



Through a Looking Glass: Lessons Learned and Future Directions for Performance Assessment

Raymond L. Pecheone and Stuart Kahl

With Assistance from Jillian Hamma and Ann Jaquith

This study was conducted by the Stanford Center for Opportunity Policy in Education (SCOPE) with support from the Ford Foundation and the Nellie Mae Education Foundation.

© 2010 Stanford Center for Opportunity Policy in Education. All rights reserved.

The Stanford Center for Opportunity Policy in Education (SCOPE) supports cross-disciplinary research, policy analysis, and practice that address issues of educational opportunity, access, equity, and diversity in the United States and internationally.

Citation: Pecheone, R., Kahl, S., Hamma, J., Jaquith, A. (2010). *Through a Looking Glass: Lessons Learned and Future Directions for Performance Assessment*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Stanford Center for Opportunity Policy in Education

Barnum Center, 505 Lasuen Mall
Stanford, California 94305
Phone: 650.725.8600
scope@stanford.edu
<http://edpolicy.stanford.edu>



Table of Contents

Introduction	1
Building on Current Approaches to Standards-Based State Performance Assessment.....	4
Lessons Learned from Past State Accountability Systems That Included	
Performance Measures	15
Promising and Emerging Assessment Practices	25
One Approach to Development of Next-Generation of State Assessment	31
Conclusions.....	41
References	42
Appendix.....	46

Introduction

Over the last decade the growing momentum to rethink test-driven accountability practices has been the focus of intense debate. Disappointment about performance of U.S. students on international tests, concern as to the nation's global competitiveness, and questions about our students' readiness to enter college and the workforce have all led to another wave of efforts to significantly reform American education.

A recurring theme in the public debate among educators, business leaders, elected officials, and community members is the need for schools to focus on a new and expanded skill set in order for American students to compete in a digital age. The discourse centers on the need to measure the core knowledge and higher-order skills critical to postsecondary learning and career success. In particular, growing emphasis on critical thinking, analytical reasoning, and communication skills has led to calls for a more balanced assessment system that includes authentic measures of student performance.

This chapter describes efforts by states to use performance assessment in large-scale state accountability systems and highlights promising practices that can a basis for broadening how the nation approaches accountability testing. The states highlighted here (see Table 1) present a window into assessment practices currently in place that can help shape the development of the next generation of assessment in this country. These pioneering efforts offer insight into the challenges and opportunities of using performance measures within the context of state assessment policy.

Background and Challenges

The nature of large-scale assessments has a significant impact on the attitudes, behaviors, and practices of students and teachers (Shepard, 2002; Wood et al., 2007; Coe et al., 1994). Research in the early 1990s showed that reliance on multiple-choice tests in a high-stakes environment can have a negative effect on instruction by reducing the complexity of task demands and the opportunities for students to develop and demonstrate certain thinking and performance skills (Cizek, 2001; Wilson, 2004; Conley, 2010; Flexer, 1991; Hiebert, 1991; Koretz, Linn, Dunbar, & Shepard, 1991; Madaus et al., 1992; Shepard, et al., 1995). In fact, the multiple-choice and constructed-response forms of a question often tap different skills. For example, there is an important difference between actually solving a quadratic equation and using the lower-level, pre-algebra skill of substituting answer options in the equation to identify the correct answer. Likewise, there is a difference between drawing and justifying one's own conclusions after reading a passage and picking the best conclusion from a set of four multiple-choice options. One of many lessons we've learned during the age of high-stakes statewide testing is that what gets tested is what gets taught.

The approach to schooling that currently dominates precollege programs has existed for well over a century. Although this model may have served us when the nation was at the height of industrialization, the model falls short when it comes to preparing students for postsecondary programs and the 21st-century workplace (NASBE, 2009; Schleicher, 2009; Alliance for Excellent Education, 2009; NCEE, 2007). Many high school students are bored in school and not motivated to learn (Quaglia Institute, 2008), resulting in a disturbing high school and college dropout rates—a leading indicator of why educational approaches and testing practices should be reformed. Furthermore, many high-stakes, statewide accountability programs currently use assessment instruments reminiscent of the minimal competency and basic skills tests employed 25 years ago (Tucker, 2009). Given that what gets tested is what gets taught, students are neither required nor offered the opportunity to demonstrate 21st-century skills (critical thinking, problem solving, communication)—not on tests or in class (Wood, Darling-Hammond, Neill, & Roschewski, 2007; Shepard, 2002).

Concern about American students' low level of engagement, as well as high school graduates' apparent lack of 21st-century skills, has led to heightened interest in curriculum-embedded performance assessment, an approach to assessment that many believe is better suited to measuring these higher-order skills (Wood et al., 2007; Tucker, 2009). For the purposes of this chapter, performance measures are defined as an opportunity for students to show how they can apply their knowledge and skills in disciplinary and interdisciplinary tasks focused on key aspects of academic learning. When people speak of performance assessment today in the context of 21st-century skills, they are often referring to more substantial activities—either short-term, on-demand tasks or curriculum-embedded, project-based tasks that yield reliable and valid scores. The most common example of such performance assessment in education is a directed writing assessment—administration of writing “prompts”—that requires students to produce essays or other forms of extended student writing. Other scorable products or performances could include responses to constructed-response questions following some activity, research reports, oral presentations, and debates.

Large-scale performance assessment is not new. In the late 1980s, dissatisfaction with nonsecure, off-the-shelf tests not designed for evaluation of school programs led states to undertake customized, statewide testing programs (Hamilton, Stecher, & Yuan, 2008; Kahl, 2008). Heavily influenced by curriculum experts, many of these programs had nonmultiple-choice components, including constructed-response items, performance tasks, and portfolios. These performance components were considered “authentic assessments” in that they were intended to engage students in “real-world” activities that they might encounter outside of school (Wiggins, 1998).

In states with the greatest emphasis on authentic assessment, teachers made extensive use of released constructed-response questions and performance tasks (Khattari, Kane, & Reeve, 1995; Koretz Barron, Mitchell, & Stecher, 1996; Coe et al., 1994). Through professional development they learned the value of evaluating actual student work for

informing instructional practice. Teachers gained the competency to use (and even develop) scoring rubrics. Supplementary curriculum companies supported these efforts by producing materials that addressed higher-order thinking skills.

The days of large-scale authentic assessments were short-lived, however. In the late 1990s, the U.S. Department of Education stepped up its efforts to enforce Title I assessment and accountability requirements, and the next reauthorization of ESEA (the No Child Left Behind Act or NCLB, enacted in January 2002) added even more teeth to the law by requiring every-child, every-year testing in reading and mathematics. The expense of testing at all the required grades, and the turnaround time for results necessary to accommodate a parental choice option added to the law, led many states to rely almost exclusively on multiple-choice test items.

It is important to ensure that the educational reforms of the future advance the cause of improved educational practice and raise standards of performance that can lead to assurance that all students are college- and career-ready. It is anticipated that in designing the next generation of assessment performance assessments will play a vital role in measuring the higher-order skills that are cited as critical to college and career success (Conley, 2010).

Building on Current Approaches to Standards-Based State Performance Assessment

Although the influence of NCLB reduced the emphasis on authentic assessment in many states, a number of states continue assessing students' learning through performance tasks and constructed-response items. This section profiles promising assessment practices that are part of current state accountability systems, which include student performance components—that is, measures affording the opportunity for students to show how they can apply their knowledge and skills in disciplinary and interdisciplinary tasks focused on standards of academic learning, in particular critical thinking. This review covers state performance assessments used for a range of purposes, from offering alternative approaches to high-stakes state assessments to formative assessments designed to improve instructional practice and student learning. Performance assessment practices in states can include on-demand, constructed-response (CR) items; real-world, classroom-embedded performance tasks; and collection of student work through portfolios.

Additionally, performance tasks can be complex projects spanning days or weeks to complete, such as science experiments; student-designed, disciplinary research inquiries; and assembly and interpretation of evidence about an historical question. A detailed description of current state assessment practices is highlighted below to illustrate the role performance assessments play in operating high-stakes state accountability systems. Specifically, this section highlights secondary education because it is the pathway to college and career success and represents the primary focus of the national debate on rethinking accountability practices.

Table 1 briefly profiles accountability practices in selected states that have maintained performance-based components in their state assessment systems. Specific states are highlighted to establish a national baseline for informing development of the next generation of assessment and accountability.

New York State

New York has a 135-year history of state-level assessment that includes both on-demand and performance tasks (New York State Education Department, 1987, p. 18). To earn a diploma in New York, in addition to completing course credit requirements students must pass commencement-level Regents Examinations in comprehensive English, global history and geography, U.S. history and government, mathematics, and science. Different cut scores on these syllabus-based, end-of-course tests are used for a local diploma and a state Regents diploma. Alternative assessments, approved by the state, can also be used for these diplomas. Additionally, the Regents have put in place a local diploma option that allows development of equivalent academic tasks, often part of a portfolio-based system, that can be substituted for the Regents Exam. All local options must be reviewed and approved by the state department of education. For example,

Table 1: Exemplars of State-Based Uses of Performance Assessment

Assessments (Percentages are Based on Number of Items)	Assessment Graduation Requirements
Connecticut	
<p>Connecticut Academic Performance Test (CAPT)</p> <ul style="list-style-type: none"> • Math (grade 10): 25% constructed response (CR), 75% multiple choice (MC) • Science (grade 10): 20% CR, 92% MC • Reading for information (grade 10): 33% CR ended, 67% MC • Response to literature (grade 10): 100% open-ended • Writing (grade 10): 70% essays, 30% MC 	<p>CAPT results must be included in district-generated graduation requirements. Generally, districts mandate a score of at least “proficient” (level 3 of 5) on the writing and mathematics assessments at a minimum.</p>
Kentucky	
<p>Kentucky Core Content Test (KCCT)</p> <ul style="list-style-type: none"> • Reading (grade 10): 6% CR, 94% MC • Mathematics (grade 11): 8% CR, 92% MC • Science (grade 11): 8% CR, 92% MC • Social studies (grade 11): 8% CR, 92% MC • On-demand writing (grade 12): 100% CR <p>Writing Portfolio (grade 12): 4 pieces developed over years</p>	<p>Students must pass the core content areas and meet standards on the writing portfolio. Note: the writing portfolio requirement is currently being phased out and will be replaced in future state assessments.</p>
New Jersey	
<p>High School Proficiency Assessment (HSPA)</p> <ul style="list-style-type: none"> • Language arts literacy (grade 11): variety of MC, CR, and performance-based tasks, including speaking • Mathematics (grade 11): 17% CR, 83% MC <p>End-of-Course Examinations</p> <ul style="list-style-type: none"> • Biology/life science: approx. 6% CR, 94% MC • Performance assessment prompt field-tested in 2008 and 2009 • Algebra I: 4% CR, 11% short answer, 85% MC <p>Special Review Assessment (SRA)</p> <ul style="list-style-type: none"> • Performance Assessment Tasks (PATs) completed in the area(s) in which student did not pass the HSPA 	<p>Students must pass either HSPA or SRA in both language arts literacy and mathematics.</p>
New York	
<p>Regents Examinations (end-of-course assessments)</p> <ul style="list-style-type: none"> • Comprehensive English: essay and MC; number varies • Global history and geography: essay CR, MC • U.S. history and government: essay, CR, MC • Mathematics B: 41% CR, 49% MC ('09) • Mathematics A, integrated algebra: 23% CR, 77% MC ('09) • Geometry: 26% CR, 74% MC ('09) • Biology: 33% CR, 67% MC ('09) • Chemistry: 38% CR, 62% MC ('09) • Earth science: performance-based assessment and written test 41% CR, 59% MC ('09) • Languages other than English (French, German, Hebrew, Italian, Latin, Spanish): speaking, CR, MC 	<p>Students must pass commencement-level Regents Examinations with a score of at least 55–64 to qualify for a local diploma or 65 for a Regents diploma in:</p> <ol style="list-style-type: none"> (1) Comprehensive English (2) Mathematics (3) Global history and geography (4) U.S. history and government (5) Science

Assessments (Percentages are Based on Number of Items)	Assessment Graduation Requirements
Rhode Island	
<p>New England Common Assessments Program (NECAP)</p> <ul style="list-style-type: none"> • Writing (grade 11): essay questions • Reading (grade 11): 14–20% constructed response, 80–85% MC ('08) • Mathematics (grade 11): 11% constructed response, 44% short answer, 44% MC • Districts must include in their local assessment system a combination of at least two of the following: graduation portfolios, exhibitions, comprehensive course assessments, or a combination thereof 	<p>Each district determines proficiency-based graduation requirements in the six core academic areas. NECAP exams must count as 1/3 of their total assessment in English and mathematics. The other measures must include at least two additional performance-based diploma assessments.</p>
Vermont	
<p>New England Common Assessments Program (NECAP)</p> <ul style="list-style-type: none"> • Writing (grade 11): essay questions • Reading (grade 11): 14–20% constructed response, 80–85% MC ('08) • Mathematics (grade 11): 11% constructed response, 44% short answer, 44% MC • Science (grade 11): inquiry task and exams 11% constructed responses, 89% MC 	<p>Student meets graduation requirements if the school board determines student has (1) met the framework or comparable standards as measured by results on performance-based assessments or (2) completed at least 20 Carnegie units, or any combination of (1) and (2) that demonstrates the student has attained the standards.</p>

the New York Performance Standards Consortium, a group of 27 secondary schools, has received a state-approved waiver allowing their students to complete a graduation portfolio in lieu of the Regents Exams. This portfolio includes a set of ambitious performance tasks—a scientific investigation, a mathematical model, a literary essay, a history/social science research paper, an arts demonstration, and a reflection on a community service or internship experience—that meet a set of common standards and are scored through social moderation processes using common scoring rubrics.

Performance-based components of the Regents Examinations include a variety of tasks. 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" (Shyer, 2009, p. 3). In history and social studies, students complete essays about document-based questions that require analysis of a set of documents and artifacts. The Regents Science Examination includes a laboratory performance test completed near the end of the course and a written test with a large number of open-ended questions (Shyer, 2009, p. 14).

Generally, at least two teachers must independently rate all Regents Examinations that lead to a Regents diploma, except mathematics, which requires at least three scorers. All teachers rate exams according to the scoring key and rubrics provided by the department of education, which have directions for scoring multiple-choice and constructed-response

questions and, if applicable, guidelines for rating the essay or performance components (Office of State Assessment, 2008). Teachers are trained to score all extended writing tasks using benchmark performances and rubrics (University of the State of New York State Education Department, 2009a, 2009b). (See New York assessment tasks in the Appendix.)

New England Common Assessment Program (NECAP)

The New England states of New Hampshire, Rhode Island, and Vermont have formed an unprecedented state collaboration around development and design of common reference examinations, which high school students in some NECAP states must pass to earn a high school diploma. In those states, along with other specific individual state requirements, all students must pass the common NECAP assessment administered in the 11th grade in reading, writing, mathematics, and science (Measured Progress, 2009). Science is tested in the spring over three sessions (New England Common Assessment Program, 2009), while the other subjects are tested over two sessions (Measured Progress, 2009). The NECAP is a hybrid assessment comprising both multiple-choice and constructed-response items. Performance-based components of NECAP entail a range of standards-based constructed-response (CR) items and performance tasks. In writing, students respond to two writing prompts that are scored using a common set of rubrics. In science, both CR and performance items are included. The science CR items “require students to respond to a question by using words, pictures, diagrams, charts, or tables to fully explain their response,” and an inquiry task asks students “to hypothesize, plan, and critique [scientific] investigations, analyze data, and develop explanations” (New England Common Assessment Program, 2009, pp.1, 15). All NECAP scorers are trained and go through a calibration process prior to scoring. The writing prompt is “scored by two independent readers both on quality of the stylistic and rhetorical aspects of the writing and on the use of standard English conventions” (Measured Progress, 2009, p. 8). The other CR answers are scored using an item-specific rubric with score point descriptions (Measured Progress, 2009). Common cut scores were established through representative expert committees in the NECAP states to ensure comparability across states. Professional development materials to support the NECAP assessment were developed by content specialists at the New Hampshire, Rhode Island, and Vermont departments of education in partnership with the Education Development Center and the National Center for the Improvement of Educational Assessment.

Through this collaboration, the NECAP states were able to maintain common content standards for the region; this fosters sharing of instructional and curriculum resources within and across state borders. In addition, NECAP states significantly lowered the cost of assessment while maintaining high standards of quality; these cost savings enabled the NECAP consortium to develop a more balanced assessment, including performance-based, constructed-response items, which would have strained the testing budgets of the individual states. Collaboration among states in test design and adminis-

tration could become a model for the country in developing innovative assessments that meet high standards of test quality, measure a broad range of skills and abilities, and are administratively feasible and cost-effective. By pooling resources, states are better able to afford development of richer performance measures designed to address the skills and abilities needed to be college- and workplace-ready in the 21st century. Finally, state-based consortia can promote and support development of “regional learning networks,” which enable teachers and administrators to share promising practices and resources across states. (See New Hampshire, Rhode Island, and Vermont assessment tasks in the Appendix.)

Vermont Local Comprehensive Assessment System

Vermont was an early pioneer in using embedded classroom assessments for accountability and to guide curriculum development. As a result of NCLB requirements, these assessments became part of Vermont’s School Quality Standards mandate, requiring each school to develop a local comprehensive assessment system “aligned with the Vermont Framework of Standards and Learning Opportunities and is consistent with the Vermont Comprehensive Assessment System, adopted by the State Board of Education in November 1996” (Vermont State Board of Education, 2000).

Each school’s local comprehensive assessment system must assess students in the required standards not covered by the state assessment (Vermont State Board of Education, 2000). There are no requirements stated for the type of assessment to be used, although the materials, items, and tasks supplied by the state for optional use are predominantly performance-based. Additionally, the department of education reviews district-based assessment systems and gives specific guidance to teachers and other educators responsible for scoring common assessments (M. Hock, personal communication, September 17, 2009). 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” (Vermont Department of Education, n.d., p.5). The state furnishes a variety of assessment tools that schools may use in developing their local comprehensive assessment system. For example, in the content areas of math and writing the state offers benchmarks, rubrics, calibration materials, and data analysis tools, to effectively use math and writing portfolios as local classroom assessments. According to the deputy commissioner of education and the director of standards and assessment, the local assessment provisions of the school quality standards are intended to place “classroom assessment at the core of the assessment system—common grade, team, school, and state assessments would round out the Local Comprehensive Assessment System” (Pinckney & Taylor, 2006, p. 1).

Although the NECAP assessment is used as the primary pathway for Vermont students to earn a diploma, they also can earn a diploma through meeting the requirements of a

performance-based local option. A student “meets the requirements for graduation if, at the discretion of each secondary school board: ‘The student demonstrates that he or she has attained or exceeded the standards contained in the Framework or comparable standards as measured by results on performance-based assessments’” (Vermont Department of Education, 2006, 2120.8.7[a]).

Over almost two decades, Vermont’s leadership in performance assessment has created a collaborative professional culture around curriculum and instruction that engages teachers in principled discussions about the quality of student work. Here is how Richard Murnane, a Harvard professor, vividly describes the conversation of Vermont teachers who come together to score student portfolios: “Often heated, the discussion focused on what constitutes good communication and problem solving skills, how first-rate work differs from less adequate work, and what types of problems elicit the best student work” (Murnane & Levy, 1996, p. 263).

Formal school-based structures designed to bring teachers together to discuss student work not only serve to deepen teacher knowledge of student skills and abilities but can change how professional development is practiced in schools and districts. Teacher-led discussions of student work are often cited as the best and most consequential professional development that can lead to higher student achievement (Darling-Hammond, Wei, Andree, Richardson, & Orphanos, 2009). (See Vermont assessment tasks in Appendix.)

Maine and New Hampshire

Like Vermont, Maine and New Hampshire established policies that include and encourage using performance assessment in conjunction with their large-scale state accountability systems. Both states have put in place assessment policies requiring that high school graduation decisions *cannot* be based solely on state high school examinations. Rather, assessment policies must be used in conjunction with other performance measures, among them curriculum-embedded performance tasks, portfolios, and other locally determined graduation indicators. In Maine, local assessments are organized around the state’s learning results in eight areas: English, mathematics, science, social studies, health and physical education, career preparation, visual and performing arts, and world language. The state offers extensive professional development to local districts in developing common performance tasks, rubrics, portfolios, and exhibitions of student work. Note that Maine has recently joined the NECAP consortia and is using the NECAP assessment as the state accountability measure. New Hampshire passed legislation to develop a competency-based system for graduation that no longer relies on Carnegie units beginning in 2008–09 (New Hampshire code of administrative rules—education, 2005). The competency-based system uses a “mastery of learning” approach that will rely on course-based performance assessments to earn high school credits both in and out of school, rather than Carnegie units. (See New Hampshire, Rhode Island, and Vermont assessment tasks in the Appendix.)

Connecticut

Connecticut has used authentic assessments for more than a decade to design the state-mandated high school examination. The Connecticut Academic Performance Test (CAPT) is administered in grade 10 in reading, writing, math, and science (Connecticut State Board of Education, 2009). By scoring student results relative to established state goals in each content area, CAPT is designed to measure progress toward the educational goals reflected in the Connecticut curriculum frameworks. Moreover, by statute CAPT must be included as one indicator of performance to support a graduation decision but cannot be used as the sole criterion for graduation. Specifically, by statute CAPT scores must be combined with other “measures of successful course completion” (2009, p. 26).

The CAPT assessment was designed as a balanced assessment with multiple-choice, constructed-response and curriculum-embedded performance tasks to assess student content knowledge. Performance-based components include a variety of item types and tasks:

- *Reading.* Reading scores are split between the reading for information and response to literature subtests. Reading for information “assesses a student’s ability to independently read, thoroughly comprehend, and thoughtfully respond to three authentic nonfiction texts” (Connecticut State Department of Education, 2006, p. 11). All constructed responses are scored by trained scorers who have met calibration standards using a 0-to-2-point rubric (2006).
- *Response to literature.* The response to literature subtest assesses students on their ability to “independently read, thoroughly comprehend, and thoughtfully respond to one authentic fictional text through four constructed-response questions” (Connecticut State Department of Education, 2006, p. 8). Two independent readers score each of the four written responses on a 6-point rubric (2006).
- *Interdisciplinary writing.* The writing assessment is administered in two 65-minute sessions. During each session, students must take a position on a stated contemporary issue, possibly in the form of a letter or editorial, and cite given sources as support for their argument. Each response is scored holistically. Two independent readers score responses using a 6-point scale (Connecticut State Department of Education, 2006).
- *Mathematics.* The mathematics assessment uses both multiple-choice and constructed-response items. The constructed-response items are weighted more heavily across the tests; therefore half the total points draw from these performance-based questions (Connecticut State Board of Education, 2009).

- *Curriculum-embedded science tasks.* In the science assessment, two real-world tasks are intended as components to be embedded in the science teachers' course curriculum. Specifically, a laboratory activity and a science, technology, and society (STS) investigation are given to schools in each of the science content strands for grades 9 and 10. Assessment tasks are aligned to state standards and curriculum frameworks and then designed to be embedded in the science curriculum. Students are required to complete these tasks in class; they are asked to formulate a hypothesis, conduct experiments, analyze data, and write a lab report to demonstrate their ability to engage in scientific reasoning. For the on-demand component of the CAPT science assessment, the specific scientific skills and processes needed to complete the embedded assessment are independently tested through use of constructed-response items aligned to the locally embedded performance tasks.

Using this testing methodology presents an innovative approach to high-stakes assessment in science using both curriculum-embedded performance tasks scored by classroom teachers and an on-demand assessment of student knowledge using a constructed-response methodology to independently measure student learning (Connecticut State Department of Education, 2010). [AU: see query in ref-list entry; is date known?] This assessment approach takes advantage of the power of performance assessment to transform classroom practice, and the need to ensure that measures of student learning are comparable and objective. Given recent development of national common core standards in English language arts and mathematics, Connecticut's CAPT assessment can serve as one "proof point" of what an assessment looks like that is designed to predict college and career success. To examine the impact of student performance on CAPT and its relationship to college and workplace success, Connecticut funded a major study tracking five cohorts of 10th grade students between 1996 and 2000 over eight years beyond high school. The study found that students scoring higher on CAPT were more likely to attend and graduate from college, and it showed a positive relationship between CAPT and workplace success (Coelen, Rende, & Fulton, 2008). (See Connecticut assessment tasks in Appendix.)

Rhode Island

To earn a high school diploma in Rhode Island, all students are required to demonstrate proficiency on both the NECAP and a locally developed school-based portfolio. Student portfolios for graduation must include a "composite measure of each student's overall proficiency for graduation in the six core academic areas" locally developed and validated in each district (Rhode Island Board of Regents for Elementary and Secondary Education, 2008, L-6-3.3). Student results on the NECAP examinations count as one-third of the components of their total assessment in English, mathematics, and science "as designated by the Board of Regents" and include "at least two additional performance-based diploma assessments" in other subject areas (2008, L-6-3.3).

Districts must include in their local assessment system “a combination of at least two of the following performance-based assessments: graduation portfolios, exhibitions, comprehensive course assessments, or a combination thereof, such as a Certificate of Initial Mastery” (Rhode Island Board of Regents for Elementary and Secondary Education, 2008, L-6-3.2). Schools must develop a review process to score performance-based diploma assessments at the local level. For exhibitions, portfolios, or Certificates of Initial Mastery to be considered part of the schoolwide diploma assessment, schools must meet state requirements such as supplying “sufficient evidence” and “using valid and reliable rubrics and/or an independent review process.” Each entry in a portfolio “should be evaluated using valid and reliable rubrics and/or a review process” (Rhode Island Department of Education & Education Alliance at Brown University, 2005, p. 2). Teachers involved in portfolio scoring must be trained and meet calibration standards on the rubric in order to reliably score student work.

To ensure “opportunity to learn,” the state department guidelines require that “[e]xisting course offerings must now give students frequent opportunities to practice applying their skills and knowledge” (Rhode Island Department of Education, 2005, p. 4). These guidelines ensure that courses prepare students “for the more formal demonstrations of proficiencies necessary to earn a diploma. Naturally, high school courses will also continue to administer routine assessments such as tests, quizzes, papers, labs and so forth” (Rhode Island Department of Education, 2005, p. 4).

Rhode Island is breaking new ground in developing a balanced assessment system that takes into account school-based portfolio assessments, in combination with an on-demand standardized assessment (NECAP). In addition, use of a judgmental weighting system to combine information from standardized assessment (NECAP) and portfolio scores illustrates one approach that states might consider to develop a composite score employing both performance and standardized test data to support a graduation decision.

Currently Rhode Island is the only state using portfolio-based performance data as a central part of a high-stakes state accountability system. The Rhode Island approach puts teachers at the heart of the assessment process and teacher scoring as the basis for judgments of student learning. Rhode Island’s use of a school-based portfolio system is rooted in a longstanding tradition of local control, but this approach to assessment also raises a number of psychometric concerns (for example, when teachers score their own students’ work, their assessment of student learning may be biased by factors other than the construct tested). Therefore, issues of reliability and validity are challenging and complex, especially in a portfolio-based system that can vary from school to school and across districts. Questions about the reliability and comparability of student performance across schools and districts must be addressed and resolved if portfolio assessment is to be used as part of a high-stakes state accountability system. (See New Hampshire, Rhode Island, and Vermont assessment tasks in Appendix.)

New Jersey

New Jersey is in transition as it develops a high school accountability system. Currently all students must pass the High School Proficiency Assessment (HSPA) or an approved alternative in order to graduate (New Jersey Department of Education, 2005). The state is moving toward more specific end-of-course content exams. Nevertheless, it has by statute developed an alternative pathway to graduation through use of a performance-based assessment system. If students do not pass the HSPA in March of their junior year during the first testing, they can take the assessment again in October and March of their senior year. Additionally, students failing to meet state standards must also begin remediation instruction in preparing for the Special Review Assessment (SRA). All students, if eligible, can take the SRA in the fall of their senior year (New Jersey Department of Education, 2008a).

The SRA is an “individually, locally administered, untimed, state-developed, locally scored assessment” (New Jersey Department of Education, 2008b, p. 7). Students must participate in a school-designed SRA instructional program for the content area(s) in which they did not meet proficiency on the HSPA prior to being administered SRA Performance Assessment Tasks (PATs). The SRA Instructional Program is continued until the SRA teacher decides that students can be successful on a PAT (2008b).

For content areas in which students do not score at least 200 on the HSPA, students must successfully complete two PATs in each content area cluster/standard. Language arts literacy has two clusters, while mathematics has four standards. Selection of PATs “is based solely on the results of the student’s first HSPA administration” (New Jersey Department of Education, 2008b, p. 6); “If a student is not successful on a specific PAT, additional PATs may be administered until the student successfully completes the required number of PATs for that content area” (p. 7). In addition, to earn a diploma “Students with disabilities who are in grade 11 . . . must participate in the HSPA or the APA [Alternate Proficiency Assessment]” (p. 12). “The Individualized Education Program (IEP) team for each student determines which assessment (HSPA or APA) the student will take for each content area addressed” and “must also determine if the student who is taking the HSPA in one or both content areas will be required to pass the HSPA in those content areas in order to graduate” (pp. 12–13).

New Jersey has been a leading state in developing and validating alternative assessment pathways to graduation. The state, through developing the SRA and APA, allows all students (including special needs students) alternative pathways to obtain a high school diploma. More importantly, the SRA alternatives offer diverse learners greater access to college through embedded performance measures that assess academic progress by using testing formats more sensitive to various learning modalities. As a result of implementing these policies, New Jersey reports one of the highest graduation rates

(83%) in this country, comprising all racial and ethnic groups. In spite of these results, concern over the reliability and comparability of the assessment results within and across districts has prompted a formal state review. Based on the review process, recommendations will be presented and adopted that serve to significantly strengthen the reliability and validity of these alternative graduation measures. (See New Jersey assessment tasks in the Appendix.)

Lessons Learned from Past State Accountability Systems That Included Performance Measures

There is much to be learned from states that include performance assessments in their accountability systems. The lessons described here are intended to guide development and implementation of statewide performance assessment components of high quality and utility. Although many states engaged in performance testing to some degree in the 1990s, the Commonwealth of Kentucky probably went the farthest in taking such testing to scale. Its program included multiple forms of performance assessment: extended constructed-response questions, “hands-on” performance tasks, and portfolios. Given Measured Progress’ contractual association with the Kentucky assessment, this discussion primarily pertains to lessons learned from that state’s efforts.

In 1990 the Commonwealth Supreme Court declared Kentucky’s educational system unconstitutional on the grounds of equity, resulting in passage of the Kentucky Educational Reform Act (KERA). KERA led to the creation of an assessment and accountability program, the Kentucky Instructional Resources Information System (KIRIS), which was required to be “predominantly performance-based.”

The initial KIRIS program had on-demand tests in seven subjects, administration of hands-on performance events (tasks) in six subjects, and portfolios in writing and mathematics. All of these were required at three grade levels, although on-demand tests for the “off” grades were offered to schools for voluntary use. At first, the high-stakes, on-demand tests at three grade levels employed multiple-choice and constructed-response items. When it was determined that the constructed-response items by themselves yielded acceptable test reliability, the multiple-choice items were not counted in the school accountability index and were dropped from the tests for a few years. (They were ultimately reinstated as new test designs evolved.)

The performance event component of the test lasted three years. The mathematics portfolios never counted toward accountability results. However, even after several redesigns of the program, the writing portfolio component survived until the recent budget crisis led to the decision to continue writing portfolios only voluntarily and not count them as part of the state accountability system. In the 2009–10 school year, the on-demand component is called the Kentucky Core Content Test. The remaining assessment components are on-demand tests in reading, mathematics, science, and social studies (with a 50/50 weighting of multiple-choice and constructed-response results) and on-demand writing. More detail on the performance and portfolio components of the original KIRIS program appears in the discussion of lessons learned, below.

Quality and Feasibility of Taking Performance Assessment to Scale

As suggested previously, in the late 1980s and early 1990s state assessment programs took the lead in developing and implementing performance assessment in pre-college academic disciplines. At the start of the so-called authentic assessment period, though, the emphasis of many of these programs was on group (school or statewide) results. For example, Connecticut and Massachusetts administered performance tasks in selected grades to samples of a few hundred students (Badger, Thomas, & McCormack, 1990; Kahl, Abeles, & Baron, 1985). Trained task administrators (local or otherwise) followed detailed instructions and used pre-packaged kits of materials. These two efforts were one-time probes, not intended for accountability. The sampling designs did not enable reporting of local results. Instead, reports focused on what the activities revealed about student understanding and on instructional implications. Thus the findings were reported in much the same way as those of pre-1980 NAEP.

In Kentucky, all students in three grades participated in performance testing during the first three years of the program. This component was discontinued because of difficulty in finding a viable approach to equating performance events across years. Trained administrators carried kits of materials to the schools and conducted testing one class at a time, with teams of three or four students working on different tasks during the same 50-minute session. In this case, students worked together for the first part of the period but individually in the latter part to produce unique, individual, scorable products that were returned to the assessment contractor for central scoring. With students taking un-equated tasks, no attempt was made to report student-level results; instead, results were aggregated and reported at the school level. Because of the high quality of the tasks and scoring, the state was able to count the results of the performance testing toward the high-stakes accountability index for every school.

Vermont and Kentucky had statewide portfolio assessment programs in writing and mathematics. In both cases, although there were specifications for the types of work samples to be permitted in the individual student portfolios, task development and selection was left to the teachers. In the small state of Vermont teachers came together to conduct central scoring; in Kentucky, teachers 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 some additional scorer training provided to teachers on a limited basis. These portfolio assessments, successful in many ways, were an attempt to make performance assessment feasible on a large scale by placing a great deal more responsibility for various aspects of the process in the hands of local educators. Many other states engaged in similar performance and portfolio assessment efforts, but the Kentucky and Vermont programs are sufficient to illustrate key points.

The efforts of the authentic assessment period taught us many lessons, many of which are explained in the next few sections. With an eye toward including performance assessments as a component of accountability measures for both individual and group re-

sults, the lessons present ways to strike a balance between local and centralized responsibilities to ensure high quality of products (the tasks or projects, associated materials, and measurement attributes), at the same time attending to the feasibility, in terms of costs in time and money, of taking performance assessment to scale.

Task Quality

Traditional state assessments typically emerge from a development process that involves multiple reviews by curriculum and assessment experts, field testing, psychometric analyses, and further review and revision. But several programs with performance components rely on local development and selection of tasks. As expected, there is varying quality in locally developed tasks. The next sections examine task quality in terms of a few critical attributes.

Alignment to Standards. One major problem with locally developed tasks was that they weren't always closely tied to standards. In all fairness, many states did not have the kind of content standards in the early 1990s that Title I mandated a few years later; many had only general curriculum guidelines. Kentucky did have content standards, but they were brand new and, more importantly, new to the local educators. Local curricula were also not aligned to the content standards. This is especially important in a high-stakes environment. It is generally agreed (and affirmed by the courts) that it is inappropriate to assess students on concepts and skills—and attach negative consequences to poor performance relative to those concepts and skills—if the students do not have an opportunity to learn and perform them. In the early 1990s, New Jersey conducted three years of “due notice testing” in conjunction with its HSPA program. This gave schools time to adjust curricula to new standards and get used to new tests before high stakes were associated with results from the testing. Clearly, a state's responsibility in this regard is to make it clear to schools via content standards just what students should know and be able to do. Of course, current work at the national level on developing common state standards will help states address this responsibility.

For performance assessment, including portfolio assessment, to count toward high-stakes accountability, there must be some central control over or monitoring of the quality of the tasks the students are asked to tackle. One critical attribute is that tasks need to be closely tied to standards. For this attribute and others discussed later, teacher training is essential.

Scorable Products. With the exception of writing portfolios, performance assessments were as unfamiliar to local Kentucky educators as the state's new standards were. The Kentucky writing portfolios were of sufficient quality to be counted toward the accountability index because teachers could easily find ways to get students to generate text. Even though writing topics for Kentucky portfolios were never standardized or common across students, the specifications for entries apparently were adequate to ensure a degree of data comparability that justified continuing to use the approach, up to the

present day (for almost two decades), as documented in the technical reports of the program (Kentucky Department of Education, 1997).

The teachers were not as successful at identifying activities in mathematics that led to rich, scorable products. As a result, the mathematics portfolios never counted toward state assessment results. Lack of quality assessments in mathematics was not a problem unique to Kentucky; it was a challenge in other states (Massachusetts, Vermont) undertaking portfolio assessment.

Unlike the assignments leading to portfolio entries in Kentucky, which were left to teachers to devise, the hands-on performance tasks administered by trained administrators sent into the Kentucky schools were developed through the same joint efforts of department and contractor staff and advisory committees that produced the items for the paper-and-pencil tests. Another advantage the Kentucky on-demand performance events had over writing portfolios was that the student products for them truly represented individual student work. By contrast, the portfolio component allowed use of the “writing process,” involving collaborative efforts of students and teachers in revision of student work. Although this practice is often encountered in nontesting, real-world settings, this is not always the case. If the assessment is used for accountability purposes, it is important to collect and evaluate evidence of an individual student’s capability.

Lesson: Another critical requirement for performance assessments for high-stakes, statewide programs is the need for tasks to yield rich, scorable products (closely tied to standards) that yield credible evidence of learning and represent the full range of individual student capabilities.

Classroom Management and Resources. Almost 30 years ago, a performance assessment project in the United Kingdom involved trained administrators going into the schools to administer performance tasks (Burstall, 1986; Burstall, Baron, & Stiggins, 1987). Students worked in pairs on the tasks, from the belief that a one-on-one situation might be intimidating to students whereas working with a peer would be more comfortable even in the presence of an adult outsider. The same approach was used in performance testing projects in Connecticut and Massachusetts during the 1980s, both of which were small-scale studies involving only a limited sample of students.

Kentucky’s performance testing was conducted on a much larger scale. The initiative required all students in three grades to participate. The approach still involved external administrators carrying kits of materials into the schools and supervising an entire classroom of students working in teams of three or four at stations where different tasks were set up. The groupings were not for the comfort of the students, but rather to make the job manageable for the administrators. Nevertheless, this component of Kentucky’s assessment program was expensive and labor-intensive for the state. Furthermore, it was burdensome for the schools. Also, because the tasks were not necessarily related to the content being taught at the time, the efforts were of little immediate instructional

value. Needless to say, the material and personnel needs of the programs described here posed challenges in terms of time, logistics, and expense. Some current-day, statewide alternate assessment programs for students with disabilities continue to have such burdensome material needs, though those programs test only a small percentage of the student population.

Lessons: Statewide performance testing of the general population of students using external administrators and supplying specialized materials is quite expensive. Cost efficiencies can be realized if (1) tasks are administered by local school personnel and (2) required materials are readily available in the schools, in homes, or online. Whether performing on-demand tasks or longer projects, if students are allowed to work in teams they should be required to produce individual scorable products, so there is no question whose work is being scored. (Working in teams for at least part of a task or project also gives the students an opportunity to demonstrate some noncognitive 21st-century skills, such as leadership, collaboration, and flexibility.) Finally, to justify the burden on teachers and students, it would be best if the tasks were curriculum-embedded. That is, they should be relevant, instructionally sensitive, and syllabus-based.

Technical (Measurement) Quality

Issues of technical quality—whether real or perceived—contributed to the demise of the authentic assessment movement more than a decade ago. The challenges of performance assessment were identified and exposed, but little effort was made to work through the technical problems of reliability and validity. Given the demands of continuous development of new tests, coordination and administration of tests statewide, and analysis and reporting, the assessment and accountability staff in the state departments of education and their contractors had enough on their plates with little time to publish and disseminate information beyond the requirements of the programs and contracts (such as evidence of high technical quality when it existed). Unsubstantiated criticism colored by biases was all too common and often left unanswered. For example, this was clearly the case with respect to human scoring, which is relied on heavily in performance assessment (see the following discussion on scoring accuracy).

Scoring Accuracy and Reliability. Some outspoken critics of human scoring of student work argue that the process is subjective and focused on values and attitudes (Schlafly, 2001). Yet the process states typically use to score student responses (to constructed-response questions, for example) has many elements designed to render it objective. Scorers do not know student names or schools. Using pre-established rubrics describing the characteristics of work earning each point value, scorers are really just being asked to categorize responses. The rubrics are developed with the tasks and both are field tested, and then improved, if necessary. Scorers generally have to have a background in the relevant subject area; they are trained on each test question using

the rubric and many samples of student work of varying quality; they have to qualify (calibrate) by achieving a certain level of accuracy on “qualifying packs” of student responses before being allowed to score for record; and even though they are scoring for the record, various approaches to blind double scoring are used to monitor their accuracy, with corrective action taken when necessary. All these practices are described in “Operational Best Practices,” a document produced jointly by state testing directors and testing company experts at the request of former Secretary of Education Margaret Spellings (Association of Test Publishers and Council of Chief State School Officers, 2010). That these best practices yield reliable results is well documented by state testing technical manuals.

There is a mistaken, but all too common belief, that tests requiring human scoring are inherently unreliable. True, human scorers evaluating complex student performance are not perfect; raters’ scores often vary by plus or minus one point. Nevertheless, students can be credited with partial knowledge on the basis of performance, which is more accurate than receiving full credit for guessing correctly on a multiple-choice question. The same level of test reliability can be achieved with 8 to 10 4-point constructed-response questions as with 50 multiple-choice items. This is a matter of fact, documented in the technical manuals associated with hundreds of state tests.

Some have touted the merits of using “artificial intelligence”—scoring student work by computer. This technique has proven useful in scoring writing samples, but it is still experimental. More evidence of reliability and validity is needed to establish its use in scoring in other content areas, and for large-scale state assessment.

The same techniques for human scoring of student responses to constructed-response questions and writing prompts are applicable to any of the scorable products resulting from performance tasks or projects. Consequently, when it comes to scoring performance assessments, there is expertise and experience. Again, the matter of taking it to scale is an issue of time and expense. A method of dealing with this problem is addressed in the next section.

Lesson: For limited measures, such as a writing sample (a one-item test), new strategies are needed to strengthen reliability analyses. An alternate solution, of course, is to use more measures; to accomplish this, multiple curriculum-embedded tasks could be the least intrusive. Moreover, for more extensive performance measures, traditional approaches to demonstrating test reliability are as appropriate as they are for all-multiple-choice tests. Also, it is not enough to create and use high-quality measures; it is also necessary to convincingly demonstrate to a variety of audiences that they are indeed of high quality, both in terms of what they measure and how reliably they measure it.

Local Scoring with Central Auditing. It has already been argued that, to justify the time and effort local educators would have to devote to performance assessments for statewide accountability, it would be best if assessments were curriculum-embedded. This would also reduce the costs associated with administering isolated tasks statewide using external administrators transporting cartons of materials to every school. The same situation applies in scoring the resulting student work. The logistics of transporting student work (which can take many forms, not just writing) for central scoring, and the scoring itself (which would be in addition to scoring that might already be necessary for on-demand constructed-response testing), would be time-consuming and expensive.

The logical solution is to have local educators score their students' work with the state (on a sampling basis) auditing the local scoring. Kentucky did this with its writing portfolio assessment. At the end of the second year of assessment, audit results showed that the scores submitted by some schools were inappropriately high. (The reason for this is explained in the next section, on the score scale.) 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 (Kentucky Department of Education, 1997). 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.

Lesson: Consistency of local scoring across schools, and therefore comparability of results, can be accomplished if states make a commitment to teacher training, as well as an audit process and associated “remediation” to yield scoring that is consistent across a whole state. There may be some audit sampling and feedback approaches that can significantly cut down on the need for remediation. For example, use of interim measures (e.g., curriculum-embedded performance assessments) throughout the year, long before accountability results must be reported, allows time for feedback to local scorers. For each performance task, schools could be asked to submit the work of just a few students for central audit scoring, and the scores from that process, reported back to the schools, could be used as benchmarks to anchor the scoring of the work of the rest of the students. This practice is consistent with the guidance in the “Operational Best Practices” document regarding real-time monitoring of scoring accuracy. Also, phasing in performance assessment components of larger assessment systems could allow time not only for a state to refine and improve audit procedures but also for local educators to internalize the state's general standards of performance.

Score Scales. Scorers of writing portfolios in the 1990s Kentucky assessment program assigned a score of 1, 2, 3, or 4 to each portfolio—scores that corresponded to the four performance levels used in the state for all subject areas. This practice was a mistake, for a number of reasons. The idea of a portfolio is to accumulate a larger body of evidence of a student’s capabilities. (Seven writing samples were initially required in every student’s portfolio.) Reducing a portfolio to a number from one to four so early in the process negated the advantages of a bigger score scale and in fact reduced the measurement value of portfolios to that of a single constructed-response item.

The Kentucky portfolio scoring approach is also what led to inflated scoring in some schools, which the audit process identified, as described in the previous section. Teachers knew that student writing had improved, but the only way they could show improvement was by assigning higher scores to the work. With so few score points, each point corresponded to a wide range of quality, and for many students their improvement was not enough to “cross the line” to the next performance level. This problem was verified in many audited schools. (Interestingly, the more limited, on-demand performance events administered in Kentucky had much larger score ranges.)

Lesson: Although no particular score range is optimal, a wide (rather than narrow) range of possible scores on a task or project would be desirable to allow more fine-grained distinctions to be made and growth or gains to be more appropriately noted. Of course, for a given task or project, several products or performances could contribute to that range (e.g., a writing sample, an oral presentation, a model, and even responses to follow-up questions). “Collapsing” the score points into fewer score ranges can then be done later for purposes of performance-level reporting and even equating. Also, different measures related to a task or project might be counted toward different subject area scores.

Equating. Kentucky discontinued the on-demand performance task component after three years because of the challenges associated with equating the tasks across years. (The program required approximately 12 unique tasks each year at each of three grade levels. Within the 12, only two or three pertained to the same subject area.) The state’s equating approach required identification of out-of-state schools willing to administer tasks from two consecutive years to their students, with each student taking one task from each of the two years. The difficulty in finding appropriate samples of students in large enough numbers, which ultimately affected the quality of the equating, coupled with opposition to performance assessment of some factions within the state led to the decision to drop this component of the program.

Consideration was given to equating the performance component of the program through the traditional, on-demand, group-administered component, that is, treating the performance tasks like new, constructed-response questions scaled with reused, constructed-response, “equating” questions. This idea was rejected, though, because

there was no reason to believe that in successive years the extent of improvement (change) in performance on the two components would be the same.

Lesson: Equating performance tasks that are more substantial than typical, on-demand, constructed-response questions is particularly challenging because the standard approach of administering both old and new tasks to the same students can be overly burdensome on teachers and students, and therefore often unfeasible. Alternative approaches should be developed and considered. For example, even though it's not ideal, pre-equating is a procedure frequently used for direct writing assessments—often one-item tests. Prompts are selected for successive years that produced similar score distributions in field testing. The same approach could be used for performance testing within a subject area, provided a large enough sample of students participate in the field testing of each task or project. Of course, with a much larger score scale, similar distributions might be difficult, but different tasks could have their 20-to-30-point score ranges collapsed to fewer score points (say, 10) that do exhibit similar distributions.

Validity. Validity theory is centered around claims about the appropriateness of the interpretation of data in relationship to student performance on a test or performance task (Cronbach, 1971; Frederiksen & Collins, 1996; Mislevy, Steinberg & Almond, 2003; Pellegrino, Chudowsky, & Glaser, 2001; Moss, Girard, & Haniford, 2006). According to Campbell and Stanley (1963), to claim test validity means that evidence obtained from the assessment provides support for interpretation of the evidence to the extent that the interpretation is stronger than any other alternative explanation (e.g., internal validity). Applying this paradigm to performance assessment focuses on the scorer collecting and presenting the evidence used to make judgments about the knowledge and skills the student exhibits. Teachers or expert scorers who have been trained and calibrated to score consistently using a scoring protocol or rubric often make judgments rooted in cognitive theory about student learning (Bransford & Darling-Hammond, 2005). According to Frederiksen and White, “by adopting systematic approaches to observing and analyzing performance, teachers can accurately and consistently code key features of student work.” They further hypothesize that “carrying out accurate and complete analyses of student work will enable the teachers to develop the evidence needed to make accurate scoring judgments of the quality of student work in reference to state curricular standards.” This chain of validity evidence to support curriculum-embedded work requires clear understanding of the skills to be measured and rich description of the performance to support interpretation of a student’s meta-cognitive ability to carry out complex performance tasks. Additionally, teachers can use this chain of evidence to inform their practice by evaluating the impact of specific instructional strategies on student learning, including subgroup analyses (Moss et al., 2006).

Further, to meet the new demands of the common core of learning (all students college- and career-ready), it can be argued that performance assessment measures of higher-

order thinking need to be a core component of the next generation of assessment and accountability in this country. Performance tasks should be designed to engage students to think critically, reason analytically, solve problems, and communicate effectively. To achieve these goals requires a more balanced view of assessment and accountability that includes both formative information—how students develop and access learning resources to complete challenging tasks—and summative judgments of student learning, based on performance tasks that are aligned to national standards and that can be used for district and state accountability purposes. Moreover, these meta-cognitive tasks are situated in the learning process in classes and therefore change the nature of the validity argument supporting use and interpretation of embedded performance tasks. A broader conception of a validity argument is suggested, beginning with a detailed description of the constructs being measured that takes into account multiple types of evidence at different levels of the scale and that is sensitive to the dynamic interaction between the student and the task as the act of inquiry continuously shapes student learning.

Teachers and students need to know the learning demands of the tasks the students are expected to master, and teachers need to create instructional opportunities for students to successfully complete the tasks. This interaction between standards and tasks in the classroom is based on teachers and students developing common understandings of the skills that will be measured and a clear description of the performance indicators to interpret the students' performance on the basis of standards, student work samples, benchmarks, and rubrics. To this end, the goal of creating more transparency in assessment is to signal to the students the knowledge and skills necessary for success in the classroom. In addition, students often engaged in self-assessment and peer assessment of student work cause a significant shift in how we think about test administration and validation. That is, using standardized administration standards for on-demand tests to ensure the “objectivity and comparability” of assessment is confounded when students have the opportunity to collaborate on the performance tasks. Research suggests that opportunities for peer collaboration coupled with formative feedback to students are a leading indicator of student knowledge of the subject matter and a strong predictor of future success (Black & Wiliam 1998, Bransford & Swartz, 2001).

Finally, as discussed earlier, performance assessment data go well beyond supplying scores. They are designed to inform students and teachers about what it is important to learn, what learning looks like, and how learning is shaped by the context of the learning environment and its learners (Engeström, 1999). As a result, broadening the conception of validity to address richer and more complex performance tasks requires considering how assessment functions in various instructional and school-based contexts and how the learner is influenced (shaped) by the learning environment.

Promising and Emerging Assessment Practices

here are many innovative assessment projects currently under way and worth noting. This section highlights three such initiatives that focus on higher-order thinking skills and are designed to predict college and workplace readiness: the College and Work Readiness Assessment (CWRA), the College Readiness and Performance Assessment System (C-PAS), and the Ohio Performance Assessment Pilot Project (OPAPP). These projects illustrate three related, but different, approaches to measuring higher-order thinking by (1) using on-demand, computer-adapted, constructed-response items (CWRA); (2) assessing cognitive strategies that enable college-bound students to learn, understand, retain, use, and apply content from a range of disciplines (C-PAS); and (3) developing course and curriculum-embedded rich projects in the core academic disciplines that can be combined with high-stakes state accountability test measures to support a high school graduation decision (OPAPP).

College and Work Readiness Assessment (CWRA)

The CWRA is a high-school-senior version of the Collegiate Learning Assessment (CLA) (Klein et al., 2007; Klein, Freedman, Shavelson, & Bolus, 2008; Shavelson, 2007, 2010). The CLA and CWRA were developed at the Council for Aid to Education by Roger Benjamin, Steve Klein, and Richard Shavelson; see Chapter 4 in Shavelson (2010) for a brief history.

The CLA was developed to measure undergraduates' learning—in particular their ability to think critically, reason analytically, solve problems, and communicate clearly. The assessment comprises performance tasks and critical writing components. The performance task component presents students with a real-world problem and an in-basket of information and asks them to either solve the problem or recommend a course of action based on the furnished evidence. The analytic writing prompts ask students either to take a position on a topic or to critique an argument.

A 90-minute, entirely constructed-response exam, the CLA is delivered over the Internet. The assessment focuses on performance at the institution level or on performance at the program level within an institution. Institution or program-level scores are reported, both as observed performance and as value added beyond what would be expected from entering student SAT scores. The CLA differs substantially—in terms of its philosophical and theoretical underpinnings—from most learning assessments, which are based on an empiricist philosophy and a psychometric/behavioral tradition. From this stance, everyday complex tasks are divided into components, and each component is analyzed to identify the abilities required for successful performance. For example, suppose that components such as critical thinking, problem solving, analytic reasoning, and written communication are identified. A separate measure of each component would then be constructed and students would take each test. At the end of testing,

student scores would be added up to construct a total score to describe their performance—not only on the assessment at hand, but also generalizing to a universe of complex tasks similar to those the tests were intended to measure.

The CLA is based on a combination of rationalist and socio-historical philosophies in the cognitive-constructivist and situated-in-context traditions. The CLA's conceptual underpinnings are embodied in what has been called a *criterion sampling* approach to measurement. This approach assumes that the whole is greater than the sum of its parts and that complex tasks require an integration of abilities that cannot be captured if divided into and measured as individual components. The criterion-sampling notion is straightforward: If you want to know what a person knows and can do, sample tasks from the domain in which she is to act, observe her performance, and infer competence and learning. For example, if you want to know whether a person not only knows the laws that govern driving a car but can also actually drive a car, don't just give her a multiple-choice test. Rather, also administer a driving test with a sample of tasks from the general driving domain (starting the car, pulling into traffic, turning right and left in traffic, backing up, parking). On the basis of this sample of performance, it is possible to draw more general, valid inferences about driving performance.

The CLA follows the criterion-sampling approach by defining a domain of real-world tasks that are holistic and drawn from life situations. It samples tasks and collects student *operant responses*. Operant responses are student-generated responses modified with feedback as the task is carried out. These responses parallel those of the CWRA.

The CWRA's twofold mission is to improve teaching and learning by using performance tasks to connect classroom practice with authentic institutional assessment and to evaluate student readiness to do college work. The CWRA employs certain performance tasks from the CLA to test high school seniors' critical thinking, analytic reasoning, problem solving, and communication abilities. To date, 49 high schools have participated in the CWRA. Currently in development are means to provide peer group comparisons and on-demand testing. Efforts are also under way to reduce the required testing time for a performance task from 90 minutes to 60 minutes, adjust the reading level to make it appropriate for a range of high school students, and administer the adapted, open-ended performance tasks in conjunction with other standardized tests of critical thinking to produce reliable individual student scores.

It is crucial to work with teachers to offer a model for how curricular and pedagogical interventions help students to develop these higher-order skills. CWRA offers "Performance Task Academies" for hands-on training grounded in the literature on learning theory, critical thinking, and authentic assessment. The academies focus on showing teachers how to create classroom projects that are a hybrid of case studies and performance-based learning, with a special focus on higher-order skills. The Performance Tasks Library is a teacher-created teaching resource.

The CWRA program has conducted preliminary research on using the assessment. Research shows that high school GPA combined with scores on CWRA performance tasks are as good as high school GPA combined with SAT scores for predicting a student's cumulative college GPA. (See Collegiate Learning Assessment tasks in the Appendix.)

The College-Readiness Performance Assessment System (C-PAS)

C-PAS was developed by David T. Conley, chief executive officer of the Education Policy and Improvement Center at the University of Oregon. It is specifically designed to create new methods to evaluate the college readiness of high school students, on the premise that current assessments and tests do not necessarily do a good job of gauging student cognitive capabilities and ability to apply strategies they will be expected to demonstrate in entry-level college courses and beyond. As American secondary school classrooms attempt to become more data-driven, and as more high school students set college as their goal, it is critical that teachers have the right set of data to enable them to make instructional decisions that prepare their students for postsecondary education, and that students have a clearer picture of their readiness for college courses. C-PAS is being designed specifically to supply this type of information to teachers and students to ensure high school instruction leads to college readiness for all students.

C-PAS is a series of curriculum-embedded performance tasks that teachers administer within the context of their curriculum and score with a common scoring guide, resulting in a performance profile for each task composed of scores from up to five key cognitive skills. The teacher separately grades the task for inclusion as a component in the course grade, thereby increasing student engagement in the task. C-PAS scores are useful to teachers in considering how well their curriculum is helping students reason, solve problems, interpret information, conduct research, and generate work with precision and accuracy. The tasks are carefully designed to encourage student development of key cognitive strategies that research identifies as being important elements of entry-level college courses.

C-PAS is deeply rooted in psychometric principles and practices in order to achieve a high degree of technical adequacy, which helps ensure that the scores generated are valid and accurate indicators of student development of key cognitive strategies associated with college success. This is achieved in a number of ways. The five key cognitive strategies are carefully analyzed using item response theory to determine the degree of interaction among them and to establish task difficulty. Scoring guides are refined so that they focus on the key attributes of each cognitive strategy. All task writers are carefully selected and then trained to use task shells to ensure the structural similarity of all tasks and to minimize task variance on extraneous dimensions. Teachers must follow common conditions of administration when introducing the tasks in class. Finally, after teachers score their students' tasks using the common scoring guides, master scorers rescore a subset of the tasks to ensure consistency of teacher scoring.

Ohio Performance Assessment Pilot Project (OPPAP)

The goal of the OPAPP is to contribute to developing a world-class assessment that raises learning expectations for all students, is balanced, and uses a multiple measures approach to assessment and accountability. One important purpose of this approach is to support improvement in instructional practice and align student achievement to college readiness standards and international benchmarks of student performance. The initial phase of the project focused on developing curriculum-embedded, teacher-managed, rich performance tasks that are content-focused, skills-driven, and aligned to college and workplace readiness standards.

OPAPP comprises:

- *Course-embedded performance assessment tasks* that measure the content knowledge and skills learned in 11th and 12th grade courses. Students complete tasks in and out of class over a period of one to four weeks as an embedded part of course curriculum. Teachers administer the tasks under the supervision of their districts and state coaches. Content area and state department curriculum experts develop each task in consultation with teachers, higher education faculty, and national content experts.
- *Performance outcomes* are the content-specific knowledge and skills described by content experts that are needed for college and career success in the 21st century. Ohio teachers, higher education faculty, and state curricular experts arrive at a consensus as to the relative importance and validity of the performance outcomes. The performance outcomes are aligned with state content standards, national content standards (NCTM, NCTE, NAS), college readiness standards (Conley, 2007), and international benchmarks. Performance outcomes are a reduced set of high-leverage content and skills aligned to the Ohio standards and are in the process (upon adoption) of being aligned to the national common core standards. Explication of performance outcomes serves as the blueprint for designing the course-embedded performance assessment tasks.
- *Common scoring rubrics* are a set of evaluative criteria aligned with the performance outcomes, designed to assess all performance tasks within a specific disciplinary focus (e.g., scientific inquiry and investigation, mathematics problem solving, English language arts inquiry, and communication).
- The *scoring system* is based on a set of training protocols and benchmarks designed to ensure high scorer reliability. Included in the up-

coming spring 2010 pilot are data collection strategies to assess score reliability, moderation procedures to audit local teacher scores, and a data feedback loop to give districts and schools formative information to improve teaching and learning. OPPAP is designed to support high-stakes decisions for graduation; the design of the scoring is therefore grounded in psychometric principles and practices to support the validity of the measure. A high degree of technical adequacy is needed, to ensure that the scores generated are credible and defensible and that they measure important dimensions of learning associated with college and career success. This is accomplished, in part, through using item response theory to examine the degree of interaction among them and to establish estimates of student ability and task difficulties.

State Sample. Fifteen pilot districts, including district consortia, were selected from among many more applicants to participate in designing, developing, and piloting the OPAPP system. These sites encompass 24 schools in urban, suburban, and rural contexts.

Use of Results. The performance assessments offer a rich, authentic measure of higher-order thinking and focus on discipline-specific thinking skills deemed necessary for college and career success. Assessment results supply formative information to teachers about student achievement of key performance outcomes, as well as credible, defensible evidence that can be used as part of a summative, high-stakes accountability decision. In addition, the approach is designed specifically to create a two-way flow of information and engagement from the classroom level to the school and district, and from the state and systems level back into the classroom. These performance tasks will be constructed consistent with state curriculum frameworks and course syllabi and will be modeled after proven assessment practices that are already in place in high-performing countries worldwide.

Development of the course-embedded content knowledge and skills assessments is designed to fit the redesign of the Ohio accountability system as:

A component of an end-of-course examination system. Evidence and scores from course-embedded, performance-based assessments are specifically designed for use in combination with state-developed, end-of-course exams. The reference exams include both constructed-response and on-demand multiple-choice items—essays and problem solutions, as well as curriculum-embedded, extended-performance tasks that may require more extended writing, research, and inquiry tasks. The tasks would be constructed by high school faculty and college faculty under the leadership of the department of education and are intended to inform the grade given to the student in the course as well as combine with end-of-course exams to support high school graduation decisions.

A way to satisfy the “senior project” component of Ohio Governor Strickland’s education reform bill. The senior project is defined to include a variety of formats. One format might be a single project in an area of deep interest to students; a second could be a graduation portfolio that includes performance assessments in a selected number of content areas—for example, subject areas chosen by the student, as is common in other countries (the United Kingdom uses O and A level exams from which students choose the areas in which to complete assessments). Students demonstrating mastery in additional content areas could receive additional diploma endorsements (“merit badges”) recognizing their outstanding achievement. These endorsements of accomplishment could be taken into account as part of a student’s application for college or in conjunction with a placement exam used by colleges to determine course eligibility. (See Ohio assessment tasks in the appendix.

One Approach to Development of Next-Generation of State Assessment

If local educators are to go beyond tests of core knowledge and skills in their teaching, then the next generation of high-stakes accountability assessment should include challenging performance tasks aligned to the demands of college and career in the 21st century. This can be accomplished by a two-pronged approach involving (1) a more rigorous statewide, on-demand test, one that includes not only multiple-choice items but also higher-order, constructed-response questions and (2) a locally administered and scored, curriculum-embedded performance assessment component that addresses skills not measurable by the statewide test. In the latter, the local educator role is more substantial than in traditional testing because of participation in implementation and scoring of performance assessment.

On-Demand Component

It is important that both assessment components model good classroom practices. As discussed earlier, sole reliance on multiple-choice tests can and will narrow curriculum and drive instruction toward tested skills (Amrein & Berliner, 2002; Shepard et al., 1995). These authors recommend that an on-demand component (either paper-and-pencil or computer-delivered) include a significant number of constructed-response questions carrying substantial weight toward the total score. Teachers tend to model the state tests in their classroom testing; as Kentucky showed, on-demand, constructed-response testing can indeed lead teachers to place greater emphasis on this format, using rubrics for scoring and gaining the benefit of seeing more actual student work. Of course, the constructed-response questions would lead to greater instructional focus on higher-order thinking skills.

Over the past few decades, several states have used a test with “common” and “matrix-sampled” questions. The former are the same across all forms of a test at a grade level and are the basis for individual scores. The matrix-sampled questions differ across forms and serve several purposes. If included for successive years, they can be used for test equating purposes. Also, matrix sampling is a means of field testing items for use in future years, replacing common items that are released or held for reuse in much later years. Embedding field-test items in operational tests constitutes the most effective means of field-testing items, because students do not know which items are operational and which are being field-tested. Consequently, student motivation is the same for both. Matrix-sampled items can also be used to bolster measurement in subtest areas for which school results are produced.

The NECAP uses a common/matrix-sampled design similar to the one used previously by several consortium states. For illustrative purposes, the essentials of the grade 5 NECAP design (per test form) are shown here.

Reading (9 forms)

- Common multiple-choice: 28 items
- Common 4-point constructed-response: 6 items
- Matrix-sampled multiple-choice equating: 14 items
- Matrix-sampled 4-point constructed-response equating: 3 items
- Matrix-sampled multiple-choice field test: 14 items

Matrix-sampled 4-point constructed-response field test: 3 items, mathematics (9 forms)

- Common multiple-choice: 32 items
- Common 1-point short-answer: 6 items
- Common 2-point short-answer: 6 items
- Common 4-point constructed-response: 4 items
- Matrix-sampled multiple-choice equating: 6 items
- Matrix-sampled 1-point short-answer equating: 2 items
- Matrix-sampled 2-point short-answer equating: 2 items
- Matrix-sampled 4-point constructed-response equating: 1 item
- Matrix-sampled multiple-choice field test: 3 items
- Matrix-sampled 1-point short-answer field test: 1 item
- Matrix-sampled 2-point short-answer field test: 1 item
- Matrix-sampled 4-point constructed-response field test: 1 item

This design allows the equating items to approximate the proportions of numerous types of items in the common test. Short-answer questions in mathematics are quickly scored items that measure skills not effectively measured by multiple-choice items for which students could arrive at correct answers without applying the skills intended to be measured. In the NECAP mathematics tests, there is some variation in the proportion of item types across grades, with greater emphasis on short-answer questions than on 4-point constructed-response items at earlier grades. The highly cited Massachusetts Comprehensive Assessment System (MCAS) has used a test design similar to the NECAP since the late 1990s.

Curriculum-Embedded Performance Component

There is growing belief that to support education of students who can compete effectively in the digital age, assessment systems must be broadened to include locally administered, curriculum-embedded performance assessments (Popham, 1999; Darling-Hammond et al., 2010). Much of what is considered “core knowledge” can be assessed by traditional summative tests, but higher-order skills that traditional tests address either inadequately or not at all should be the focus of curriculum-embedded performance tasks designed to measure higher-order cognitive ability.

Here is a list of characteristics, practices, or steps of an approach to a performance component that capitalizes on the valuable lessons from the past. Full implementation could require three to five years.

1. The state posts models online, tried-and-true tasks and projects calling for individual, scorable products closely aligned to standards, with a total score range of at least 20 points for each task or project. The tasks use materials and other resources readily available in schools, homes, or online. The posting also includes sample student work, scoring rubrics, and specifications for tasks and projects.
2. Teachers use the state-provided tasks and projects in their own instruction and as models for tasks and projects they develop themselves to submit to the state for review. The state also conducts professional development training sessions, using both online and train-the-trainer or coaching models.
3. The state reviews, selects, rejects, revises, and furnishes feedback to teachers for their submissions.
4. The state selects high-quality tasks or projects for pilot testing, collects associated student work, and then posts the tasks and projects, rubrics, and sample student work online for local use. This development, vetting, field testing, and posting sequence is ongoing.
5. The state holds back (does not post) selected tasks or projects, saving them for later use in the local performance assessment component of accountability testing.
6. The state posts a set of tasks or projects for schools to administer within a specified time frame. Teachers score the resulting student work and submit the scores to the state.
7. Each school identifies a low-, middle-, and high-performing student for each task or project and submits the work of those students to the state via electronic portfolio platforms. The teachers' scoring for those students is audited (rescored) by content specialists.
8. Audit scores are sent back to the schools, and local personnel adjust scores of their students to be consistent with the "benchmarks" obtained through the audit process.
9. The next year, steps 6, 7, and 8 are repeated three times, with the tasks and projects for each round chosen to coincide as closely as possible to the time during the year when relevant instruction is given.

10. The results of the performance component are combined with those of the on-demand assessment component, thereby contributing to both student- and school-level results.
11. States support and supply resources for creating learning networks that build and spread educator capacity to strengthen instructional practice by means of effective use of performance assessments.

Policy Considerations

Policy implementation research carried out in a various contexts makes it clear that simply changing the testing requirements and collecting more outcome data does not assure that data will be used to make meaningful and effective decisions about curriculum and instruction. Developing a reliable and valid performance measure is a necessary condition for making more defensible decisions about student learning and more evidence-based decisions about programs. It is also necessary to develop practical guidelines for local and state policy that support strategic use of student performance data in systematically gathering evidence of student achievement on challenging tasks that can measure college and workplace readiness skills.

Costs. Under the current testing and scoring paradigm, there is no question that non-multiple-choice testing can be more expensive than multiple-choice testing. The benefits of assessing higher-order skills and of professional development of teachers involved in scoring should be weighed heavily in decisions about testing programs, along with cost-saving measures that can be implemented even with performance assessments.

Testing companies’ constructed-response scoring systems are particularly efficient and reliable and therefore probably the method of choice for end-of-year summative tests. Here are rough guidelines that can help state officials estimate scoring costs:

Scorer Time and Costs*		
<i>Type of Measure</i>	<i>Training Time</i>	<i>Scoring Time</i>
4-point C-R	2–3 hours	1 minute
Writing sample	Half-day	5 minutes
Performance tasks**	1 day	50 minutes

Notes: * Temporary readers used by testing companies earn \$12 to \$15 per hour.

** It is assumed a “task” yields multiple scorable products.

The number of readings of a student’s work is also relevant. Often a small percentage of constructed-responses are double-scored for quality control purposes, although writing samples and portfolios are typically all double-scored. There are, of course, additional costs associated with scoring leadership. There might be one senior reader or table

leader for every 10 or so readers. Also, there might be a content leader (one per subject), a permanent staff member chargeable to the contract during benchmarking (finding exemplar responses for training), training, and live scoring. Some time from other scoring leadership staff would also be involved for on-demand, constructed-response scoring. Already used effectively for scoring writing samples, computer scoring in the near future may also significantly reduce scoring costs.

For interim assessments, particularly curriculum-embedded performance assessments, teacher scoring is desirable for many reasons, cost savings among them. If the tasks or projects are truly curriculum-embedded, they are part of regular instruction, and there are no additional scorer costs. (Also, the teacher-assigned scores can count toward student course grades.) Scoring training for teachers and the scoring itself can become professional development time, an approach in Australia and in some U.S. states. For years, Maine awarded recertification credits to scorers of writing samples. (The scoring was accomplished at centralized sites.) Of course, there are costs associated with the trainers' time and materials. Online training, however, can be economical. There would also be costs associated with an audit process, by which samples of student work from curriculum-embedded performance assessment are scored.

States can also cut testing costs by joining state consortia for their entire assessment program or program components. The savings are more significant for smaller states whose fixed costs (for test development and program management, for example) are a significant proportion of total costs. Those costs could be shared equally among consortium states. Variable (per student) costs (e.g., materials production and scoring) benefit from economies of scale. Large states do not benefit quite as much economically from joining consortia because, depending on their size, their fixed costs might be insignificant compared to their variable costs and because they already have economies of scale with respect to their variable costs.

Time. As a result of NCLB legislation, many states have abandoned non-multiple-choice formats in their on-demand testing. Cost may be more of a factor in this action than testing or scoring time. It is not widely known that it takes far greater time to create clean data files for final analysis than to produce the scores on constructed-response questions. The latter can be done in days or a few weeks, depending on the number of students and responses. Student-testing materials are not all returned on time, and return instructions are not always followed explicitly. In most of today's programs, which use new tests every year, all student materials and results have to be accounted for and grouped correctly before final results are determined. Traditionally, this takes a great deal of time. For general achievement measures not designed to furnish the kind of rich, diagnostic information teachers need day to day, turnaround time of a few weeks or even months may be acceptable, given that the reports are general profiles of achievement. Timeliness of results from curriculum-embedded components should not be an issue; teachers have immediate access to the scores they assign. Furthermore, because the assessments are curriculum-embedded, they are part of the regular instructional

program. As a result, instructional time is not lost to testing. The scores can be reported and audited in plenty of time for combining with end-of-year, on-demand results for accountability purposes.

State and Local Capacity. Curriculum-embedded assessment is most effectively accomplished if it is consistent with regular instructional practice. Many would argue that for this to be the case schools must implement major reforms. Pre-service and in-service training is critical in assessment literacy, formative and summative assessment practices, and use of data. School leaders at the local and state levels must play a major role in school reform. At the time of this writing, although state resources have been significantly cut back Race-to-the-Top monies are one key source (or possibly the only source) of funding, to support developing the next generation of state accountability, at least for initial design, field testing, and validation of new approaches to state assessment.

Reframing Professional Development. At the core of classroom-based performance assessment work is the explicit intention to build both the assessment literacy and the capacity of teachers to use classroom-embedded assessments. High-quality, classroom-embedded assessments can help shape curriculum and instruction, better informing teacher decisions around learning in order to have the greatest potential to improve the performance of students and schools. A recent study of the 2003-04 school and staffing survey, as well as data from the MetLife Survey and the 2007-08 NSDC Standards and Assessment Inventory (Darling-Hammond et al., 2009.) gives a detailed profile of prevalent professional development practices in the United States, as well as teacher access to high-quality professional development across states and within particular school contexts.

The analysis is sobering. Although the United States has made progress in some areas, such as mentoring novice teachers and deepening teacher content knowledge, the general U.S. approach to professional development is rooted in using short-duration, topically based sessions with little or no follow-up. Schools rarely have a coherent professional development plan with formal structures and supports that foster job-embedded professional development in a collegial, school-based setting. This lack of structure is made evident by low ratings of the usefulness of most professional development opportunities, and teacher perceptions of a lack of collective decisionmaking and problem solving to make professional development practices relevant and meaningful. Effective professional development programs worldwide tend to have some features in common. In high-performing OECD nations, professional development is structured to offer teachers extended learning opportunities and actively engage professional communities in research on relevant education topics—both to learn from one another through mentoring or peer coaching and collectively to guide curriculum, instruction, and professional learning decisions at the school level.

To prepare all children to meet the academic and higher-order thinking demands embedded in the design of performance assessment, it is important to rethink how pro-

fessional development is practiced in this country. In 1990, the psychologist Robert Glaser called the type of teaching and learning needed to support higher-order thinking *adaptive pedagogy*. He proceeded to argue that 21st-century schools must depart from a selective mode, “characterized by minimum variation in the conditions of learning,” in which a “narrow range of instructional options and a limited number of ways to succeed are available.” The switch must be made “to an adaptive mode in which the educational environment can provide a range of opportunities for success—modes of teaching are adjusted to individuals—their backgrounds, talents, interests, and the nature of past performance.”

In summary, professional development organized around building high-quality relationships and professional communities appears to pay significant dividends in deepening teacher knowledge and practice, resulting in improved performance of schools and students. Teachers must have significant opportunities to participate in and influence all aspects of the overall assessment project, from development and scoring the assessment(s) to participating in action research on the impact of performance assessment on teaching and learning. Their participation will improve knowledge of the factors shaping implementation and outcomes and will generate strategies, lessons learned, and evidence that others can use to support instructional improvements in other contexts.

Technology. Moving to a more balanced, multiple-measure accountability system depends, in part, on developing intelligent technologies to capture and transform information that goes beyond simple test scores to include both formative and summative student performance data, ranging from simple text to digital media (embracing exhibitions of student work). Many data management systems currently in use yield accessible and relevant demographic and test score data. These systems, however, are not generally structured to produce actionable, “just in time” evidence of academic factors that schools, districts, and states can use to guide curriculum, instruction, and assessment. More importantly, these systems do not generally focus on putting classroom-based formative and diagnostic data in the hands of teachers and educators to support struggling students and to continuously monitor student progress over time—so called early-warning and on-track measures. Ready access to actionable data embedded in the school’s culture and norms can guide development of preventive and proactive strategies to strategically target resources to high-leverage areas of need, which will lead to improved student outcomes and school improvement.

A smart technology system is designed to extend and “move beyond parallel play” and create viable multimedia tools for teachers and schools to advance 21st-century learning through active use of information and communication technologies that are changing how people share, develop, and process information in this digital age. A platform that integrates traditional and nontraditional data (such as performance data and digital media) is critically needed to support high-performing schools in the future. In a policy brief, Larry Pinkus (AEE, 2009) discussed the potential impact of a data-based

system capable of generating actionable early warning data on schools, especially low performers:

Examining the pattern of early warning indicators can unearth systemic weaknesses and enable schools and districts to address them head on. Early warning data can help to identify schools that face high concentrations of incoming off track students; have clusters of students with certain academic risks factors; have a history of contributing risk factors . . . this kind of data can help educators pinpoint and address problems at the school and student level systemically.

Technologies that will enable educators to use and capture classroom performance data as the data unfold in real time can be powerful tools to support strategic decision making for an immediate difference at the classroom, school, and district levels. Accomplishing such change is possible if intelligent technology platforms are developed to build capacity and foster effective development of accountability tools to support structures having a positive and lasting impact on curriculum, instruction, and assessment.

Credibility. Although Title I legislation played a role in the partial retreat from performance assessment in the 1990s, other factors contributed as well, notably misinformation, low assessment literacy, and regulations related to then-current policies. Two attributes of good performance assessment discussed previously are absolutely critical: close alignment to standards and measurement quality. Having those attributes, however, is not enough; people have to know about them. Unfortunately, despite articles galore during the authentic assessment era on the merits of performance assessment, there was a distinct absence of research reports in the published literature. As a result, hard evidence of technical quality, positive impacts, and feasibility and affordability is lacking. State assessment programs using performance assessments on a large scale had such evidence, but it was not available in the early years as new psychometric, scoring, and standard-setting techniques were being developed. Even when the evidence was produced, it was often buried in technical reports that reached a very limited audience. Commissioned studies of program impact on instruction and student performance saw no greater exposure than technical manuals did.

A curriculum-embedded performance activity as envisioned here would yield several products scored much the same as constructed-responses or writing samples. Consequently, corrections must be made to current misconceptions about constructed-response scoring associated with performance assessment. As suggested in an earlier section, it is inappropriate to claim that a multiple-choice test is more reliable than a human-scored, constructed-response test. With high-quality items of either type of test, the number of score points per item and number of items per test become the critical factors for reliability.

Unfortunately, during the authentic assessment period the psychometric community was not united in supporting performance assessment. Even though scoring techniques evolved and highly reliable results were achieved on constructed-response items and richer embedded performance tasks both internationally and in the United States, use of state-of-the-art techniques in the United States has been more limited, leading to skepticism because of limited visibility and documentation of quality. The lesson here is that performance assessment advocates must be proactive and do a much better job of publicizing evidence of the merits and technical quality of performance assessment. Furthermore, the psychometric community should embrace the challenge of performance assessment and support it by tackling any psychometric problems this testing approach poses, rather than merely citing them. The impact of high-stakes testing on classroom practice is significant. Teachers, quite reasonably, want to prepare their students for those tests. To that end, their own classroom tests model their state's approach to assessment. State testing programs and the testing industry have made classroom teachers reliant on multiple-choice items for their own tests, caused them to see less actual student work, and produced and encouraged testing (and instruction) that ignores valued standards (and skills) not easily assessed via the multiple-choice format.

Philosophical or ideological differences are also problematic today. Interestingly, at a recent conference Gong (2009) identified challenges faced by the Kentucky program of the 1990s and explained how “many powerful state groups and individuals opposed KIRIS.” Some believed the states’ “valued outcomes” (now called “content standards”) and assessments measuring them intruded on “parental prerogatives and personal privacy” and reflected a “conspiracy to produce compliant workers.” He also reported mathematics and literacy wars, which in essence were about low-level versus higher-level cognitive skills. Today, these same wars are going on in the unfortunate debate about core knowledge versus 21st-century skills—unfortunate because the battles continue even though the two sides acknowledge the importance of both perspectives.

Whether we are dealing with measurement or philosophical issues, it is clear that if progress is to be made then proponents of curriculum-embedded performance assessment and any other innovations associated with it must prove their case. This was not done in the 1990s, when programs were implemented before proper foundations could be laid. Fortunately, performance assessment advocates are not starting at ground zero this time.

Political Will. Gong also explained how the old Kentucky program got “caught in political election battles,” becoming the victim of extreme partisan politics. Changes in legislative and education department leadership and “an adversarial relationship between the Department of Education and a legislative oversight group” also contributed to the demise of KIRIS. Clearly, a decade later the politicization of education is playing a role in shaping the future of education, and more specifically educational assessment.

It remains to be seen what the large-scale assessment landscape will look like in the second decade of the 21st century.

It is critical that states involve “frontline implementers”—teachers and students—in developing a more balanced assessment system. It is also important to give teachers opportunities to learn from assessments of their own students. These frontline users must be the key champions of reform for it to succeed. They have to carry the message to all key constituents and build understanding and support, as well as be involved in correcting misinformation and political posturing.

Conclusions

The nation has adhered to an assessment paradigm that principally favors one form of accountability, driven by a restrictive focus on developing standardized tests to meet proficient standards of performance as defined by NCLB. Instead of naming winners and losers among competing policies and competing testing philosophies, it's time to take an inclusive and bold step forward by focusing efforts on keeping what works in NCLB, fixing its problems, and broadening the prevalent conception of accountability to include performance measures of higher-order thinking. Other high-performing nations, such as Australia, Canada (Alberta), Hong Kong, Singapore, Sweden, and the United Kingdom, are implementing or developing more balanced, content- and skills-driven accountability systems that use classroom-based performance measures in combination with national examinations to assess student learning. Although several of these countries stepped up to the new global and economic realities and reformed their education systems, the United States appears to have some catching up to do. Warning signs such as a high dropout rate, high attrition in college, and student achievement levels nationally that are far below those in other countries highlight the need to establish and support high standards of performance for students, schools, districts, and states.

This paper describes the lessons learned from states' current and past efforts to use performance assessments. It builds on those lessons to offer a new vision for making performance assessments an integral part of a statewide, multiple-measure, balanced assessment system. Identifying promising practices, as well as missteps, in building performance assessment for statewide use can inform and guide development of the next generation of assessment in this country. The lessons learned can inform policymakers and practitioners alike. Future work to develop new measures of student learning should be accompanied by an equally vigorous effort to develop and evaluate a system of technical, organizational, and human resource supports for states and institutions to enable them to make better use of accountability data, including performance data. Designed and used well, development of the next generation of state accountability systems has the potential to strengthen instruction, curriculum, and assessment as well as serve as a catalyst to reform schools and districts.

References

- Alliance for Excellent Education. (2009). The high cost of high school dropouts: What the nation pays for inadequate high schools. Issue Brief. Washington, DC: author.
- Amrein, A. L., & Berliner, D. C. (2002, March 28). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, 10(18).
- Association of Test Publishers & Council of Chief State School Officers. (2010). *Operational best practices*. Report produced at the request of former Secretary of Education Margaret Spellings.
- Badger, E., Thomas, B., & McCormack, E. (1990). *Background summary: Beyond paper and pencil*. Malden: Massachusetts Department of Education.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice*, 5(1), 7–74.
- Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80, 139–144.
- Bransford, J. D., & Darling-Hammond, L. (2005). Preparing teachers for a changing world: What teachers should learn and be able to do. San Francisco: Jossey-Bass.
- Bransford, J. D., & Schwartz, D. L. (2001). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24, 61–100.
- Burstall, C. (1986, Spring). Innovative forms of measurement: A United Kingdom perspective. *Educational Measurement: Issues and Practices*, 17–22.
- Burstall, C., Baron, J., & Stiggins, R. (1987). *The use of performance testing in large-scale student assessment programs*. Paper presented at the Education Commission of the States' 17th Annual Assessment Conference .
- Campbell, D. T., & Stanley, J. C. (1963). Experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.
- Chronbach, L. J. (1971). Test validation. In E. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.
- Cizek, G. J. (2001, Winter). More unintended consequences of high-stakes testing. *Educational Measurement, Issues and Practice*, 20(4), 19–28.
- Coe, P., Leopold, G., Simon, K., Stowers, P., & Williams, J. (1994). *Perceptions of school change: Interviews with Kentucky students*. Charleston, WV: Appalachia Educational Laboratory.
- Coelen, S., Fulton, D., & Rende, S., (2008, April). *Next steps: Preparing a quality workforce*. Storrs, CT: Department of Economics and Connecticut Center for Economic Analysis, University of Connecticut.
- Conley, D. T. (2010). *College and career ready: Helping all students succeed beyond high school*. San Francisco: Jossey-Bass.
- Conley, D. T. (2007). *Redefining college readiness*. Eugene, OR: Educational Policy Improvement Center.
- Connecticut State Board of Education. (2009). *Connecticut academic performance test program overview*. Retrieved October 2, 2009 from http://www.csde.state.ct.us/public/cedar/assessment/capt/resources/misc_capt/2009%20CAPT%20Program%20Overview.pdf
- Connecticut State Department of Education. (2006). *CAPT third generation handbook for reading and writing across the disciplines*. Retrieved October 2, 2009 from <http://www.sde.ct.gov/sde/cwp/view.asp?a=2618&q=320866>
- Connecticut State Department of Education. (2007, August 28). Science curriculum-embedded tasks, CAPT: Generation III. Retrieved October 2, 2009, from <http://www.sde.ct.gov/sde/cwp/view.asp?a=2618&q=320892>
- Darling-Hammond, L., Wei, R.C., Andree, A., Richardson, N., & Orphanos, S. (2009, February). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. Palo Alto, CA: National Staff Development Council and the Stanford Center for Opportunity Policy in Education.

- Darling-Hammond, L., Pecheone, R., Jacquith, A., Schultz, S., Walker, L., & Wei, R.C. (2010). Developing an internationally comparable balanced assessment system that supports high-quality learning. In *National conference on next generation assessment systems*. Retrieved September 28, 2010, from <http://www.k12center.org/rsc/pdf/Darling-HammondPechoneSystemModel.pdf>
- Engeström, Y. (1999). Activity theory and individual and social transformation. In Y. Engeström, R., Miettinen, & R. Punamäki (Eds.), *Perspectives on activity theory* (pp. 19–38). Cambridge, UK: Cambridge University Press.
- Flexer, R. J. (1991). *Comparisons of student mathematics performance on standardized and alternate measures in high-stakes contexts*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Frederiksen, J. R., & White, B. Y. (1997). Cognitive facilitation: A method for promoting reflective collaboration. *Proceedings of the second international conference on computer support for collaborative learning* (pp. 53–62). Toronto: University of Toronto.
- Frederiksen, J. R., & Collins, A. (1996). Designing an assessment system for the workplace of the future. In L. B. Resnick, J. Wirt, & D. Jenkins (Eds.), *Linking school and work: Roles for standards and assessment* (pp. 193–221). San Francisco: Jossey-Bass.
- Glaser, R. (1990a). *Testing and assessment: O tempora! O mores!* Paper presented at the Horace Mann lecture at the University of Pittsburgh .
- Glaser, R. (1990b). Toward new models for assessment. *International Journal of Educational Research*, 14(5), 475–483.
- Gong, B. (2009). *Innovative assessment in Kentucky's KIRIS system: Political considerations*. Paper presented at the Best Practices in State Assessment workshop sponsored by the National Academy of Sciences, Washington, DC, and at the Board on Testing and Assessment session “What is the status of innovative assessments? Political considerations.”
- Hamilton, L., Stecher, B., & Yuan, K. (2008). *Standards-based reform in the United States: History, research, and future directions*. Washington, DC: Center on Educational Policy, Rand Corporation.
- Hiebert, E. H. (1991). *Comparisons of student reading performance on standardized and alternative measures in high-stakes contexts*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- K–12 literacy, restructuring of the learning environment at the middle and high school levels, and proficiency based graduation requirements (PBGR) at high schools, L-6-3 C.FR. (2008).
- Kahl, S. (2008, June). *The assessment of 21st-century skills: Something old, something new, something borrowed*. Paper presented at the CCSSO National Conference on Student Assessment, Orlando, FL.
- Kahl, S., Abeles, S., & Baron, J. (1985). *Results of the 1984–85 Connecticut assessment of educational progress in science: Implications for improving local science programs*. Paper presented at the National Science Teachers Association area meeting, Hartford.
- Kentucky Department of Education. (1997). *KIRIS accountability cycle 2 technical manual*. Retrieved from contractor files.
- Khattari, N., Kane, M., & Reeve, A. (1995, November). How performance assessments affect teaching and learning. *Educational Leadership*.
- Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review*, 31(5), 415–439.
- Klein, S., Freedman, D., Shavelson, R., & Bolus, R. (2008). Assessing school effectiveness. *Evaluation Review*, 32, 511–525.
- Koretz, D., Barron, S., Mitchell, K., & Stecher, B. (1996). Perceived effects of the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: Rand Corporation.
- Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

- Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4–12: Executive summary*. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Measured Progress. (2009). *New England Common Assessment Program 2008–2009 technical report*. Dover, NH: author.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.
- Moss, P. A., Girard, B., & Haniford, L. (2006). Validity in educational assessment. *Review of Research in Education*, 30, 109–162.
- Murnane, R., & Levy, F. (1996). *Teaching the new basic skills: Principles for educating children to thrive in a changing economy*. New York: Free Press.
- National Association of State Boards of Education. (2009). *Reform at a crossroads: A call for balanced systems of assessment and accountability*. Report of the NASBE Study Group on Assessment Systems for the 21st-Century Learner, Arlington, VA.
- National Center on Education and the Economy. (2007). *Tough choices, tough times: The report of the New Commission on Skills of the American Workforce*. Washington, DC.
- New England Common Assessment Program. (2009). *2009 Test administrator manual—Grade 11 science*. Retrieved October 2, 2009 from http://education.vermont.gov/new/pdfdoc/pgm_assessment/newcap/manuals/science/admin_manual_09_grade_11.pdf
- New Hampshire code of administrative rules-education*, 306.27(d) C.F.R. (2005). Retrieved October 2, 2009, from <http://www.ed.state.nh.us/education/laws/documents/306Adopted.pdf>.
- New Jersey Department of Education. (2005). *October 2005 and March 2006 HSPA cycle I and cycle II score interpretation manual*. Retrieved October 2, 2009 from <http://www.state.nj.us/education/assessment/hs/sim.pdf>
- New Jersey Department of Education. (2008a). *March 2009 high school proficiency assessment: Student preparation booklet*. Retrieved October 2, 2009 from <http://www.state.nj.us/counties/cumberland/0610/schools/distschools/senior/guidanceimages/HSPA%20Student%20Prep%20Booklet%20%2708.pdf>
- New Jersey Department of Education. (2008b). *Special review assessment administration manual 2008–2009 school year*. Retrieved October 2, 2009 from <http://www.state.nj.us/education/assessment/hs/sra/man.pdf>
- New York Commissioner of Education. (n.d.). *Department-approved alternative examinations acceptable for meeting requirements for a local or Regents diploma*. Retrieved from <http://www.emsc.nysed.gov/osa/hsgen/list.pdf>
- New York State Education Department. (1987). *History of Regents Examinations: 1865 to 1987*. Retrieved September 30, 2009, from <http://www.emsc.nysed.gov/osa/hsinfo/gen/hsinfo/genarch/rehistory.htm>
- New York State Education Department, Office of State Assessment. (2008). *Regents Examinations, Regents competency tests, and second language proficiency examinations: School administrator's manual*. Retrieved September 30, 2009 from <http://www.emsc.nysed.gov/osa/sam/secondary/sam08-pdf/nysed-sam08.pdf>
- Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Pinckney, E., & Taylor, G. (2006). *Standards and assessment memorandum*. Retrieved October 2, 2009 from http://education.vermont.gov/new/pdfdoc/pgm_curriculum/local_assessment/assessment_guidance_030106.pdf
- Popham, W. J. (1999). Why standardized test scores don't measure educational quality. *Educational Leadership*, 56(6), 8–15.
- Quaglia Institute. (2008). *My Voice student report 2008*. Portland, ME: Quaglia Institute for Student Aspirations.
- Rhode Island Board of Regents for Elementary and Secondary Education. (2008, September 3). *Regulations of the Board of Regents for Elementary and Secondary Education: K-12 Literacy, Restructuring of the Learning Environment at the Middle and High School Levels, and Proficiency Based Graduation Requirements (PBGR) at High Schools*, L-6-3.3. Use of proficiency measures for high school gradu-

- ation. Retrieved September 15, 2009, from <http://www.ride.ri.gov/Regents/Docs/RegentsRegulations/HS%20Regulations%20September,%202008.pdf>
- Rhode Island Department of Education. (2005). The Rhode Island high school diploma system: All kids well prepared for high-performing, bright futures. Retrieved September 21, 2010 from http://www.ride.ri.gov/HighSchoolReform/DOCS/PDFs/HIGH%20school%20reform/HSDiploma_v071405.pdf
- Rhode Island Department of Education & Education Alliance at Brown University. (2005). *Required graduation portfolio elements*. Retrieved October 2, 2009 from http://www.ride.ri.gov/HighSchool-Reform/DSLAT/pdf/por_040103.pdf
- Schlaflly, P. (2001). Dumbing down and developing diversity. *Phyllis Schlaflly Report*, 34(8). Retrieved September 28, 2010 from <http://www.eagleforum.org/psr/2001/mar01/psrmar01.shtml>
- Schleicher, A. (2009). International assessment of student learning outcomes. In L. Pinkus (Ed.), *Meaningful measurement: The role of assessments in improving high school education in the twenty-first century*. Washington, DC: Alliance for Excellent Education.
- Shavelson, R.J. (2010). *Measuring college learning responsibly: Accountability in a new era*. Stanford, CA: Stanford University Press.
- Shepard, L. A. (2002). The hazards of high stakes testing. *Issues in Science and Technology*, 19(2), 53–58.
- Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J. (1995). *Effects of introducing classroom performance assessments on student learning*. CSE technical report 394. Boulder: Center for Research on Evaluation, Standards, and Student Testing (CRESST) and University of Colorado at Boulder.
- Shyer, C. (2009). *August 2009 Regents Examinations and Regents competency tests*. Retrieved September 30, 2009 from <http://www.emsc.nysed.gov/osa/08-09memo/jun-aug-09/724/563-809.pdf>
- Tucker, B. (2009). *Beyond the bubble: Technology and the future of student assessment*. Washington, DC: Education Sector.
- University of the State of New York State Education Department. (2009a). *Information booklet for scoring Regents Examinations in global history and geography and United States history and government*. Retrieved September 30, 2009, from <http://www.emsc.nysed.gov/osa/08-09memo/jun-aug-09/730/541hg-809.pdf>
- University of the State of New York State Education Department. (2009b). *Information booklet for scoring the Regents Comprehensive Examination in English*. Retrieved September 30, 2009, from <http://www.emsc.nysed.gov/osa/08-09memo/jun-aug-09/730/541e-809.pdf>
- Vermont Department of Education. (n.d.). *Core principles of high-quality local assessment systems*. Retrieved October 2, 2009 from http://education.vermont.gov/new/pdfdoc/pgm_curriculum/local_assessment/core_principles_08.pdf
- Vermont State Board of Education manual of rules and practices: School quality standards*, 2000 C.F.R. § 2120 (2006).
- Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve*. San Francisco: Jossey-Bass.
- Wilson, M. (Ed.). (2004). *Towards coherence between classroom assessment and accountability*. Chicago: University of Chicago Press.
- Wood, G. H., Darling-Hammond, L., Neill, M., & Roschewski, P. (2007). *Refocusing accountability: Using local performance assessments to enhance teaching and learning for higher order skills*. Briefing paper prepared for members of the Congress of the United States. Stewart, OH: Forum for Education and Democracy.

Appendix

Performance Assessment Exemplars

Collegiate Learning Assessment

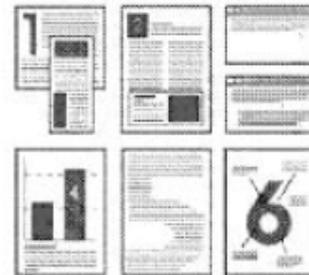
Performance Task

Published 2007

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 and 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.

**CAPT Science Open-Ended Item:
Enzyme Investigation**

Enzyme Investigation

A group of students hypothesized that adding an enzyme to applesauce would produce more juice than adding an enzyme to mashed pears. The students wrote the following procedure for their investigation.

Procedure:

1. Place a coffee filter in each of two plastic funnels and place each funnel in a separate beaker.
2. Put 113 g of applesauce in one filter-covered funnel.
3. Put 113 g of peeled, mashed pears in one filter-covered funnel.
4. Add 3 drops of enzyme A to the applesauce and stir for one minute.
5. Add 3 drops of enzyme B to the mashed pears and stir for one minute.
6. Allow the fruit to sit for 10 minutes.
7. Measure and record the amount of juice contained in each beaker.
8. Repeat the procedure exactly for a second trial to verify data.

The data collected during the investigation are shown in the table below.

Type of Fruit	Juice Produced (mL)		
	Trial 1	Trial 2	Average
Applesauce	12	11	11.5
Mashed Pears	13	12	12.5

The students claimed that their original hypothesis was correct.

- a) Explain why the credibility of the students' claim should be questioned.
- b) Describe two changes that the students should make to their procedure that would allow their original hypothesis to be more accurately tested and/or would ensure the accuracy of their results.

Write your answer in your answer booklet.

Rubric for *Enzyme Investigation*

Possible Correct Responses:

Credibility Problem:

- The procedure allowed for more than one variable.
- The students used different enzymes in each type of fruit.
- Data doesn't support the claim.
- Other acceptable responses

Changes:

- Use the same enzyme in each type of fruit (either A or B, but not both).
- Use both enzymes on each fruit.
- Add a control to the investigation (a sample of each fruit to which no enzyme is added).
- Perform additional trials.
- Other acceptable responses

3-Point Rubric:

Score 3

The response provides an explanation for why the credibility should be questioned and describes two changes the students could make that would allow their original hypothesis to be more accurately tested and/or would ensure the accuracy of their results.

Score 2

The response provides an explanation for why the credibility should be questioned and describes one change the students could make that would allow their original hypothesis to be more accurately tested and/or would ensure the accuracy of their results.

-or-

The response fails to provide or provides an incorrect explanation for why the credibility should be questioned, but describes two changes the students could make that would allow their original hypothesis to be more accurately tested and/or would ensure the accuracy of their results.

Score 1

The response provides an explanation for why the credibility should be questioned, but fails to correctly describe any changes.

-or-

The response fails to provide or provides an incorrect explanation for why the credibility should be questioned, but describes one change.

Score 0

The response describes little or no accurate or relevant information related to the enzyme investigation.

Strand IV: Cell Chemistry and Biotechnology

Expected Performance: D INQ.2 Read, interpret, and examine the credibility and validity of scientific claims in different sources of information.

Connecticut

Science Curriculum-Embedded Task

Global Interdependence Laboratory Investigation, Released 2007

Student Name: _____ Class: _____



Acid Rain

Laboratory Investigation
Student Materials

Acid Rain

Connecticut

Science Curriculum-Embedded Task

Global Interdependence Laboratory Investigation, Released 2007

Student Materials

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:]

Proposed building materials:

Containers with lids	Limestone chips
Graduated cylinder	Marble chips
Vinegar (simulates acid rain)	Red sandstone chips
pH paper/meter	Pea stone
Safety goggles	Access to a balance

Designing and Conducting Your Experiment

- 1. In your own 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 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.

Safety note: Students must wear approved safety goggles and follow all safety instructions. When you have finished, your teacher will give you instructions for cleanup procedures, including proper disposal of all materials.

Communicating Your Findings

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

- **A statement of the problem you investigated; a hypothesis (“If . . . then . . . because . . .” statement) that described what you expected to find and why. Include 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, 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 lack of validity of your conclusions.** Also, discuss how your experiment could be improved if you were to do it again.

Kentucky

Kentucky Core Content Test

Mathematics Open-Response Item, Released 2007

How Tall Is the Broccoli?

At the beginning of the summer, Shanna bought broccoli plants that were 5 inches tall and planted them in her garden. The nursery tag says the plants will grow at an average rate of 2 inches per week.

- In your Student Response Booklet, make a table that shows the expected height of Shanna's plants at the end of 1, 2, 3, and 4 weeks. Label weeks as n and height as t .
- Write an algebraic equation for a **sequence** that can be used to find the height, t , of Shanna's broccoli plants at the end of week n during the summer.
- Use your equation from **part b** to find the height, in inches, of Shanna's broccoli plants at the end of the 12th week.

Scoring Guide

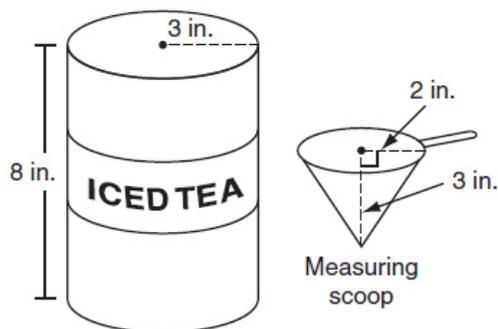
SCORE	DESCRIPTION
4	Student scores 4 points.
3	Student scores 3 – 3.5 points.
2	Student scores 2 – 2.5 points.
1	Student scores 0.5 – 1.5 points.
0	Student's response is totally incorrect or irrelevant.
Blank	No student response.

Score Points

Part a:	score 1 point OR score 0.5 point	correct table with labels table with 3 correct values OR correct graph with exact points (not estimates) graphed OR correct values not in a table
Part b:	score 2 points OR score 1.5 points OR score 1 point OR score 0.5 point	correct equation correct expression OR correct equation with undefined variables (other than n and t) correct verbal rule OR correct equation with “ \times ” used as a multiplication symbol OR correct expression with undefined variable (other than n) correct expression with “ \times ” used as a multiplication symbol
Part c:	score 1 point OR score 0.5 point	correct answer (based on answer to part b) with work correct answer with no work or with work not based on answer to part b OR incorrect answer due to calculation error

New Hampshire, Rhode Island, Vermont
 New England Common Assessment Program
 Mathematics Short-Answer Item, Released 2008

This diagram shows a cylindrical container of iced tea mix and a cone-shaped measuring scoop.



One level measuring scoop of iced tea mix makes one pitcher of iced tea. How many pitchers of iced tea can be made from this full container of iced tea mix? Show your work or explain how you know.

Scoring Guide

Score	Description
2	Student gives the correct answer, 18 , and provides appropriate work or explanation.
1	Student gives correct answer but does not provide appropriate work or explanation. OR Student's work or explanation shows correct strategy in solving the problem but contains an error in computation.
0	Response is incorrect or contains some correct work that is irrelevant to the skill or concept being measured.
Blank	No response

Everyday Life at the End of the Last Ice Age Informational Writing (Report)

A student wrote this fact sheet about life 12,000 years ago, at the end of the last ice age. Read the fact sheet. Then write a response to the prompt that follows.

Everyday Life at the End of the Last Ice Age

- people lived in bands of about 25 members
- lived mainly by hunting and gathering
- shared decision-making fairly equally among members in a band
- each person skilled in every type of job
- diet: small and large mammals, fish, shellfish, fruits, wild greens and vegetables, grains, roots, and nuts
- approximately 10,000 years ago woolly mammoth became extinct
- nomadic based on time of year or movement of animal herds
- cooked meat by roasting it on a spit over a fire or by boiling it inside a piece of leather secured by a twig
- gathered herbs
- made everything themselves: tools, homes, clothing, medicines, etc.
- worked about 2–3 hours a day getting food
- worked about 2–3 hours a day making and repairing tools and clothes
- spent remainder of day relaxing with family and friends
- told stories, danced, sang, and played games
- owned very few possessions
- no concept of rich or poor
- communicated through art (painting and sculpture) and the spoken word
- buried their dead and had concepts of religion and an afterlife
- sometimes adorned themselves with ornaments and decorations such as jewelry, tattoos, body painting, and elaborate hairstyles

What would a person from 12,000 years ago find familiar and/or different about life today? Select relevant information from the fact sheet and use your own knowledge to write a report.

New Hampshire, Rhode Island, Vermont
 New England Common Assessment Program
 Informational Writing Prompt, Released 2008

Scoring Guide:

Score	Description
6	<ul style="list-style-type: none"> • purpose is clear throughout; strong focus/controlling idea OR strongly stated purpose focuses the writing • intentionally organized for effect • fully developed details; rich and/or insightful elaboration supports purpose • distinctive voice, tone, and style enhance meaning • consistent application of the rules of grade-level grammar, usage, and mechanics
5	<ul style="list-style-type: none"> • purpose is clear; focus/controlling idea is maintained throughout • well organized and coherent throughout • details are relevant and support purpose; details are sufficiently elaborated • strong command of sentence structure; uses language to enhance meaning • consistent application of the rules of grade-level grammar, usage, and mechanics
4	<ul style="list-style-type: none"> • purpose is evident; focus/controlling idea may not be maintained • generally organized and coherent • details are relevant and mostly support purpose • well-constructed sentences; uses language well • may have some errors in grammar, usage, and mechanics
3	<ul style="list-style-type: none"> • writing has a general purpose • some sense of organization; may have lapses in coherence • some relevant details support purpose • uses language adequately; may show little variety of sentence structures • may have some errors in grammar, usage, and mechanics
2	<ul style="list-style-type: none"> • attempted or vague purpose • attempted organization; lapses in coherence • generalized, listed, or undeveloped details • may lack sentence control or may use language poorly • may have errors in grammar, usage, and mechanics that interfere with meaning
1	<ul style="list-style-type: none"> • minimal evidence of purpose • little or no organization • random or minimal details • rudimentary or deficient use of language • may have errors in grammar, usage, and mechanics that interfere with meaning
0	Response is totally incorrect or irrelevant.
Blank	No response

WRITING
PAT-LWPP-0015

STUDENT ID No: _____

PILOT MATERIALS FOR OCTOBER 2002
New Jersey Department of Education
HSPA/SRA LANGUAGE ARTS LITERACY
Performance Assessment Task (PAT)

All HSPA/SRA Performance Assessment Tasks (PATs) are secure assessment instruments and may NOT be used as instructional materials. Each HSPA/SRA Performance Assessment Task may be administered to an individual student only ONCE. The PATs must be kept in locked storage at all times when not in use.

Writing Situation:

In a recent job interview, you were asked about a goal you set for yourself and how you accomplished it. You were unable to answer the question and now you are concerned that you will not get the job.

You decide to write a letter to the employer to complete your interview.

Directions for Writing:

Write a letter to the employer. Describe a goal you set for yourself and how you accomplished it. Convince the employer that you would be a good person for the job.

Materials/Resources:

- Paper, pencil or pen
- Access to a word processor or computer, if desired
- Writing Prompt

Techniques for PAT Scoring:

- New Jersey Registered Holistic Scoring Rubric (a 1-to-6 point scale)

Writing Prompt 1: Score Scale Point 2

The response indicates a **LIMITED COMMAND** of written language. The writing samples in this category:

- | | |
|----------------------------------|--|
| CONTENT/
ORGANIZATION | <ul style="list-style-type: none">• may not have an opening and/or a closing. These responses will exhibit an attempt at organization. In other words, there will be some evidence the writer attempted to control the details. The responses relate to the topic, but in some papers, the writer drifts away from the primary focus or abruptly shifts focus. In other papers, there is a single focus, but there are few, if any, transitions, making it difficult to move from idea to idea. Details are presented with little, if any, elaboration—highlight papers. |
| USAGE | <ul style="list-style-type: none">• may have numerous problems with usage, but they are not totally out of control. |
| SENTENCE
CONSTRUCTION | <ul style="list-style-type: none">• may demonstrate excessive monotony in syntax and/or rhetorical modes. There may be numerous errors in sentence construction. |
| MECHANICS | <ul style="list-style-type: none">• may display numerous severe errors in mechanics. |

Writing Prompt 1: Score Scale Point 3

The response indicates a **PARTIAL COMMAND** of written language. The writing samples in this category:

- | | |
|----------------------------------|---|
| CONTENT/
ORGANIZATION | <ul style="list-style-type: none">• may not have an opening and/or closing. These responses relate to the topic and usually have a single focus. Some of these papers may drift from the focus or abruptly shift focus; however, in these papers, at least one of the subjects focused on clearly meets the criteria for a “3.” For example, some “3” papers are sparse—they have several details with a little elaboration, but they are organized and controlled; some “3” papers will ramble somewhat, repeating ideas resulting in a lengthy response that would otherwise be sparse; and other “3” papers have elaborated ideas and details, but the writing sample is interrupted by organizational flaws/lapses or by a lack of transition between ideas or between clusters of ideas. |
| USAGE | <ul style="list-style-type: none">• may display (a) pattern(s) of errors in usage. |
| SENTENCE
CONSTRUCTION | <ul style="list-style-type: none">• may demonstrate little variety in syntax structure and/or rhetorical modes. There may be errors in sentence construction. |
| MECHANICS | <ul style="list-style-type: none">• may display (a) pattern(s) of errors in mechanics. |

Writing Prompt 1: Score Scale Point 4

The response indicates an **ADEQUATE COMMAND** of written language. The writing samples in this category:

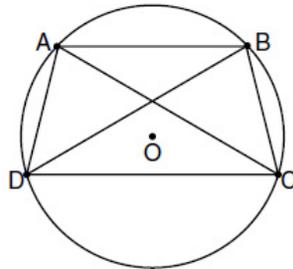
- | | |
|----------------------------------|--|
| CONTENT/
ORGANIZATION | <ul style="list-style-type: none">• generally will have an opening and a closing. The responses relate to the topic. They have a single focus and are organized. There is little, if any, difficulty moving from idea to idea. Ideas may ramble somewhat, and clusters of ideas may be loosely connected; however, an overall progression is apparent. In some papers, development is uneven, consisting of elaborated ideas interspersed with bare, unelaborated details. |
| USAGE | <ul style="list-style-type: none">• may display some errors in usage, but no consistent pattern is apparent. |
| SENTENCE
CONSTRUCTION | <ul style="list-style-type: none">• may demonstrate a generally correct sense of syntax. They avoid excessive monotony in syntax and/or rhetorical modes. There may be a few errors in sentence construction. |
| MECHANICS | <ul style="list-style-type: none">• may display some errors in mechanics, but these errors will not constitute a consistent pattern, nor do they interfere with the meaning of the response. |

Part IV

Answer the question in this part. The correct answer will receive 6 credits. Clearly indicate the necessary steps, including appropriate formula substitutions, diagrams, graphs, charts, etc. For the question in this part, a correct numerical answer with no work shown will receive only 1 credit. [6]

- 38 In the diagram below, quadrilateral $ABCD$ is inscribed in circle O , $\overline{AB} \parallel \overline{DC}$, and diagonals \overline{AC} and \overline{BD} are drawn.

Prove that $\triangle ACD \cong \triangle BDC$.



Rubric for Item 38

- [6] A complete and correct proof that includes a concluding statement is written.
- [5] A proof is written that demonstrates a thorough understanding of the method of proof and contains no conceptual errors, but one statement or reason is missing or incorrect.
- [4] A proof is written that demonstrates a good understanding of the method of proof and contains no conceptual errors, but two statements or reasons are missing or incorrect.
- [3] A proof is written that demonstrates a good understanding of the method of proof, but one conceptual error is made.
- [2] A proof is written that demonstrates an understanding of the method of proof, but one conceptual error is made and one statement or reason is missing or incorrect.
- or*
- [2] Some correct relevant statements about the proof are made, but three or four statements or reasons are missing or incorrect.
- [1] Only one correct relevant statement and reason are written.
- [0] The “given” and/or the “prove” statements are rewritten in the style of a formal proof, but no further correct relevant statements are written.
- or*
- [0] A zero response is completely incorrect, irrelevant, or incoherent or is a correct response that was obtained by an obviously incorrect procedure.

New York
Regents Examinations in Physics
Constructed-Response Item, Released 2001

Base your answers to questions 69 through 71 on the information below.

You are given a 12-volt battery, ammeter A , voltmeter V , resistor R_1 , and resistor R_2 .
Resistor R_2 has a value of 3.0 ohms.

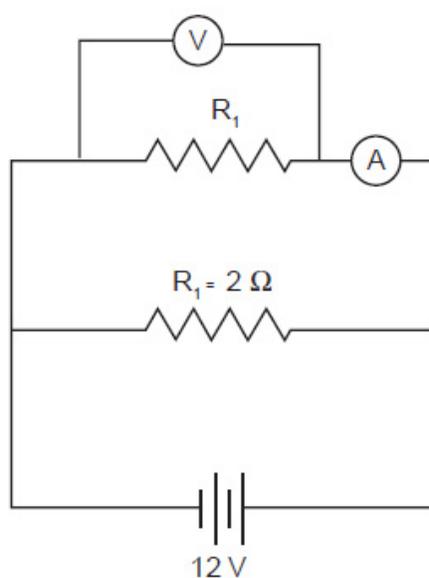
- 69 Using appropriate symbols from the *Reference Tables for Physical Setting/Physics*, draw and label a complete circuit showing:
- resistors R_1 and R_2 connected in parallel with the battery [1]
 - the ammeter connected to measure the current through resistor R_1 , only [1]
 - the voltmeter connected to measure the potential drop across resistor R_1 [1]
- 70 If the total current in the circuit is 6.0 amperes, determine the equivalent resistance of the circuit. [1]
- 71 If the total current in the circuit is 6.0 amperes, determine the resistance of resistor R_1 . [Show all calculations, including the equation and substitution with units.] [2]

New York
Regents Examinations in Physics
Constructed-Response Item, Released 2001

69 Allow a total of 3 credits, allocated as follows:

- 1 credit for R_1 and R_2 connected in parallel with the battery
- 1 credit for the ammeter connected in series with R_1 , only
- 1 credit for the voltmeter connected in parallel with R_1 or equivalent position

Example of Acceptable Response



70 Allow 1 credit for 2.0 V or 2 V .

New York

Regents Examinations in United States History and Government Document-Based Essay, Released 2000

Directions: Using information from the documents provided, and your knowledge of United States history, write a well-organized essay that includes an introduction, several paragraphs, and a conclusion.

Historical Context:

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.

Task:

Using information from the documents 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.

Guidelines:

Be sure to:

- Address all aspects of the *Task* by accurately analyzing and interpreting at least four documents
- Incorporate information from the documents in the body of the essay
- Incorporate relevant outside information throughout the essay
- Richly support the theme with relevant facts, examples, and details
- Write a well-developed essay that consistently demonstrates a logical and clear plan of organization
- Introduce the theme by establishing a framework that is beyond a simple restatement of the *Task* or *Historical Context* and conclude the essay with a summation of the theme

New York

*Regents Examinations in United States History and Government
Document-Based Essay, Released 2000*

New York

Regents Examinations in United States History and Government Document-Based Essay, Released 2000

5 [continued]

- The plan of organization will generally take one of two approaches:
 - (1) The student will discuss advantages, then disadvantages, of industrialization and then the impact upon two groups, or
 - (2) The student will discuss advantages and disadvantages of industrialization upon one group and then discuss advantages and disadvantages of industrialization upon a second group
- Introduces the theme by establishing a framework that is beyond a simple restatement of the *Task* or *Historical Context* and concludes with a summation of the theme

4

- Discusses all three tasks, although the discussion of one task may be less complete than the discussion for the other two tasks
- Discussion includes accurate information from at least *four* documents in the body of the essay
- Incorporates relevant outside information
- Uses relevant facts, specific examples and details, but discussion may be more descriptive than analytical
- Is a well-developed essay, demonstrating a logical and clear plan of organization
- Introduces the theme by establishing a framework that is beyond a simple restatement of the *Task* or *Historical Context* and concludes with a summation of the theme

3

- Addresses most aspects of the *Task* or addresses all aspects of the *Task* in a limited way. May discuss only two of the tasks; i.e., advantages of industrialization, disadvantages of industrialization, or how specific groups were affected by industrialization
- Uses or refers to some of the documents in the body of the essay
- Incorporates little or no relevant outside information
- Includes some facts, examples, and details, but discussion is more descriptive than analytical
- Is a satisfactorily developed essay, demonstrating a general plan of organization. Essay may not distinguish between advantages and disadvantages of industrialization or specific groups
- Introduces the theme by repeating the *Task* or *Historical Context* and concludes by simply repeating the theme

2

- Attempts to address some aspects of the *Task*, such as only discussing disadvantages of industrialization
- Makes limited use of the documents—discussion may only restate contents of documents
- Presents no relevant outside information
- Includes few facts, examples and details; discussion may contain some inaccuracies
- Is a poorly organized essay, lacking focus; could contain digression or extraneous information
- Fails to introduce or summarize the theme

1

- Shows limited understanding of the *Task* with vague, unclear references to the documents
- Presents no relevant outside information
- Includes little or no accurate or relevant facts, details, or examples
- Attempts to complete the *Task*, but demonstrates a major weakness in organization
- Fails to introduce or summarize the theme

0

- Fails to address the *Task*, is illegible, or is a blank paper

Ohio

Ohio Performance Assessment Pilot Project Physics (spring/late) Performance Task, Released 2009

How Things Work Student Materials

Introduction

Our world depends upon devices, gadgets, and instruments which impact how we live our lives. Each of these devices, gadgets, or instruments is based on scientific principles and concepts that can be used to explain the physics of everyday life. Think about all the devices, gadgets, and instruments you use every day. Have you ever really thought about how they work?

Your Task

You are a member of an engineering team and you will be using reverse engineering to determine how something works. Reverse engineering is the process of discovering the technological principles of a device, gadget, or instrument by analyzing its structure, function, and how it operates. It involves taking something apart and analyzing its components to figure out how each part interacts to successfully operate. In this task, your team will select a device, gadget, or instrument that you would like to study based on some criteria provided by the Project Manager (your teacher). Your team will write a description of your selected device, gadget, or instrument without stating its name to see if your classmates can identify what you are investigating. After some initial research, your team will deconstruct your object to determine how it works. In your explanation of how the object works, you will need to include an analysis of at least one energy transformation that occurs when your object is in operation and explain how you can apply at least three physics principles you have learned this year to your object. Your team will research an earlier version of a device, gadget, or instrument that was used to fulfill the same basic purpose and compare the characteristics, type of technology advancements, and physics principles utilized by the two objects (historical and current day). Finally, each member of the team will write their own research paper describing what you have learned about how your device, gadget, or instrument works.

Task Overview

Task Part	What You Need to Do	Product
1	Select a device, gadget, or instrument for investigation	Research paper
2	Write a 100-word essay describing your object	
3	Deconstruct, label, and explain how the components of your object interact to make it work	
4	Conduct research to learn how your object works and explain at least one energy transformation that occurs when your object is operating	
5	Identify at least three physics principles and/or concepts that helps to explain how your object works	
6	Compare your object with an earlier device, gadget, or instrument that was used to serve the same purpose	
7	Write an individual research paper using what you learned in Parts 2–6	
8	Reflect on your learning	Essay
9	Group presentation	Optional*

*Your teacher will decide whether you will be doing this portion of the performance assessment task.

Ohio

Ohio Performance Assessment Pilot Project Physics (spring/late) Performance Task, Released 2009

Part 1: Select a device, gadget, or instrument that you would like to investigate (**Team Activity**). As a team, select an object that is a system of components. An excellent object will enable you to:

- Learn about the object by taking it apart and analyzing its operations (reverse engineering)
- Remain safe working with the object, and it meets the approval of your teacher
- Identify at least one type of energy transformation when your object is in normal operation
- Make relevant connections to at least three physics principles and/or concepts

Part 2: Describe your device, gadget, or instrument to the class without naming it (**Individual Activity**). Write a 100-word essay describing your device, gadget, or instrument without naming it. You want to completely explain the appearance of your object, its function or purpose, how it operates, and why people use it. Your challenge is to see if your classmates can guess your device, gadget, or instrument based on your description.

Part 3: Deconstruct and label the components of your device, gadget, or instrument (**Team Activity**). Using tools provided by your teacher, take your object apart to examine the parts. Identify all the components of your object. Attach the components to a poster board and label each part.

Part 4: Research a variety of sources (experts, Internet, textbook, manufacturing manuals, etc.) to find out everything there is to know about the operation of your device, gadget, or instrument (**Team Activity**). While researching the object:

- Determine the purpose and/or function of each component or subassembly
- Describe how the components work together
- Explain the sequence of events that occur in order for the object to work
- Explain how energy is transformed when the object operates. Specifically, determine where the energy originates and what happens to the energy

Part 5: Connect and describe at least three physics principles and/or concepts to the operation of your device, gadget, or instrument (**Individual Activity**). Describe the physics principle and/or concept that helps to explain how and why your object works. Imagine you are explaining this to a peer who is not currently taking a physics course. You want to fully explain any vocabulary that you use so they will understand your explanation.

Part 6: Identify and research about a previously used device, gadget, or instrument that was used to serve the same general purpose (**Team Activity**). Compare the earlier version with your object, specifically:

- Describe the characteristics of each system
- Identify any unique components and the function of each component or subassembly

(continued)

- List the sequence of events that occur in order for each object to work
- Explain how energy is transformed when each object operates
- Describe any changes in technology and/or materials used in the objects
- Compare and explain the physics principles and/or concepts used in each system

Ohio

Ohio Performance Assessment Pilot Project

Physics (spring/late) Performance Task, Released 2009

Part 7: Write an individual research paper explaining everything you learned about your device, gadget, or instrument (**Individual Activity**). This report will include a summary of all the information that you researched and discovered in Parts 2 to 6. Specifically:

- Include a brief description of your device, gadget, or instrument (Part 2)
- Provide a diagram, picture, or photo of your deconstructed object with all the components and/or subassemblies labeled (Part 3)
- Explain how your object works and the transformation of energy that occurs when your object is operating (Part 4)
- Apply at least three physics principles and/or concepts to explain how and why your object works (Part 5)
- Compare the systems of an earlier version with your object to describe changes in technology and use of materials (Part 6)
- Check any written materials and visuals to ensure that you have used proper vocabulary and proper scientific conventions
- Cite all of your references using the format selected by your teacher

Part 8: Reflect on your learning (**Individual Activity**). Write an essay reflecting on your learning over the course of completing this performance assessment, specifically explaining what you:

- a) Learned about how your device, gadget, or instrument works and how it applies to at least three physics principles and/or concepts
- b) Discovered about energy transformations
- c) Used as strategies for learning, thinking, and producing work that were effective and those that did not work so well
- d) Learned about investigative skills and/or your understanding of scientific inquiry
- e) Contributed to your group work, the strengths of your team, and how the interactions within your group could be improved in the future

Part 9: Present your findings—**OPTIONAL (Team Activity)**. You will be asked to make an oral presentation on what you learned about your object through this investigation. When preparing your presentation:

- a) Consider the audience, estimate their current knowledge of the topic, and prepare your materials so they can understand your findings
- b) Provide a clear overview of your investigation (purpose, procedures, analysis, and findings) so that it has an impact on the audience and will help them to learn about your investigation
- c) Display the data using appropriate graphs, tables, visuals, etc.
- d) Check any written materials and visuals to ensure that you have used proper formulas and proper scientific convention
- e) Cite all of your references using the format selected by your teacher

Vermont
Information Technology Performance Assessment Task

Grade Cluster 9-12 Task # 5	
The Grade Cluster & Task #:	Grade Cluster 9-12, Task # 5
The Product:	Graphs and Charts to support a conclusion
The Components:	Spreadsheet, Graphing Calculator
The Task:	<p>Students will create a simulation or model using a graphing calculator. They will use digital tools to capture this data. (e.g., light from a fluorescent bulb, temperature from a variety of liquids over time, etc.) They will then analyze the data and import it into a spreadsheet. Students will manipulate the data appropriately, construct charts to justify their conclusions, and report the results visually.</p> <p>Rationale: Real-world scientific process is best accomplished using real-world tools. With the ability of technological tools to capture and manipulate data, students will focus on the understanding rather than the recording of data. Once the results are measured and scenarios understood, the learner will engage in presenting their findings. True learning occurs when explanation of that learning is evident.</p>
IT1 - Basic Operations & Concepts	
<ul style="list-style-type: none"> Using digital tools to capture images <u>and other information</u> (e.g., temperature, light, sound, etc.) and import them into a computer. 	
IT2 - Social, Ethical & Human Issues	
<ul style="list-style-type: none"> Not assessed in this task. 	
IT3 - Productivity Tools	
<ul style="list-style-type: none"> Creating a spreadsheet from a blank page, including formulas and functions (MIN, MAX, ROUND), formatting cells (e.g., numeric, monetary, percent, values). <u>Documenting spreadsheets with named cells and comments</u> Creating a graphical representation appropriate to the numerical data (e.g., scatter plot, x-y) Manipulating format (e.g., resizing rows and columns, font, colors, hiding grid) <u>Referencing formulas from other worksheets</u> Using a graphing calculator and grade appropriate applications/ functions (e.g., graphing, statistics, tables, equations, matrix). 	
IT4 - Communication	
<ul style="list-style-type: none"> Not assessed in this task. 	
IT5 - Research, Problem Solving & Decision Making	
<ul style="list-style-type: none"> Justifying decisions made, e.g. representing data, formatting, setting up formula, selecting criteria for search Creating and using simulations or models, (e.g., spreadsheet to design “what if” scenarios) 	

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.

¹ This directions page guides students towards the "proficient" level (level "3") for this CBA. To help students reach "excellent" (level "4"), please refer to the rubric or, if available, the graphic organizer.

² Students may do a paper or presentation in response to the CBA provided that for either format, there is documentation of this response that someone outside their classroom could easily understand and review using the rubric (e.g., a videotaped presentation, an electronic written document).

Washington
 Civics Classroom-Based Assessment
 Constitutional Issues, Released 2008

High School - Constitutional Issues CBA Rubric (Recommended for 11th Grade)

GLE (EALR)	PASSING		NOT PASSING	
	4 - Excellent	3 - Proficient	2 - Partial	1 - Minimal
1.4.1. Analyzes and evaluates ways of influencing local, state, and national governments to preserve individual rights and promote the common good. (11 th Grade) (EALR 1.4. Understands civic involvement)	States a position on the issue that <ul style="list-style-type: none"> Includes a proposal for balancing individual rights and the common good AND Includes an analysis of how to advocate for this position. 	States a position on the issue that <ul style="list-style-type: none"> Evaluates or considers the interaction between individual rights and the common good AND Includes an analysis of how to advocate for this position. 	States a position on the issue that <ul style="list-style-type: none"> Evaluates or considers the interaction between individual rights and the common good but does NOT include an analysis of how to advocate for this position. 	States a position on the issue that addresses individual rights OR the common good.
1.1.1. Analyzes and evaluates the ways in which the US Constitution and other fundamental documents promote key ideals and principles. (11 th Grade) (EALR 1.1. Understands key ideals and principles...)	Provides reason(s) for the position supported by evidence. The evidence includes: <ul style="list-style-type: none"> An analysis of how the Constitution promotes two or more specific ideals or principles logically connected to the issue. 	Provides reason(s) for the position supported by evidence. The evidence includes: <ul style="list-style-type: none"> An analysis of how the Constitution promotes one specific ideal or principle logically connected to the issue. 	Provides reason(s) for the position supported by evidence. The evidence includes: <ul style="list-style-type: none"> A reference to the Constitution WITHOUT an analysis of how the Constitution logically connects to the issue. 	Provides reason(s) for the position. The evidence includes: <ul style="list-style-type: none"> A reference to the Constitution that is partial or unclear.
1.1.2. Evaluates how well court decisions and government policies have upheld democratic ideals and principles in the United States. (11 th Grade) (EALR 1.1. Understands key ideals and principles...)	The evidence for the position includes: <ul style="list-style-type: none"> A detailed evaluation of how well a court case OR a government policy upheld a constitutional principle related to the issue, including: A discussion of competing viewpoints related to the case or policy. 	The evidence for the position includes: <ul style="list-style-type: none"> An evaluation of how well a court case OR a government policy upheld a constitutional principle related to the issue. 	The evidence for the position includes: <ul style="list-style-type: none"> A description of a court case or a government policy WITHOUT an evaluation of how it upheld constitutional principles related to the issue. 	The evidence for the position includes: <ul style="list-style-type: none"> A description of a court case or a government policy that is partial or unclear.
5.4.1. Evaluates and interprets other points of view on an issue within a paper or presentation. (EALR 5.4. Creates a product...)	The evidence for the position includes: <ul style="list-style-type: none"> A fair interpretation and a refutation of a position on the issue that contrasts with the student's own. 	The evidence for the position includes: <ul style="list-style-type: none"> A fair interpretation of a position on the issue that contrasts with the student's own. 	The evidence for the position includes: <ul style="list-style-type: none"> A description of a position(s) on the issue with no evaluation. 	The evidence for the position includes: <ul style="list-style-type: none"> A description of another position on the issue that is partial or unclear.
5.4.2. Creates strategies to avoid plagiarism and respects intellectual property when developing a paper or presentation. (10 th Grade) (EALR 5.4. Creates a product...) 5.2.2. Evaluates the validity, reliability, and credibility of sources while researching an issue or event. (EALR 5.2. Uses inquiry-based research...)	<ul style="list-style-type: none"> Makes explicit references within the paper or presentation to four or more credible sources that provide relevant information. Cites sources within the paper, presentation, or bibliography. 	<ul style="list-style-type: none"> Makes explicit references within the paper or presentation to three credible sources that provide relevant information. Cites sources within the paper, presentation, or bibliography. 	<ul style="list-style-type: none"> Makes explicit references within the paper or presentation to two credible sources that provide relevant information. Cites sources within the paper, presentation, or bibliography. 	<ul style="list-style-type: none"> Makes explicit references within the paper or presentation to one credible source that provides relevant information. Cites the source within the paper, presentation, or bibliography.

* OSPI recommends that this CBA be used at a particular grade level and thus, the GLEs included in the rubric are for that grade. However, if the CBA is used at another grade level within the grade band (3-5, 6-8, or 9-12), the GLEs may need to change to match the appropriate content.
 ** Please also refer to the document "Scoring Notes for Secondary Social Studies CBAs" when evaluating student work.

High School
Recommended
for 12th Grade

You and the Economy CBA

Responsible citizenship requires careful consideration of the role people play in the local, national, and global economy. Evaluate the career choices available to you, and their possible effects on your local, national, and international economy, as well as yourself.

Directions to students¹

In a cohesive paper or presentation², you will:

- State a position on which career choices would be best for you.
- Provide reasons for your position that include:
 - An analysis of how your career choices will affect the local, national, and/or global economy with one or more examples.
 - An analysis of how the economic system may affect your economic choices with two or more examples.
- 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.

¹ This directions page guides students towards the "proficient" level (level "3") for this CBA. To help students reach "excellent" (level "4"), please refer to the rubric or, if available, the graphic organizer.

² Students may do a paper or presentation in response to the CBA provided that for either format, there is documentation of this response that someone outside their classroom could easily understand and review using the rubric (e.g., a videotaped presentation, an electronic written document).

Washington
Economics Classroom-Based Assessment
You and the Economy, Released 2008

High School – You and the Economy CBA Rubric (Recommended for 12 th Grade ¹)				
GLE (EALR)	PASSING		NOT PASSING	
	4 – Excellent	3 – Proficient	2 – Partial	1 – Minimal
5.4.1 Evaluates positions and evidence to make own decisions in a paper or presentation. 5.1.1 Analyzes the short-term and long-term implications of decisions made affecting the global community.	States a position on which career choices would be best for the student that includes: <ul style="list-style-type: none"> An analysis of what the student's career options tell us about the global economy. 	States a position on which career choices would be best for the student.	States a position on career choices WITHOUT indicating which would be best for the student.	Describes career choices without taking a position.
2.1.1 Analyzes how economic choices made by groups and individuals in the global economy can impose costs and provide benefits. 5.2.1 Evaluates the plausibility of an analysis of implications of decisions for the global community.	Provides reason(s) for the position supported by evidence. The evidence includes: <ul style="list-style-type: none"> An analysis of how the student's career choices will affect the local, national, and/or global economy with two or more examples. 	Provides reason(s) for the position supported by evidence. The evidence includes: <ul style="list-style-type: none"> An analysis of how the student's career choices will affect the local, national, and/or global economy with one example. 	Provides reason(s) for the position supported by evidence. The evidence includes: <ul style="list-style-type: none"> An analysis of how the student's career choices will affect the local, national, and/or global economy WITHOUT specific examples. 	Mentions how the student's career choices will affect the local, national, and/or global economy WITHOUT an analysis.
2.2.1 Analyzes and evaluates the advantages and disadvantages of different economic systems for countries and groups of people. 2.4.1 Analyzes and evaluates how individuals affect and are affected by the distribution of resources and sustainability.	The evidence for the position includes: <ul style="list-style-type: none"> An analysis of how the economic system may affect the student's economic choices with two or more examples. AND <ul style="list-style-type: none"> An analysis of how the distribution of resources or issues of sustainability in the global economy affect the student's career choices. 	The evidence for the position includes: <ul style="list-style-type: none"> An analysis of how the economic system may affect the student's economic choices with two or more examples. 	The evidence for the position includes: <ul style="list-style-type: none"> An analysis of how the economic system may affect the student's economic choices with one example. 	The evidence for the position includes: <ul style="list-style-type: none"> An analysis of how the economic system may affect the student's economic choices WITHOUT specific examples.
5.4.2. Creates strategies to avoid plagiarism and respects intellectual property when developing a paper or presentation. (10th Grade) (EALR 5.4. Creates a product...) 5.2.2 Evaluates the validity, reliability, and credibility of sources while researching an issue or event. (EALR 5.2. Uses inquiry-based research.)	<ul style="list-style-type: none"> Makes explicit references within the paper or presentation to four or more credible sources that provide relevant information. Cites sources within the paper, presentation, or bibliography. 	<ul style="list-style-type: none"> Makes explicit references within the paper or presentation to three credible sources that provide relevant information. Cites sources within the paper, presentation, or bibliography. 	<ul style="list-style-type: none"> Makes explicit references within the paper or presentation to two credible sources that provide relevant information. Cites sources within the paper, presentation, or bibliography. 	<ul style="list-style-type: none"> Makes explicit references within the paper or presentation to one credible source that provides relevant information. Cites the source within the paper, presentation, or bibliography.

¹OSPI recommends that this CBA be used at a particular grade level and thus, the GLEs included in the rubric are for that grade. However, if the CBA is used at another grade level within the grade band (3-5, 6-8, or 9-12), the GLEs may need to change to match the appropriate content.
²Please also refer to the document "Scoring Notes for Secondary Social Studies CBAs" when evaluating student work.

A Letter to the Publisher

Student Copy

Name _____ Period _____

Date _____ Teacher _____

School _____

Total Score / 8

Citing concerns over some of the advertising found in the school's magazine collection, your school librarian has asked you to join a committee of students to examine the issue. As part of your participation in this committee, you will analyze four magazine advertisements whose messages concern you and then write a letter to their publisher. The results of this task should persuade the publisher to include advertisements with fewer negative messages in the magazines.

Part 1: Advertising and Negative Effects

Pre-Writing

Pre-writing—will not be scored.

For each of the four advertisements, provide one negative effect on teens. You must list four *different* negative effects. Next, identify the part of the advertisement that leads to the negative effect.

Negative Effect on Teens	Negative Part of Advertisement
1.	1.
2.	2.
3.	3.
4.	4.

Washington
Health and Fitness Classroom-Based Assessment
Letter to the Editor, Released 2009

Part 2: Advertising and Positive Effects

Pre-writing—will not be scored.

For each of the four advertisements, describe **one change** that would result in a positive effect on teens rather than a negative effect. Next, describe how each change would result in the **intended positive effect**.

Change to Advertisement	Intended Positive Effect
1.	1.
2.	2.
3.	3.
4.	4.

Washington
Health and Fitness Classroom-Based Assessment
Letter to the Editor, Released 2009

Score / 8

Washington
 Health and Fitness Classroom-Based Assessment
 Letter to the Editor, Released 2009

Rubric 1– Used to score Part 1 of A Letter to the Publisher

(EALR 3) The student analyzes and evaluates the impact of real-life influences on health.

4	<p>4-point response: The student accumulates a total of 7–8 value points.</p> <p>The student:</p> <ul style="list-style-type: none"> • Provides one negative effect that each of four advertisements may have on teens (1 point per effect = 4 possible value points) • Identifies one part of each advertisement that may lead to each of the four negative effects or negative health problems (1 point per example = 4 possible value points). <p>Example:</p> <p>One advertisement shows a thin female teenager who is in the company of several thin muscular teenage males. The body image, or images, that are represented in this ad may lead to eating disorders such as bulimia.</p>
3	3-point response: The student earns 5-6 value points.
2	2-point response: The student earns 3-4 value points.
1	1-point response: The student earns 1-2 value points.
0	0-point response: The student shows little or no understanding of the task.

Rubric 2– Used to score Part 2 of A Letter to the Publisher

(EALR 3) The student analyzes and evaluates the impact of real-life influences on health.

4	<p>4-point response: The student accumulates a total of 7–8 value points.</p> <p>The student:</p> <ul style="list-style-type: none"> • Suggests one change for each of the four advertisements that may result in a positive effect on teens (1 point per change = 4 possible value points) • Describes how each of the four changes would result in the intended positive effect (1 point per description = 4 possible value points). <p>Example:</p> <p>An advertisement that previously showed only thin teenagers wearing a particular brand of jeans, having fun at a party, may be changed into teenagers with an assortment of body types having fun at a party while wearing a particular brand of jeans. This advertisement may communicate the message that teenagers of all body types can have fun.</p>
3	3-point response: The student earns 5-6 value points.
2	2-point response: The student earns 3-4 value points.
1	1-point response: The student earns 1-2 value points.
0	0-point response: The student shows little or no understanding of the task.

Washington

Visual Arts Classroom-Based Performance Assessment

A Zoo Mug, Released 2008

A local zoo is accepting proposals for a novelty mug promoting the zoo. The zoo desires that these mugs be fun, functional, and decorative. The zoo has asked high school art students to submit an actual prototype for the mugs containing a functional handle, at least two levels of sculptural relief, and two or more decorative textures.

The zoo requests that each artist also submit a detailed planning pencil sketch for the mug prototype. The theme of your mug design must represent zoo animals and environments without the use of words or typography.

The zoo staff explains that you must meet the following task requirements when creating your zoo mug prototype:

Creating

- The overall design of the mug must be in a theme that promotes the zoo. (Examples: mammals, reptiles, insects, birds, natural environments)
- Prior to beginning the actual **functional** piece or work, you must create at least one sketch of your mug, indicate by using arrows and labels the **theme**, the handle, and the **sculptural** and **relief components**.
- In the actual functional mug include three levels of relief that enhance the overall zoo theme of the mug. Two levels of relief should project off the surface of the mug through forming processes such as **appliqué, modeling**, and carving and protrude into the surface or background through forming techniques such as carving, stamping and impressing.
- In the actual functional mug, include a variety of two or more **textures** in addition to the smooth or flat surface of the mug.
- Use at least two different forming methods, such as **pinching, coiling, extruding, slab building, or throwing**.

Performing

- Create a **functional** and **decorative** standard-size mug, approximately 4" in height and 3" in diameter, with a 2½" handle, made out of clay, which promotes the zoo.
- Select a primary forming method deliberate to your design such as pinching, coiling, extruding, slab building, or throwing.
- Walls of the mug should be in **proportion** to its size and be even throughout.
- Handle must be in **proportion** to the size and thickness of the mug. The handle and lip must be smooth and comfortable for the user.
- Construction seams (sides, bottom, handle) are crafted so that the mug is functional and will not leak.

Washington

Visual Arts Classroom-Based Performance Assessment

A Zoo Mug, Released 2008

Planning/Sketch of Mug

- 1) Draw a detailed sketch of your Zoo Mug design. You may draw a cutaway or silhouette view. Your plan may be a series of sketches showing your mug from different views.
- 2) By using arrows and labels, indicate at least two ways your mug design supports the zoo theme.
- 3) By using arrows and labels, indicate the handle and the sculptural and relief components as part of your sketch.



Washington

Visual Arts Classroom-Based Performance Assessment

A Zoo Mug, Released 2008

Response Sheet

The zoo staff explains that you must also meet the following task requirements when responding about your mug:

1. Name the forming processes and explain how you used them to create two levels of relief

First level of relief

Name of technique: _____

How you used the technique:

Second level of relief

Name of technique: _____

How you used the technique: _____

2. Identify at least two specific forming methods you used to construct the mug, such as pinch, coil, slab/drape, and throw. Explain why you chose each method for specific parts of the mug.

Method one: _____

Explanation: _____

Method two: _____

Explanation: _____

3. Explain in detail how your mug design is both functional and decorative.

Functional _____

Decorative _____

1. Identify two textures on your mug and what techniques you used to create the two textures.

Texture one:

Where: _____

What technique: _____

Washington

Visual Arts Classroom-Based Performance Assessment

A Zoo Mug, Released 2008

Texture two:

Where: _____

What technique: _____

5. Explain two ways your planning pencil sketch influenced your work as you constructed your mug.

First way: _____

Second way: _____

Washington

Visual Arts Classroom-Based Performance Assessment

A Zoo Mug, Released 2008

Scoring Guide Grade 10 Visual Arts A Zoo Mug

Creating Rubric (1.1.1, 1.1.2, 2.1, 2.2, 3.1, 3.2, and 4.5)

4	A 4-point response: The student demonstrates an understanding of the creative process by successfully meeting all of the four task requirements below: The sketch of the mug includes the theme, the handle, and the sculptural and relief components The overall design of the mug is completed in a theme that promotes the zoo (examples: mammals, reptiles, insects, birds, natural environments) The actual mug includes two levels of relief that enhance the overall zoo theme of the mug The actual mug includes a variety of two or more textures
3	A 3-point response: The student demonstrates an adequate understanding of the creative process by successfully meeting three of the four task requirements listed above.
2	A 4-point response: The student demonstrates a partial understanding of the creative process by successfully meeting two of the four task requirements listed above.
1	A 4-point response: The student demonstrates a minimal understanding of the creative process by meeting one of the four task requirements listed above.
0	A 4-point response: The student demonstrates no understanding of the creative process by meeting none of the four task requirements listed above.

Performing Rubric: Demonstration of Hand-Building Skills (1.1.1, 1.1.2, 2.1, 2.2, 3.2, and 4.5)

4	A 4-point response: The student demonstrates mastery of hand-building skills by successfully meeting all of the four task requirements below: Used at least two different forming methods to create three distinct levels of relief Walls of the mug are in proportion to its size and are even throughout Handle is in proportion to the size and thickness of the mug; the handle and lip are smooth and comfortable for the user Construction seams (sides, bottom, handle) are crafted so that the mug is functional and will not leak
3	A 3-point response: The student demonstrates adequate mastery of hand-building skills by successfully meeting three of the four task requirements listed above.
2	A 2-point response: The student demonstrates partial mastery of hand-building skills by successfully meeting two of the four task requirements listed above.
1	A 1-point response: The student demonstrates minimal hand-building skills by successfully meeting one of the four task requirements listed above.
0	A 0-point response: The student demonstrates no understanding of hand-building and meets none of the task requirements listed above.

Washington

Visual Arts Classroom-Based Performance Assessment

A Zoo Mug, Released 2008

Note: EALRs 3 and 4 are naturally and authentically embedded in the prompts and rubrics of this assessment, even when not specifically measured.

Responding Rubric (2.1, 2.2, 2.3, 3.2, and 4.5)

4	A 4-point response: The student demonstrates effective communication in an artistic response process by successfully completing four or five of the requirements listed below: <ul style="list-style-type: none">• Name the forming processes and explain how you used them to create two levels of relief• Identify two specific forming methods used to construct the mug; explain why you chose each method for specific parts of the mug• Explain how the mug design is both functional and decorative• Describe where and what techniques were used to create two textures on your mug excluding the smooth or flat surface of the mug• Explain two ways the detailed pencil sketch influenced the work as the mug was constructed
3	A 3-point response: The student demonstrates effective communication in an artistic response process by successfully completing three of the five requirements listed above.
2	A 2-point response: The student demonstrates effective communication in an artistic response process by successfully completing two of the five requirements listed above.
1	A 1-point response: The student demonstrates effective communication in an artistic response process by successfully completing one of the five requirements listed above.
0	A 0-point response: The student meets none of the requirements listed above.

Scoring Notes

- This assessment is best done after a significant amount of technique is covered, taught, and practiced in making functional and sculptural containers.
- Consider allowing time during the assessment window for drying, to enable students to work with the clay in a leather hard state for techniques such as carving or adding slab appliqué.
- Consider the atmosphere for drying as it will affect the assessment time and product outcome.
- Attach a photograph of the zoo mug (greenware or bisque ware) next to the sketch for ease of scoring.
- Teacher(s) or scorers may develop and use their own scoring tool for ease of scoring multiples assessments.
- Handle does not include a level of relief unless it is an obvious sculptural component such as using the trunk of an elephant as a handle.
- Incised lines count as a texture only, and not a level of relief.

Washington

Visual Arts Classroom-Based Performance Assessment Snack Time, Released 2008

Your school newspaper editor is seeking to showcase a photograph for a feature article about teen food choices. You are a photographer for your school newspaper. The editor has requested that you photograph a food item or group of items that interest teenagers. The selected photograph must be in sharp focus and draw the viewer into the image. The photograph also must display a strong focal point and rule of thirds, and use shallow depth of field. In addition you are required to show a contrast of light and dark values that will emphasize the food item or items to the target audience.

The newspaper editor requires the following elements in your photograph of a food item or items:

The subject in sharp **focus**

- Use of the **rule of thirds** in composing the focal point
- A demonstration of shallow **depth of field**
- A range of value; cast **shadows and highlights** and/or reflection through use of directional lighting techniques
- A printed color or black and white image printed/developed on 4" x 6" paper or larger

After you have completed your photograph, the newspaper editor requires you to:

- Describe in detail how you used and/or created:
- Rule of thirds to emphasize a focal point
- Shallow depth of field
- Give one example of how you used directional studio lighting techniques or computer software techniques/tools to create contrast and a range of values for cast shadows, highlights, and/or reflections
- Describe in detail the food item(s) used and why you placed the item or items in that arrangement
- Use photographic/compositional visual arts vocabulary correctly

The art editor has allowed you time to complete the photographic composition. You will have 20–30 minutes to complete your written response.

Washington

Visual Arts Classroom-Based Performance Assessment Snack Time, Released 2008

Response Sheet

You are expected to use design/art and photographic vocabulary correctly in your written responses. As you describe the use of techniques in your composition, it is important to refer directly to what is in your actual photograph.

1. Give one example of how you used directional studio lighting techniques or computer software technique/tools to create contrast and/or a range of values for cast shadows, highlights, and/or reflections.

2. Describe in detail how you used the rule of thirds to emphasize the focal point.

3. Describe in detail how you used a shallow depth of field.

Describe in detail the item(s) and why you placed the item(s) in that arrangement.

Washington
 Visual Arts Classroom-Based Performance Assessment
 Snack Time, Released 2008

Scoring Guide
 Grade 10 Visual Arts
 Snack Time
 2008

Creating Rubric: Elements and Principles Rubric (1.1.1, 1.1.2, 2.1, 2.2, 3.1, 3.2, and 4.5)

4	A 4-point response: The student combines all <u>four</u> of the following photographic design elements and principles on 4" x 6" paper or larger to depict their food item(s): Demonstrates sharp photographic focus on the focal point of the food item(s) Uses rule of thirds to emphasize the focal point on the food item(s) Creates shallow depth of field to emphasize the focal point Uses directional studio lighting and/or computer software techniques/tools to create contrast and a range of values for cast shadows, highlights and/or reflections
3	A 3-point response: The student combines three of the four photographic design elements and principles listed above.
2	A 2-point response: The student combines two of the four photographic design elements and principles listed above.
1	A 1-point response: The student combines one of the four photographic design elements and principles listed above.
0	A 0-point response: The student combines none of the photographic design elements and principles listed above.

Responding Rubric (2.1, 2.2, 2.3, 3.2, and 4.5)

4	A 4-point response: The student describes in detail all four points using photographic/compositional visual arts vocabulary correctly: Gives an example of how directional studio lighting and/or computer software techniques/tools are used to create contrast and a range of values for deep shadows, contrast, and/or reflection Describes how the rule of thirds is used to create a focal point Describes how a shallow depth of field is used and/or created Describes in detail the food item(s) used and why they were placed in that arrangement
3	A 3-point response: The student describes in detail three of the four points listed above.
2	A 2-point response: The student describes in detail two of the four points listed above.
1	A 1-point response: The student describes in detail one of the four points listed above.
0	A 0-point response: The student describes in detail none of the points listed above.

Reading Work Sample Task Form

Instructions: Please fill in the following sections using the corresponding information listed on the Work Sample Documentation Form (WSDF). The purpose of the Work Sample Task Form is to confirm the information previously stated on the WSDF, to provide a consistent template for describing your instructions to the student and to help the scorers identify the tasks in the binders.

Work Sample #: _____

On demand

Title and Author of Text: **Silk: Caterpillar Thread**

Write the target for each strand in the task in the boxes below.

	Literary Comprehension	Literary Analysis	Literary Thinking Critically
<input type="checkbox"/> Literary Text			

OR

	Informational Comprehension	Informational Analysis	Informational Thinking Critically
<input checked="" type="checkbox"/> Informational Text	IC13	IA17	IT19

Reading Task Instructions (If student work is on the same page as the task instructions, you do not need to complete this box):

Reading responses are expected to be written in complete sentence and paragraphs.

In paragraphs, answer each of the following in a reading response to "Silk: Caterpillar Thread."

1. (IC13) Infer why silk is such a valuable fabric. Use text based evidence to support your response.
2. (IA17) What effects did Silk Road have on the world? Use text based evidence to support your response.
3. (IT19) Critique China's decision to keep their silk production a secret. Use text based evidence to support your response.

Washington
Collection of Evidence
Reading Prompt, Released 2008

Informational Rubric

	Evidence of Skill			
	4	3	2	1
Informational Comprehension				
IC11 Main Idea	Discusses a main idea and uses supporting details from throughout the text to demonstrate an overall understanding	Describes a main idea and uses supporting details from the text	Identifies an idea and includes a detail about the text	States an idea or a detail about the subject
IC12 Summary	Summarizes by providing an overarching statement that connects details, facts and information from throughout the text	Summarizes by including information from throughout the text	Retells by including details, facts, or information from the text	States a detail, a fact, or information about the subject
IC13 Inference / Prediction	Infers and/or predicts about the subject to explain connections and demonstrate understanding of the text	Infers and/or predicts about a subject using details to support the inference	Identifies an inference and/or a prediction about the text	States information to suggest an inference or prediction about the subject
IC14 Informational Vocabulary	Discusses critical vocabulary by explaining its meaning and how it contributes to the overall context of the text	Describes critical vocabulary to demonstrate understanding of the text	Defines critical vocabulary from the text	States information about the vocabulary or the subject
Informational Analysis				
IA15 Text Features	Analyzes text features to demonstrate understanding of the text	Explains text features to support understanding of the text	Connects text features to the text	States a text feature about the subject
IA16 Compare / Contrast	Analyzes similarities and/or differences using examples to demonstrate understanding of relationships in the text	Explains similarities and/or differences to show understanding of ideas in a text	Determines similarities and/or differences in the text	States a similarity and/or difference about the subject
IA17 Cause / Effect	Analyzes text to interpret the relationship between cause(s) and effect(s) and make connections to demonstrate understanding of the text	Explains cause and effect relationship to demonstrate understanding of the text	Determines cause and effect relationship in the text	States cause(s) OR effect(s) about the subject
Informational Evaluation				
IT18 Author's / Text's Purpose	Evaluates author's purpose and/or intended audience in order to judge the effectiveness of the text	Explains author's purpose and/or intended audience to demonstrate understanding of the text	Identifies author's purpose and/or intended audience in the text	States a purpose and/or audience
IT19 Evaluation	Evaluates reasoning of ideas/themes and makes judgments to demonstrate understanding of the overall context of the text	Explains reasoning of ideas/themes to demonstrate understanding of the text	Identifies an opinion OR general statement about the subject and provides reasoning from the text	States an opinion OR general statement about the subject
IT20 Extends Beyond the Text	Extends information beyond the text by explaining generalizations OR drawing conclusions and connects to larger concepts/ideas	Extends information beyond the text by explaining generalizations OR drawing conclusions to demonstrate understanding of the text	Makes a general extending statement about the subject by using information from the text	Makes a general extending statement about the subject

1

Revised 2/26/09

Writing Work Sample Task Form

Instructions: Please fill in the following sections using the corresponding information listed on the Work Sample Documentation Form (WSDF). The purpose of the Work Sample Prompt Form is to confirm the information previously stated on the WSDF, to provide a consistent template for describing your instructions to the student and to help the scorers identify the tasks in the binders.

Work Sample #: _____

___ On Demand

Work Sample Title: Unrecognized Heroes _____

Purpose: Expository Persuasive

Writing Task Instructions; (if student work is on the same page as the task instructions, you do not need to complete this box.)

Task or classroom assignment:

In your history (or social studies) class, you notice that your teacher has not taught you and your classmates about heroes from cultures under-represented in history (or social studies).

Write a multiple-paragraph letter to your teacher persuading him or her to include the contributions of these heroes as part of your course work.

***High School
 Content, Organization, & Style Scoring Guide***

Points	Description
4	<ul style="list-style-type: none"> • Maintains consistent focus on topic and has selected and relevant details • Has a logical organizational pattern and conveys a sense of completeness and wholeness • Provides transitions which clearly serve to connect ideas • Uses language effectively by exhibiting word choices that are engaging and appropriate for intended audience and purpose • Includes sentences, or phrases where appropriate, of varied length and structure • Allows the reader to sense the person behind the words
3	<ul style="list-style-type: none"> • Maintains adequate focus on the topic and has adequate supporting details • Has a logical organizational pattern and conveys a sense of wholeness and completeness, although some lapses occur • Provides adequate transitions in an attempt to connect ideas • Uses adequate language and appropriate word choices for intended audience and purpose • Includes sentences, or phrases where appropriate, that are somewhat varied in length and structure • Provides the reader with some sense of the person behind the words
2	<ul style="list-style-type: none"> • Demonstrates an inconsistent focus and includes some supporting details, but may include extraneous or loosely related material • Shows an attempt at an organizational pattern, but exhibits little sense of wholeness and completeness • Provides transitions which are weak or inconsistent • Has a limited and predictable vocabulary which may not be appropriate for the intended audience and purpose • Shows limited variety in sentence length and structure • Attempts somewhat to give the reader a sense of the person behind the words
1	<ul style="list-style-type: none"> • Demonstrates little or no focus and few supporting details which may be inconsistent or interfere with the meaning of the text • Has little evidence of an organizational pattern or any sense of wholeness and completeness • Provides transitions which are poorly utilized, or fails to provide transitions • Has a limited or inappropriate vocabulary for the intended audience and purpose • Has little or no variety in sentence length and structure • Provides the reader with little sense of the person behind the words
Z	<ul style="list-style-type: none"> • Response is "I don't know"; response is a question mark (?); response is one word; response is only the title of the prompt; or the prompt is simply recopied

High School Conventions Scoring Guide

Points	Description
2	<ul style="list-style-type: none"> • Consistently follows the rules of Standard English for grammar and usage • Consistently follows the rules of Standard English for spelling of commonly used words • Consistently follows the rules of Standard English for capitalization • Consistently follows the rules of Standard English for punctuation • Exhibits the use of complete sentences except where purposeful fragments are used for effect • Indicates paragraphs consistently
1	<ul style="list-style-type: none"> • Generally follows the rules of Standard English for grammar and usage • Generally follows the rules of Standard English for spelling of commonly used words • Generally follows the rules of Standard English for capitalization • Generally follows the rules of Standard English for punctuation • Generally exhibits the use of complete sentences except where purposeful fragments are used for effect • Indicates paragraphs for the most part
0	<ul style="list-style-type: none"> • Mostly does not follow the rules of Standard English for grammar and usage • Mostly does not follow the rules of Standard English for spelling of commonly used words • Mostly does not follow the rules of Standard English for capitalization • Mostly does not follow the rules of Standard English for punctuation • Exhibits errors in sentence structure that impede communication • Mostly does not indicate paragraphs
Z	<ul style="list-style-type: none"> • Response is "I don't know"; response is a question mark (?); response is one word; response is only the title of the prompt; or the prompt is simply recopied

Principles of Holistic Scoring:

- **Density:** We weigh the proportion of errors to the amount of writing done well. This includes the ratio of errors to length.
- **Variety:** We consider the range of errors across the categories included in the rubric (usage, grammar, spelling, capitalization, punctuation, sentence formation, and paragraphing).
- **Severity:** We weigh basic errors more heavily than higher level errors. We also weigh basic spelling and sentence formation errors more heavily.

Linda Darling-Hammond, Co-Director
Stanford University Charles E. Ducommun Professor of Education

Prudence Carter, Co-Director
*Stanford University Associate Professor of Education and
(by courtesy) Sociology*

Carol Campbell, Executive Director



Stanford Center for Opportunity Policy in Education
Barnum Center, 505 Lasuen Mall
Stanford, California 94305
Phone: 650.725.8600
scope@stanford.edu

<http://edpolicy.stanford.edu>