fin.pdf
- 49 Collegiate Learning Assessment. (2010). *CLA: Returning to learning*. Retrieved March 15, 2010, from <http://www.collegiatelearningassessment.org/>.
- 50 Bennett, R.E., Persky, H., Weiss, A.R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project (NCES 2007-466)*. Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved February 20, 2017, from <http://nces.ed.gov/pubsearch/pubsinfo>.

[asp?pubid=2007466](#); Deane, P. (2006). Strategies for evidence identification through linguistic assessment of textual responses. In D.M. Williamson, R.J. Mislevy, & I.I. Bejar (Eds.). *Automated scoring of complex tasks in computer-based testing* (pp. 313-362). Mahwah, NJ: Lawrence Erlbaum Associates.

51 Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review*, 31(5), 415–439.

52 Darling-Hammond, L. & Falk, B. (2015). Supporting teacher learning through performance assessment. In L. Darling-Hammond & F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning* (pp. 277-310). San Francisco: Jossey-Bass.

53 See Daro, V. & Wei, R.C. (2015, June 29). How can teachers learn deeply? By scoring student assessments. *Education Week*. Retrieved February 26, 2017, from http://blogs.edweek.org/edweek/learning_deeply/2015/06/how_can_teachers_learn_deeply_by_scoring_student_assessments.html.



One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
voice: 202.336.7000 | fax: 202.408.8072