Performance Assessments For Learning:
The Next Generation of State Assessments

Ruth Chung Wei
Susan E. Schultz
Raymond Pecheone,
Stanford Center for Assessment, Learning, & Equity (SCALE)
Stanford University

For an updated version of this paper or more information about the Ohio Performance Assessment Pilot Project, please contact the first author (rchung@stanford.edu).

Performance Assessments For Learning: The Next Generation of State Assessments

Introduction

With the goal of re-envisioning the state's assessment and accountability system to support improvements in the rigor and quality of curriculum and instruction, to help students prepare for college, and to signal college readiness and careers, Ohio developed an assessment system that reflects a balance among teacher-generated[1] and standardized assessments. Teacher-generated, curriculum-embedded assessments reflect the ongoing curriculum and reduce the tendency to view assessment as divorced from the instructional program, while the inclusion of standardized assessments in the system provides a common metric for assessing student progress and for comparing results among schools and buildings. At the initiation of the project, the Ohio Department of Education (ODE) leadership articulated several ways in which performance assessments might fit within the Ohio assessment and accountability systems[2]:

- **As part of a system of course examinations.** Evidence from performance-based assessments will be combined with evidence from open-ended reference tests (end-of-course exams). The exams should include both open-ended on-demand tests – essays and problem solutions -- and curriculum-embedded tasks that require more extended writing, research, and inquiry. These scores are combined to create the final examination score. The tasks will be constructed by high school faculty and college faculty under the aegis of the Department of

[1] While ODE used the term "teacher-generated" in its original concept paper, SCALE prefers the use of the term "curriculum-embedded" to connote that these tasks are completed with closer proximity to instructional units of study and that these tasks are designed by experts in consultation with teachers.

[2] Since the beginning of this project, the policy framework has evolved following changes in state governance and leadership. The new policy framework undergirding the OPAPP work seems to be driven by the state's engagement in the Race To The Top federal grant program and in the state's involvement in the two national assessment consortia (SMARTER Balanced and PARCC).

Education, and they can be used to inform the grade given to the student in the course. However, no single measure should be used to make a pass/fail or graduation decision.

- **As an alternative means for students to demonstrate subject matter mastery** in lieu of

  A) Carnegie Units, per the Ohio statute authorizing such demonstration of mastery by performance;

  B) Components of the Ohio Graduation Test (OGT);

  C) Components of the ACT (if it replaces the OGT).

- **As a way to satisfy the "senior project"** component of (then) Governor Strickland's reform agenda. One format might be a single project in an area of deep interest to students. A second format might be a graduation portfolio that includes performance assessments in a minimum number of content areas (for example, three areas chosen by the student, as is common in countries that use 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 ("honors") recognizing their outstanding achievement. These endorsements could be taken into account as part of a student's application for college and/or in conjunction with a placement exam used by colleges to determine course eligibility (based on agreements made with higher education institutions).

In 2008, Ohio undertook a statewide pilot of performance-based assessments developed in partnership with the Stanford Center for Assessment, Learning, and Equity (SCALE) at Stanford University and with the input of educators and stakeholders from across the state. The purpose of this pilot project is to develop and try out performance-based assessments that are designed to both measure and promote students' learning of content and skills that will prepare them to be successful in college and in careers. Ohio's leaders had articulated a policy imperative that calls for a revision of its existing assessment system to include both tests of content knowledge and rigorous classroom based assessments and projects. The intent of this more balanced assessment system is to support the development of a challenging, relevant and rigorous curriculum that is benchmarked to national (common core) and international standards of performance, and at the same time, promotes improvements in instructional practice and enriches students' learning experiences in ways that lead to higher levels of student success in college and beyond.

The purpose of the Ohio Performance Assessment Pilot Project (OPAPP) is to contribute to the development of an assessment system that raises expectations for learning for all students, is balanced and uses a multiple measures approach to assessment and accountability. One important purpose of this assessment approach is to support improvements in instructional practice and to link student achievement with international benchmarks of student performance. The initial phase of the project was focused on building prototypes of the performance assessment instruments (performance outcomes, scoring rubrics, and tasks) that are aligned with college and career readiness expectations as well as the core content knowledge and skills of each discipline, including a set of curriculum-embedded, teacher-managed, rich performance tasks that are both content-focused and skills-driven.

Since the launch of the project in the fall of 2008, performance assessments (now known as "learning tasks") in three content areas (English language arts, mathematics, and science) were built through the involvement of Ohio educators and stakeholders across 30 schools, and piloted with 147 teachers (approximately 40-50 in each content area). The first year pilots of the "learning tasks" were completed in the spring of 2010 and provide the basis for the reliability and validity studies conducted to assess the technical quality of the tasks and scoring procedures. A second round of piloting was completed in 2010-11, with the addition of a new component of the OPAPP system -- "assessment tasks". Completed tasks from the fall 2010 pilot of "learning tasks" and from the spring 2011 "assessment tasks" were scored and analyzed.[3]

In Phase II of this project (begun in July 2011), the ODE plans to expand the performance assessment system to include history/social science and career technical education, build a task bank that will provide the foundation for ongoing work, expand the network of districts, schools, and educators in the state participating in the pilot project, and further build the capacity of local educators, administrators, coaches, and regional assistance centers to carry on this work into the future, and build a technology platform that will support the scaling-up and implementation of a performance-based assessment system.

The purpose of this paper is to 1) describe the design of the performance assessments, the assessment system, and the intended processes/protocols for implementation; 2) reflect on lessons learned from the implementation of these assessments, and 3) generate principles for performance assessment design and implementation guidelines that best support educative uses of the assessments. Results from this pilot provide important insights about the *real-world application of research-based principles of formative assessment* and how a state initiative can use research-based and educative assessment practices to *move teaching and learning forward for the common good.*

[3] Results of these reliability and validity analyses may be found in the AERA 2012 conference paper by Wei, Cor, Arshan, & Pecheone, "Can Performance Assessments Be Reliable and Valid?". The current paper focuses on the lessons learned from the design and implementation of the learning tasks only, as the assessment tasks were designed and implemented with a summative purpose in mind.

Conceptual Framework

Rationale for performance-based assessment

A growing number of business and education leaders recognize the importance of the kinds of assessments that are used to evaluate student learning. Fadel, Honey, and Pasnik (2007), for example, have suggested that the workplace of the 21st century will require "new ways to get work done, solve problems, or create new knowledge" (p.1), and that how we assess students will need to be largely performance-based in order to evaluate how well students are able to apply content knowledge to critical-thinking, problem-solving, and analytical tasks. Likewise, David Conley, in his book, *College Knowledge* (2005), reports that higher education faculty valued "habits of mind" even more than content knowledge, including the ability to think critically and analytically, to independently draw inferences and reach conclusions, and to solve problems.

More than standardized tests of content knowledge, performance-based tasks have the potential to directly measure these cognitive abilities. Performance-based assessments require students to use high level thinking to perform, create, or produce something with transferable real-world application. Research has also shown that they provide useful information about student performance to students, parents, teachers, principals, and policy-makers (Matthews, 1995; Koretz et al., 1996; Vogler, 2002).

Curriculum-Embedded Performance Assessments as "Assessment For Learning"

There is an extensive literature supporting the idea that how you assess student achievement drives the way teachers teach and what students learn (e.g., Black & Wiliam, 1998; Stiggins, 2002). Based on this research, schools engaged in high school restructuring and building new high school models have looked for alternative ways of measuring student achievement that would challenge teachers and students alike to produce work that demonstrates more than minimum competencies. These alternative assessments are designed to measure not only discipline-specific content, but also the college and career readiness skills, which are described by David Conley (2007) as Key Cognitive Strategies, and by the Partnership for 21st Century Skills as essential future workforce skills. These skills include problem solving, critical thinking, communication, collaboration, analysis, interpretation, inquisitiveness, reasoning, and accuracy and precision. Evidence from high-performing schools from across the nation supports the idea that rigorous performance assessments can support students' successful completion of high school and enrollment in college (Wasley, et al., 2000; Darling-Hammond, Ancess, & Ort, 2002; Foote, 2005; Barron & Darling-Hammond, 2008; Stevens, et al., 2008).

Black and Wiliam's (1998) extensive review of research on effective uses of assessment (those associated with large effect sizes in student achievement gains) identified several features of assessment practice that are more likely to lead to improvements in student learning, including: clearly articulating learning targets, providing actionable and specific feedback, and providing opportunities for self-assessment. Likewise, Stiggins (1994) identified several aspects of classroom assessment practice that supports student learning: clear articulation of learning targets, building assessments that can accurately reflect student learning, using assessments to help students build confidence in themselves and to allow students to take control of their own learning, providing frequent descriptive feedback, making continuous adjustments to instruction based on the results of assessment, regularly engaging students in self-assessment, and regularly communicating with students and parents about their achievement and learning.

Building a Validity Argument for Performance Based Assessment

Cronbach (1971) argues that validity has to do with the meaning or interpretation of scores, as well as consequences of score interpretation (for different persons or population groups, and across settings or contexts). Therefore, score validity is examined in light of alternative contexts for assessment use (for low-stakes formative vs. high-stakes summative purposes), as well as for different populations of students where sample sizes make this possible. Further, according to the most recent conceptions of validity, validation involves an interpretive argument that specifies the proposed and intended uses of test scores as well as a validity argument that provides evidence that the interpretive argument is coherent, that its inferences are reasonable, and that its assumptions are plausible (Kane, 2005). Messick's (1989) criteria for evaluating the validity of performance assessments include content, substantive, structural, generalizability, external, and consequential aspects of validity. Likewise, Linn, Baker, and Dunbar's (1991) criteria for building a validity argument include consequences, fairness, transfer and generalizability, cognitive complexity, content quality, and content coverage. In this paper, we focus on only two criteria for validity cited by Messick (1989): *consequential validity*; and *substantive theories, process models, and process engagement*. In another paper presented at this conference (Wei, Cor, Arshan, and Pecheone, "Can performance assessments be reliable and valid?), several other commonly defined criteria for supporting a validity argument were explored, based on the availability of data from the OPAPP pilots -- *content relevance and representativeness; scoring models as reflective of task and domain Structure; generalizability and the boundaries of score meaning; convergent and discriminant correlations with external variables; and fairness.*

Methods/Data-Sources

In this paper, we describe Ohio’s performance-based assessment system and the protocols for implementing the performance tasks, and examine the ways in which the design of the assessment system and the protocols/guidelines for implementing and using the assessments are consistent with research-based principles of "Assessment For Learning" described above. Specifically we describe the assessment design features, development processes, and protocols that were used to design the performance assessment system, provide selected examples of the assessments and their design features, and describe the implementation of the performance assessments over two pilot years.

Assessment Design Features FOR Learning

Four performance assessments in English language arts, seven in mathematics, and eight in science inquiry were designed and piloted during the state pilot. These performance assessment tasks were designed to be ***embedded in the curriculum***, meaning that they were designed to be implemented by teachers in the course of teaching their standards-based instructional units. The tasks were designed to take 1-2 weeks to complete, with some in-class, collaborative components and final work products completed individually and independently. In contrast to on-demand assessments, the curriculum-embedded nature of the performance task work inherently allows for formative assessment practices to be applied. The tasks are designed to ***scaffold students' learning toward completion of the final task*** through the production of interim work products. For example, in the English language arts tasks students read and took notes on required or self-selected texts, and had an opportunity to engage in discussions with classmates on those texts prior to using the texts in their final essays. In the science and English language arts learning tasks, ***collaboration within groups was built into the tasks*** to allow students to learn from their peers. The tasks provide the opportunity for students to get ***immediate feedback from teachers and peers on their work***, and for the teacher to ***make adjustments to their instruction*** based on early drafts of students' work. To support teacher learning, they were ***given agency to select*** two tasks to pilot (one fall and one spring)[4] and received professional development and coaching related to those two tasks. These tasks, while standardized across classrooms, were designed to be ***flexible for teacher use***. In the case of the English language arts tasks, teachers and students were given a ***choice of texts*** to use for the task or teachers were permitted to substitute required texts with comparable texts.

In terms of task implementation, teachers were provided with ***individual discretion about how they would implement the performance tasks*** and how it would fit into their curricula. However, they were also provided with ***guidelines around the appropriate level of supports, instructional scaffolds, and feedback*** that they could provide to students without decreasing the rigor and challenge of the tasks. Prior to the pilot period, all teacher participants were provided with two days of ***professional development*** in which they learned about the tasks, what was expected in the tasks, and how to best scaffold the task to their students, depending on their students' characteristics. In addition, pilot teachers were provided with limited ***ongoing coaching*** throughout the year to support their implementation of the tasks. Following scorer training and scoring sessions, teachers were also provided with ***time to analyze student work as well as reflect on their task piloting experiences*** to think about implications for their instruction and future task implementation.

[4] This was true in English language arts which had four tasks to select from and in mathematics which had seven tasks to select from. In science, only two tasks were designed for each of four courses (biology, environmental science, chemistry, and physics), and so science teachers did not have choice in this case.

The performance assessments were designed to be scored using genre-specific, descriptive four-level analytic rubrics. ***Because of their analytic design, the rubrics can be used to provide detailed feedback to students about their performance***. Since the rubrics are genre-specific rather than task-specific and the same dimensions are scored across tasks, it becomes possible to track a student's progress along the same dimensions of performance over time across years and courses in the same discipline (e.g., science inquiry, math problem-solving). Teachers were asked to ***introduce the rubric to the students*** so that they know the evaluative criteria that will be used to assess their work in advance. Teachers were also encouraged to provide students with ***opportunities for revision*** of their performance prior to submission.

The Development Process for the OPAPP "Learning Tasks"

Building Valid Tasks Through the Input of Stakeholders, Implementing Teachers, Higher Education Faculty, and Content Specialists

Project Orientation and Identifying Performance Outcomes. On February 3-4, 2009, SCALE staff and ODE content leads conducted a Performance Assessment Design Studio training with multiple goals, specifically to: (1) introduce OPAPP participants to the project and to provide an overview of the key elements of performance assessment; (2) identify the constructs to be measured within each subject area; and (3) gather teachers' input to develop a set of performance outcomes that would guide the development of rubrics and learning tasks.

SCALE and ODE content leads worked with their content area teams to identify the constructs to be measured in the performance assessments to be designed. Participants generated lists of what it meant to be ready for college and careers in each subject area and higher education faculty members reflected on what is expected of entering college freshmen and provided feedback on the ideas generated by the groups. To identify the constructs to be measured by the learning tasks, participants in each content area brainstormed the enduring understandings, essential skills, and key habits of mind that they wanted all students to know and be able to do before graduating from high school. Pulling common features from these generated lists resulted in agreement on an initial list of task parameters (general specifications for the kind of performance tasks to be generated, e.g., in ELA, the use of multi-modal texts; in science, an inquiry-based task). In addition, participants examined a sample of existing performance assessment tasks to generate task parameters.

Performance Outcomes were defined as the academic knowledge, behaviors, and skills that students are expected to learn and demonstrate in a performance task. They serve as the blueprints for the performance tasks and establish the criteria for scoring performance. To generate Performance Outcomes, participants were given opportunity to study sample performance outcome statements from prior performance assessment projects, e.g., the Envision Schools Graduation Portfolio assessments, New Standards Performance Standards) as well as standards statements from state and national standards documents (e.g., NCTM standards, IRA/NCTE standards, NRC Science standards, 21st Century Skills Standards).

Within content area teams, participants generated a list of performance outcomes that captured commonly valued performance criteria aligned with the aforementioned standards, college and career readiness expectations, and key concepts in the disciplines. Based on participant input, SCALE generated a preliminary list of performance outcomes by domain for each content area, solicited feedback from OPAPP content leads, and generated a draft set of performance outcomes. OPAPP participants and non-OPAPP teachers at their schools (some of whom intended to participate in the project by piloting tasks) were invited to review and comment on the draft performance outcomes. The planning team felt it was essential to gather input from both project and non-project teachers to build buy-in from piloting teachers as well as teachers who may be involved in the project in the future. This review process was also meant to build validity into the instruments by building on consensus in the field around the performance outcomes to be measured by the OPAPP tasks.

The on-line survey used to solicit feedback on the performance outcomes listed each domain with the appropriate performance outcomes, and asked two questions, a) "To what degree does the performance outcome above represent critical understandings and skills necessary for student success in either college or work? and b) "To what degree does the performance outcome above represent critical understandings and skills that are core to the discipline of [content area]?" Then the survey asked participants to respond to two final questions: c) "Overall, to what degree are the performance outcomes aligned to the Ohio state content standards?" and d) "Overall, to what degree are the performance outcomes aligned to expectations for college and career readiness?" Respondents evaluated the performance outcomes using a four point likert scale (1= Not at all, 4= Is truly a critical outcome). Roughly equal numbers of OPAP participants and non-participants provided feedback on the ELA and science performance outcomes. Feedback for the math performance outcomes came primarily from OPAPP participants.

Overall, respondents agreed the performance outcomes within the different domains were critical for student success in either college or work and represented understandings and skills core to the discipline. For both questions, the ELA performance outcome mean ratings ranged from 3.4 to 3.7 on a four point scale. The science performance outcome means followed the same trend with most ratings between 3.2 and 3.8, with two exceptions. The ratings of the mathematics performance outcomes were slightly lower than those for ELA and science, ranging from 2.9 to 3.6.

Across subject areas, respondents reported high levels of agreement that the performance outcomes were aligned with the Ohio state content standards with slightly higher mean agreement levels for ELA (3.7) than for math (3.4) or science (3.3). In addition, mean agreement levels by content area indicate that respondents believe the performance outcomes were well aligned to expectations for college and career readiness with mean agreement levels of 3.8 for ELA and mathematics, and 3.6 for science.

SCALE content leads, with the input of the ODE content leads and teachers, revised the performance outcomes and circulated them for additional feedback from OPAPP participants. OPAPP participants were also encouraged to circulate the performance outcomes within their departments to solicit feedback from other teachers in their content fields.

In English language arts, a total of 34 OPAPP and non-OPAPP participants submitted survey results. With the exception of one domain, respondents gave high ratings (mean scores ranging between 3.3 and 3.8 on a four-point scale) to 83% of the performance outcomes. The domain labeled, “Reflection on the process of textual production” accounted for the lower mean ratings between 2.9 and 3.2. Significant changes were made to this domain in the performance outcomes; it was not dropped, however, because many of the OPAPP teachers insisted that it was core to the discipline and there was consensus in the group that Reflection was an important 21st Century skill that needed to be taught more consistently in English classrooms and that leaving it out would suggest that it was not important.

In mathematics, 19 OPAPP participants responded to the performance outcomes survey. Three of the four domains were rated high to moderately high with mean ratings ranging from

3.1 to 3.6. Comments related to these three domains concentrated on clarifying phrases used within the rubric and concerns about students' mathematical skills or intellectual maturity. Mathematical Reasoning had the lowest mean scores of 2.6 to 3.3 with comments focusing on whether teachers should expect students to "go beyond the task at hand" and whether "proofs" are essential for college success.

In science, a total of 34 OPAPP participants and non-participants provided feedback on the performance outcomes. All domains were rated high with mean ratings ranging from 3.4 to 3.7, except "Reflect on the Process and Findings" which was moderately high. One respondent representing special populations voiced concern about the cognitive demands that these outcomes might present for different groups of students. Most of the other comments focused on suggested word changes, eliminating or combining redundant outcomes, and moving outcomes from one domain to another.

Designing Rubrics. The primary goal for the second Performance Assessment Design Studio (March 4, 2009) was to acquire some initial input for designing the rubrics. Participants examined a sample performance task and samples of student work, using a group jig-saw approach so that different groups saw a variety of task types and student samples. During this exercise, participants were asked to:

1) Identify the performance outcomes evident within the task demands
2) Predict which performance outcomes would be evident in samples of student work
3) Examine the samples of student work looking for evidence of the performance outcomes
4) Categorize student work samples as "Developing", "Proficient", or "Advanced"

5) List the characteristics of the student work under each category to be used as a basis for developing a rubric
6) Prepare a graphic organizer on poster paper to be shared with the larger group

SCALE staff facilitated a whole group discussion enabling each group to present their ideas, to provide feedback on the sample task outcomes, to debate the qualities of performance for each level, and to develop general agreement about the rubric level descriptors, particularly for the "college/career ready" level.

SCALE staff generated a preliminary draft version of a rubric. A continuum of performance for each outcome was distributed across four scale levels: emerging, developing, proficient, and advanced. Proficient was identified as the performance level for "college and career readiness". SCALE staff incorporated as much of the wording and input from the OPAPP participants as possible. ODE content staff also reviewed and provided input on the rubrics, which were also used to make additional revisions. The preliminary rubrics were circulated to OPAPP participants and they were asked to provide input via Survey Monkey prior to the next Design Studio session.

Forty-three OPAPP participants provided comments on the ELA rubrics. Overall mean ratings across all the domains ranged from 2.7 to 3.3 on a four-point scale. Mean ratings by domain are indicated on the following table. OPAPP participants' comments focused on suggesting word choices, wanting to clarify distinctions between levels, and expressing concerns that the expectations might be too high for high school students.

For mathematics, 20 OPAPP participants responded to the rubric survey. Respondents' mean ratings were between 3.2 and 3.6, indicating overall satisfaction with the preliminary math rubrics. The majority of the comments focused on rewording within the levels, noticing lack of distinction between levels, expressing concern about high expectations, and questioning the meaning and appropriateness of some of the criteria.

Thirty-two OPAP participants responded to the survey for the science rubric. Overall the participants gave high ratings to the entire preliminary rubric with mean ratings between 3.3 and 3.7. Most of the comments involved suggesting word changes, requesting clarification of terms or phrases, asking for more distinction between levels 3 and 4, expressing concerns about the amount of time, and questioning whether level 4 expectations were too high. One person representing special populations continued to express concerns about the ability of vocational and IEP students to achieve the proficient level.

Gathering Input for Learning Task Development. The third Performance Assessment Design Studio (April 29-30, 2009) focused on finalizing the performance outcomes, clarifying feedback on the preliminary rubric, discussing the importance of alignment between the performance outcomes, rubrics, and learning tasks, generating specific design ideas as the basis for designing the learning tasks, and considering students' learning needs to provide access and support for all students.

Participants deconstructed a sample learning task using a series of questions to assist with their analysis of the task. The purpose of this activity was to help participants learn the components of a learning task and to be able to evaluate performance-based tasks in the future. The series of questions required participants to: identify the core disciplinary content and skills that could be demonstrated by the task, determine the extent to which the task would require students to engage in 21st Century Skills for college and career readiness, evaluate the accessibility of the task to all students and to discuss the strengths and weaknesses of the sample tasks. After completing this activity, participants completed a table indicating the alignment between the draft performance outcomes and the sample tasks. Each group shared their findings to generate a common list of the essential features of an OPAPP performance-based learning task.

On the second day participants generated specific ideas for the learning tasks. In the science group, participants worked in course-specific teams (e.g., chemistry, biology). In English language arts, participants were grouped based on the common themes or texts they worked with (based on the curriculum unit overviews they had brought with them) (e.g., the American Dream, the Role of Media, the Hero). In science and ELA, participants agreed on topics/units taught in the fall and spring that would be appropriate for designing performance-based learning tasks. Then participants brainstormed a list of student learning needs, types of instructional supports for students to succeed with these types of learning tasks, and required resources. Finally, participants started to sketch out the essential components of the learning tasks, based on the previously generated list. Posters with participants' input for the learning tasks were collected and used by SCALE staff in designing the first drafts of the learning tasks. SCALE and ODE content leads also gathered additional feedback on the revised performance outcomes and rubric drafts so SCALE staff could finalize the revisions of both documents.

In the mathematics group, participants did not engage in a task development process but in a task *selection* process. Groups were presented with 80 or so existing performance tasks in math, some brought by the SCALE content lead and some brought by the OPAPP participants themselves. Each group quickly reviewed sets of tasks in a jigsaw fashion and rated the tasks. By the end of the day, approximately 18 tasks were selected as finalists for the pilot based on the consensus of the group. Final selections were made by the SCALE and ODE content leads based on an evaluation of the task quality, rigor, and content appropriateness. Some tasks were adopted (original versions intact, e.g., *Skeleton Tower*, now adapted and renamed *Sally's Sugar Stack*, *Maximum Volume*), some were adapted (*Grazing Area, Body Surface Area, Wheelchair Ramp*), and several were redesigned significantly (*Heating Degree Days*, *Open for Business*).

These curriculum-embedded learning tasks focus on constructs essential to the core disciplines of English language arts, mathematics, and science. For example in English language arts, students applied their understanding of a central theme in American or world literature to a task that requires analyzing, interpreting, and responding to texts, and ultimately synthesizing their own ideas about the texts/theme. The mathematics tasks require students to develop an approach to solve everyday problems and apply mathematical concepts and skills to execute the solutions. In science, the tasks assessed students' understanding of scientific inquiry and their ability to design and conduct an investigation.

For example in the science task *Got Relieve It?*, students are given the role of employees for a chemical company called Achoo-B-Gone. In the contextualization of the task, they are told their team has been working for the past year to create a new drug that will instantly relieve cold symptoms. The new product, Relieve IT, is in the final testing stages before being sent to the Food and Drug Administration (FDA) for human trials. Part of the Federal Drug Administration (FDA) approval requires the team to share their current knowledge about acids and bases and to provide all of their experimental data on the pH levels of Relieve IT. The FDA provides a preliminary report identifying concerns of potential negative human side effects due to the product's pH levels. The FDA wants to know what the team will do to "fix" the product before initiating human trials. Students must synthesize their current knowledge about acids, bases, and neutralization reactions. They then design and conduct an experiment, as a group, to determine the pH of the Relieve IT product and to determine which unknown solution (1, 2, 3, or 4) or combination of solutions can be used to neutralize any excess acid or base. Finally they prepare an individual formal lab report which includes recommendations for "fixing" the pH levels of Relieve It.[5]

OPAPP participants and ODE content staff provided extensive feedback on the draft tasks before, during, and after implementation of the tasks with their students. The feedback focused primarily on the formatting of the learning tasks, questions about the appropriate level of instructional scaffolding to provide students prior to the learning tasks, and suggestions for re-wording to make the task and prompts more accessible for all students.

In summary, the three sessions encompassing the Performance Assessment Design Studio accomplished four major goals resulting in draft learning tasks in a four month period. First, SCALE staff presented a framework and design principles enabling OPAPP participants to learn about performance based assessments. Second, OPAPP participants, ODE content staff, and SCALE experts co-constructed a set of performance outcomes aligned with national/state content standards as well as the 21st Century Skill Standards. Third, the performance outcomes served as a blueprint for developing content-specific analytic rubrics describing a continuum of student performance. Last, the scoring rubrics provided the framework for the development of the learning tasks. OPAPP participants identified key content and provided input on the design of the learning tasks. At each step, OPAPP participants and ODE content staff provided extensive feedback on the draft instruments before piloting them with students.

Preparing Teachers to Implement Tasks

When SCALE was contracted to provide technical assistance for the project, it was hired on the basis of its expertise in providing technical support around the design, implementation,

[5] See Appendix A for additional sample learning tasks from each content area.

and scoring of performance assessments, and not as a professional developer. In fact, professional development was not part of the scope of work in SCALE's initial contract. However, it became clear early in the project that in order for teachers to implement the OPAPP learning tasks as an embedded part of curriculum and to pilot test the tasks as intended, the capacity of teachers to engage in inquiry-oriented instruction and to embed performance assessments in their curriculum would need to be built through professional development.

SCALE and ODE partnered to design and execute a professional development plan that would support the implementation of tasks by teachers and schools involved in the pilot. Teachers had been asked to pilot two performance tasks over the course of the first academic year, one during the fall of 2009 and another during the spring of 2010. The fall 2009 pilot was considered a low-stakes "practice run" that would allow teachers to try out one performance task and learn from the experience to support more effective implementation of the spring 2010 pilot. All teachers interested in participating in piloting OPAPP learning tasks were required to attend the professional development - the orientation to the tasks, as well as the scorer training and scoring sessions that were to follow the fall and spring pilots. In the end, approximately 50 teachers in each content area implemented OPAPP learning tasks in 2009-10.

Since there was only one opportunity to introduce all of the tasks to teachers in the fall 2009 and there were limited SCALE and ODE personnel with enough expertise to lead the professional development, the project leads decided on a "trainer of trainers" approach by involving the OPAPP coaches. During the week of August 17-21, 2009, all coaches were provided with an intensive five-day session during which they were introduced to the performance tasks and helped to co-design a professional development plan for participating teachers. The introduction to the performance tasks varied by content area, but all were intensive and required that coaches either engage in completing the performance tasks themselves (mathematics, science) or in part of the instructional arc that would support teachers' and students' understanding of the performance tasks (ELA). Since the science tasks involved, in some cases, the use of laboratories to complete the tasks, the science coaches' training session was held at a lab space donated by an Ohio State University professor. The other coaches' training sessions were held at the Educational Service Center of Central Ohio (ESCCO) which had hosted the initial development sessions in the spring of 2009.

The first day of the five-day coaches' training was a common training day for all content-area coaches, co-led by the Buck Institute for Education (BIE), which had been recruited by SCALE to lend their expertise in the implementation of project based learning and to adapt a portion of their training module for this project. BIE modeled approaches to introducing the performance tasks to students in ways that lead to student engagement in the tasks (i.e., providing a "hook") and provided images of what inquiry-based instruction looks like in practice. In the afternoon, the group broke out into content-area teams to explore these topics in content-specific and task-specific contexts.

The remaining four days were spent in content area teams, engaging in a "deep dive" of all of the tasks in the content areas. Each content area team took a different approach in defining the roles of the coaches. For the purpose of facilitating the professional development sessions for teachers, each coach was designated to be the "expert" on at least one performance task and to facilitate or co-facilitate professional development sessions on that one task. This was more true for ELA and science coaches, and less so for mathematics coaches as described below.

In science, each coach helped to facilitate with SCALE content lead the day-long professional development session for the other coaches on the two course-specific performance tasks for which they were designated "experts". During each of the remaining four days of the coaches' training, coaches engaged in completing the two tasks in biology (*Invasive Species* and *Medical Mysteries*), environmental science (*Renewable Resources* and *Got Clean Water*), chemistry (*Is It Physical or Chemical?* and *Got Relieve It?*), and physics (*Energy Efficient Vehicles* and *How Things Work*). The coaches not only had the chance to deconstruct and experience each of the labs and tasks, but they debriefed each experience and used those experiences as the launching pad to generate and problem solve topics that were likely to be instructional issues for piloting teachers. They spent the last part of their sessions together planning for the teachers' professional development on the tasks, making sure coaches felt confident in their plans. The science teachers' professional development sessions were very similar in content with the science coaches' professional development on the tasks, except that teachers had a full day to engage with one task (whereas the coaches had half a day for each task), so that they became familiar with two tasks over two days. While SCALE content lead co-facilitated the teacher professional development for the two chemistry tasks, the other coaches had responsibility for facilitating the other three course-area tasks. The teachers' professional development in science was held on October 8-9, 2009 at a high school, which provided facilities to enable teachers engaging in the science inquiry labs.

In mathematics, coaches' expected roles were somewhat different. Rather than relying on the coaches to become experts on a particular performance task or set of tasks, the choice was made to rely on the expertise of the SCALE content leads to deliver the professional development on all of the tasks to teachers, with the coaches taking a supporting role. During the remaining four days of the coaches' session, the mathematics content leads led coaches in engaging in each of the seven mathematics performance tasks (solving the problems), discussing the various possible solution paths to solving the problems, and anticipating likely issues to arise among teachers and students, and designing the professional development for teachers. The teachers' professional development was similarly designed. During the first part of the first day of the two-day session, held on September 28-29, 2009, all teachers were presented with one math learning task - *Heating Degree Days* - and engaged in solving the task. The teachers discussed their solution strategies with each other at tables. The mathematics content leads debriefed the task, anticipating likely issues that might arise with students as they completed the tasks. During the remainder of the two-day session, teachers had a choice of attending one of two break-out sessions to engage in and learn about three additional tasks. In total, each teacher learned about four of the seven mathematics learning tasks - *Heating Degree Days* plus three additional tasks. This allowed them to have greater choice in determining which two tasks to pilot in the fall and spring. This approach relied the least on the coaches, who supported the teacher professional development as co-facilitators, but were trained primarily to provide support to teachers as they piloted the tasks.

In English language arts, the approach to introducing the performance tasks was the least like the other content areas because the amount of time it would take for coaches or teachers to actually complete the tasks as designed and then debrief that experience was too long in comparison to the time available for the professional development. Instead of attempting to take participants through completing the tasks, the decision was made to model approaches to introducing the tasks to students, identifying strategies for scaffolding the tasks, and providing resource materials to teachers to support task implementation and tools for scaffolding. SCALE's ELA content lead, with the support of a staff member from the Buck Institute for Education, designed and led the professional development training for the ELA coaches. Using one of the tasks, *Americans Dreaming*, SCALE's ELA content lead modeled approaches to introducing and scaffolding the task to students, providing resources to teachers to support those approaches to implementing the task. Time was spent "unpacking the task" in terms of the student's perspective (i.e., what do I have to do in this task?), teacher's perspective (i.e., what do I have to do to support student success on this task?), and coach's perspective (i.e., what do I have to do to support teachers implementing the task?). The remaining three days were provided to coaches and the ODE team to plan the professional development for teachers, using the model provided by the SCALE ELA content lead. Each coach was responsible for planning and executing a one-day session on their assigned task, delivering the same professional development session twice over the two days of the teacher professional development session (September 21-22, 2009). In this way, teachers had the choice of participating in two out of four sessions offered and were prepared to implement two out of the four ELA tasks.

Performance Task Implementation

The performance tasks as originally conceived by ODE and SCALE were designed to be curriculum embedded, meaning that they would be implemented by teachers in the context of their existing curriculum units over a period of one to two weeks of classroom time. What "curriculum embedded" means with regard to the OPAPP learning tasks varied by content area. In addition, the tasks varied in length of time for completion by content field, with mathematics tasks taking a few days of class time (with students also working on final write-up of the solutions independently outside of class time), science inquiry tasks taking approximately one to two weeks of class time, and ELA tasks taking between two to four weeks, depending on the teacher.

In ELA, tasks were "embedded" as though they were mini-instructional units in themselves, with teachers introducing the required and optional texts, working through the texts with students together, helping students to complete the culminating tasks through instructional scaffolds, and guiding students through the writing process. Because teachers were allowed to select some of the required texts for the tasks, the scope of the task might vary depending on the length and difficulty of the texts (e.g., a set of images, poetry, short story, or essay versus a film, play, or a novel). The scope of the tasks also varied by task selected - responding to a series of short essays and writing one's own essay of the same genre (as in the task *Employing the Personal*) takes less time than responding to an epic hero narrative such as the *Odyssey* and writing an anthology introduction (as in the task *The Hero's Journey*) that must incorporate multiple hero narratives. Analyzing a set of media images portraying youth (advertisements, songs, television shows) (as in the task *Constructing the Self*) takes less time than analyzing a set of texts that portray the American Dream (e.g., *The Great Gatsby*, *Death of a Salesman*), interviewing and completing a profile of an "American Dreamer". Therefore, some tasks might take a shorter amount of time (2 weeks) while others might take up to 4 weeks or longer depending on the texts selected and how a teacher chooses to embed the task within existing units of study.

In science, it was intended that the learning tasks would be "embedded" as a culminating task in an existing instructional unit that was already commonly taught. For example, prior to introducing the Chemistry task *Got Relieve It*, teachers were to provide the instructional background on the content to be assessed (e.g., acids and bases) as well as practice with laboratory skills needed to complete the performance task (e.g., testing the pH of a solution, using other solutions to neutralize an acid or a base) PRIOR to administration of the Chemistry learning task. In most cases, this was true, although there were a very small handful of teachers who did not understand this and administered the learning tasks as *instructional units in themselves* without having provided the prerequisite background instruction (e.g., the biology task *Medical Mysteries* requires knowledge of human anatomy systems). This instructional preparation should not be considered as part of the time it takes to administer the task. In either case, the time that it should take to administer the science learning tasks revolves around the time involved in introducing and implementing a science lab (usually about 2-3 periods of class time) followed by the opportunity to analyze the data and write up the individual lab report both in and out of class. In most cases, lab reports were typed, requiring time in a computer lab or outside of class time to complete. Therefore, the total time involved could range from 5-10 days both in and out of the classroom.

In mathematics, tasks were not administered as part of or as a culminating assessment for an instructional unit because the content of the tasks were not designed to be aligned with particular units in the 11th-12th grade curriculum (Algebra 2, Pre-Calculus). Rather, the tasks were selected and designed to require the application of prior math content learned (Algebra, Geometry) but with a "cognitive load" that would require the cognitive maturity of an 11th grader. The tasks were intentionally selected and designed NOT to align with specific units in Algebra 2, Pre-Calculus, or Calculus (although some of the tasks could be solved using skills learned in these courses) because they were not designed to be administered as "unit tests", where students would automatically know what mathematics skills and algorithms to apply, reducing the cognitive complexity of the tasks. Instead, the tasks were selected or designed to present students with novel problem where the most difficult cognitive demand was to decide what set of mathematical content and skills to apply to the situation. These tasks could, in effect, be "embedded" at any point in an 11th or 12th grade mathematics course as long as the content knowledge required had been covered previously. The mathematics tasks could take as little as one class period or multiple class periods (up to five class periods) depending on the academic preparedness of the students and the amount of time provided for students to produce their individual responses. On some tasks (e.g., *Sally's Sugarstack, Grazing Area*), a handwritten solution was sufficient whereas in others, students were directed to type up their solutions in a narrative form (e.g., *Heating Degree Days, Open for Business*) to explain their reasoning and communicate their solutions through narrative, numeric representations, figures, and symbols.

Across the three content areas, teachers were directed to provide students with in-class time to work collaboratively with their peers on the math problems, science labs, or textual analyses (ELA), but were also expected to produce their own independent responses in submitting a final product. While it was not required in math and science, students were encouraged to type their responses on a computer, which necessitated time in a computer lab, a library with access to a computer, or at home. This would require more independent work time outside of class. During class time, teachers were encouraged to introduce the tasks and scoring rubrics, answer questions about the tasks, and support students' beginning work on the tasks, including small or whole class discussions about the required texts (ELA), small group work on solving the math problems, and all lab work in science.

Teachers were also encouraged to provide opportunities for students to submit drafts of their work and to provide feedback that would support students' improvement of their drafts prior to final submission. Given that these were performance tasks that had primarily a formative purpose (and no high stakes attached to them), teachers were encouraged to provide instructional scaffolding for the tasks in ways that make the tasks accessible to all students, including special needs, ELL, and gifted learners. This would mean that teachers might break down the tasks to support students getting started on the tasks, provide graphic organizers for students to use as they interacted with a text, or they might using grouping strategies to support peer interaction around sense-making activities such as analyzing a set of texts or solving a problem. Or a teacher could provide practice on different tasks with a similar set of task demands (mathematics). In science, teachers were directed to teach the pre-requisite content and skills needed to complete the science inquiry tasks PRIOR to assigning the performance task.

At the same time, there were also limits placed on the kinds and levels of scaffolding and support teachers could provide to students, and students were required to submit an attestation that the work they were submitting was their own work, and not completed by a friend, parent, or teacher. OPAPP distributed a document "Teacher Guidelines for Task Implementation" during the Task Orientation sessions in the fall 2009. This document provided teachers with some structure for determining how much instructional scaffolding and teacher support was appropriate and setting limits on them so as not to reduce the cognitive complexity and demands of the learning tasks. Under “scaffolding the task for students” category, the guidelines asked teachers to review the performance outcomes and rubrics with students prior to implementing the task, to provide an organizing visual to assist students in understanding the flow of learning and assessment activities as well as the work to be produced, to practice thinking strategies, ways of working, and approaches to problem solving on other learning tasks prior to the performance assessment task, and to use pedagogical principles that support curriculum access and learning for all students, and to make appropriate accommodations for English language learners, special education students as well as gifted students. Guidelines referring to the appropriate amount of teacher support and feedback requested that teachers check-in periodically with students on their tasks and to provide formative feedback in the form of questions or comments without providing specific directions on how to approach a problem or suggesting a potential solution strategy, build-in opportunities for students to monitor their own progress and reflect on their work, provide constructive feedback focused on the big ideas in the performance outcomes and rubrics, and consider allowing students to revise their work which would be attached to their original work sample.

Scoring Protocols and Teacher Involvement

The scoring scales for the learning tasks in ELA and science consisted of content specific, analytic rubrics with four distinctive levels describing a continuum of student performance. Level 3 was defined as the "proficient" or "college/career ready" level for students completing high school. Descriptors at this level were carefully crafted with the input of teachers to signal that a student is ready for college-level work. Level 4 was defined as "exceptionally advanced" or already doing college-level work. Level 2 was defined as developing proficiency, with some strengths and some weaknesses in performance. Level 1 was defined as "emerging" with attempts made to demonstrate proficiency but with significant weaknesses in performance. The analytic design of the rubric enables teachers to provide detailed feedback to students about their performance, determine gaps in students' understanding, and guide their planning for future instruction. Using genre-specific rubrics (e.g., science inquiry, ELA inquiry and communication) instead of task-specific rubrics enables teachers to monitor a student's progress along the same dimensions of performance over time and across courses within the same discipline. OPAPP teachers were encouraged to share the learning task rubrics with their students both to provide them with clear criteria for completing the project and as a tool for them to evaluate their own performance. Teachers were also encouraged to allow students to revise their work prior to final submission.

In contrast, the math scoring scales relied on a point-based rubric designed to address the specific requirements of a particular task. For example, in the task *Open for Business,* a student could earn up to five points on the following "item", with partial credit given based on the number of correct component responses:

Determines the linear relationship of demand and price for Laptop, Xbox and Ipod.(5 points)

Laptop: **$q = (-3/50)p + 69$** Xbox**:** **$q = (-12/25)p + 185$** Ipod: **$q = (-8/25)p + 99$**

Stereo: **$q =(-14/19)p +131.53$** Calculator: **$q = (-14/10)p + 150$**

Partial Credit: 1 to 4 errors in slopes or intercepts (up to 3 points)

Partial Credit: 3 slopes correct (1 point)

Across tasks, students could earn a range of 1-5 points across a range of "items" (between 11-19 items totaling between 25 and 40 points). Although each scoring rubric varies by task, teachers are able to use the rubrics to provide feedback to students about strengths and weaknesses in their performance, and scores can be used by teachers to analyze their instructional decisions and determine next steps.

In all content areas, SCALE content leads led benchmarking teams in reviewing student work samples from the pilot schools to identify exemplar papers ("benchmarks") representative of each scale score on the rubric, for each implementation cycle - fall and spring. Using the rubric, each exemplar paper was scored and the evidence to support each score was recorded on a scoring form. Prior to training teachers to score, SCALE staff conducted scorer training sessions with ODE content staff and coaches to calibrate all project leads on the benchmarked samples.

To support inter-rater reliability and comparability of scores across sites, teachers participated in a scoring training and calibration process using the benchmarked samples. In the scoring training, OPAPP teachers independently examined the benchmarks of student work, selected evidence from the samples of student work, and assigned a score for each dimension using the rubric’s score level descriptions. Teachers compared their scores with the benchmarking team's scores and discussed any discrepancies between the scores based on evidence gathered from the samples. Teachers also commented on rubric related issues which resulted in minor revisions to the rubric throughout the process. After repeating this process for each set of samples, with 1-3 samples representing each score level (1 sample for each level in ELA and science and 2-3 samples for each level in mathematics), participants independently scored at least one calibration sample, followed by additional practice scoring samples to further calibrate scorers. OPAPP teachers were considered "calibrated" when a majority of their individual "item" scores AND the overall score level of the sample (e.g., "Level 2" or "Level 3") matched the benchmarked scores. Regardless of whether or not they calibrated, participants were asked to score at least one class set of students’ papers and to submit them to SCALE for analysis.

Findings

While the OPAPP learning tasks and their intended use, including scoring protocols, were designed with the goal of supporting teaching decisions that allow for the formative promise of the performance assessments to be realized, there were often unintended outcomes, as well as wide variation in the ways in which the learning tasks were used. The following section describes the outcomes and lessons of implementation from four standpoints: 1) the professional development provided to participating teachers; 2) performance task implementation; 3) consequences for participating teachers and students; and 4) scoring protocols and teacher involvement in scoring.

Outcomes of Professional Development

While the primary professional development for implementing the OPAPP learning tasks was offered at the beginning of the academic year, additional opportunities were embedded in other events designed to engage OPAPP teacher participants. During scorer training sessions in January 2010 (on the fall 2009 tasks) and April 2010 (on the spring 2010 tasks), teachers were provided opportunities to discuss and share what they were learning from examination of student responses to inform their instruction and strategies for implementing the tasks. The mathematics team led by facilitators employed by SCALE to provide technical support in mathematics) went a step further and introduced the idea of planning "re-engagement lessons" following the examination of student work to support further student mastery of the content and skills measured in the tasks.

Overall, participating teachers responded positively to the professional development offered by the project on the tasks and task implementation. SCALE administered an evaluation of the "task orientation" session (the professional development required of all implementing teachers in the fall of 2009), and across all three content areas, participants provided positive ratings of their experiences during these sessions.[6] On a scale of 1 to 4 with 1 representing

[6] Full results and analyses of participants' evaluation ratings of these sessions can be found in the external evaluator's final report of the 2009-10 pilot phase evaluation on pages 53-69 (Woodruff, Zorn, Castañeda-Emenaker, & Sutton (2010b)).

"Strongly Disagree" and 4 representing "Strongly Agree", ratings of the task orientation sessions on average were between "Agree" and "Strongly Agree" across the three content areas (See Table 1 below). Although the difference across the content areas is non-significant, it appears that the science task orientation sessions and facilitation were rated the highest.

Table 1

Mean Ratings of Task Orientation Sessions by Content Area (Sessions held in early fall 2009)

Item	Content Area	N	Mean	Std. Dev.
The task activities designed to help us "experience the task as learners" helped me to better understand the content and skill demands of the performance task.	Science	50	3.59	0.44
	Math	41	3.54	0.44
	ELA	37	3.45	0.45
The task facilitator(s) led debriefs and discussions about the performance task in ways that deepened my understanding of the task.	Science	51	3.57	0.47
	Math	40	3.49	0.42
	ELA	38	3.51	0.39
The task facilitator(s) effectively answered questions about the task and task implementation.	Science	49	3.50	0.46
	Math	39	3.40	0.42
	ELA	38	3.49	0.46
Overall, I feel that the task orientation session for this task helped me to understand the task and how to implement the task.	Science	51	3.55	0.44
	Math	40	3.38	0.43
	ELA	38	3.47	0.44

Note: Rating scale was 1=Strongly Disagree, 2=Disagree, 3=Agree, 4=Strongly Agree

Professional development - lessons learned. While the feedback on the professional development offered was mostly positive, there were also lessons learned based on results of the pilot about the adequacy of the professional development provided. One of the biggest challenges was the lack of time for providing teachers with an adequate level of exposure to the content of the tasks as well as the pedagogical strategies necessary for preparing students for success on the tasks. Teachers were released from school only two days in the fall of 2009 to participate in professional development on the tasks. With the exception of the 60 "lead teachers" (20 ELA, 20 math, 20 science) who had participated in the development of the assessment instruments, encompassing five days of contact in the spring 2009, the majority of the 150 or so teachers involved in the pilot had had no previous experience with the OPAPP assessments and little experience with performance assessment in general. While implementing curriculum-embedded performance assessment requires a fairly dramatic change in instructional approach for most teachers, the project had access to little time and few resources to provide teachers with the ongoing pedagogical training that would be necessary to see actual changes in teaching practice. Research suggests that the most effective professional development involves extended and ongoing contact time (at least 50 hours or more), embedded in teachers' work, and supported by coaching and professional learning communities (see Darling-Hammond, et al., 2009 for a recent review of research on effective professional development). However, the amount of time and resources available for professional development in the project was quite limited. Therefore, while acknowledging the need for more extensive and ongoing professional development around the use of inquiry-based teaching methods, the approach taken by this project in the first pilot year was to provide a two-day professional development session encompassing about 15 hours of contact time, with a few hours of voluntary follow-up with coaches. With limited contact hours, the project leads in mathematics and science made a decision to focus primarily on developing teachers' content knowledge in order to provide them with an "even playing field" with regard to the content measured by the tasks. This was particularly true in mathematics and science, and less so in English language arts, where the focus was more on providing pedagogical strategies for scaffolding the tasks with students. The difference between the 60 "lead teachers" who had been part of the pilot from the beginning of the development phase and the others who had begun participation in the project during the fall 2009 professional development on task implementation was observed both anecdotally and in teachers' ratings of their professional development experiences. The "lead teachers" were much more likely to understand the rationale and purposes of the pilot and the intent of the performance tasks since they had been involved in developing the performance outcomes, rubrics, and tasks from the beginning and had had time to deepen their understanding of performance-based assessment.

It is unlikely that it would be feasible to include all future OPAPP participants in the design and development of the assessment instruments (tasks, rubrics), or even in direct professional development and scoring opportunities led by expert facilitators. However, the OPAPP teachers who began their participation during the development stage benefited from a ***more extended and ongoing engagement*** with the project than did teachers who began participation during the pilot phase. These OPAPP teachers benefited from approaches to professional development that required a ***"deep dive" into a performance task***, completing the task themselves and engaging in discussions with other teachers to "unpack" the tasks and what they require of students and what they assume about the kinds of instructional scaffolds teachers should provide to best prepare students for this kind of work. Last, teachers benefited from the ***opportunity to share their implementation experiences and lessons learned with peers, and to score and analyze their own students' work for the purpose of critically examining implications for instruction*** ("re-engagement lessons"). These three features of professional development should continue to be an integral part of the professional development that is offered by OPAPP.

Outcomes of Performance Task Implementation

There was apparently wide variation in the ways that teachers implemented the learning tasks. This was evident from the student work samples that were submitted by teachers, as well as based on self-report from teachers and coaches and content leads' observations. In some cases, it was apparent from student work that teachers overly prescribed the texts (in ELA tasks) when students were to be given choice in making text selections. For example, during the fall 2009 pilot, in one school where there was a district-approved set of readings for a humanities course that integrated historical writings concerning the democratic values of the colonial United States, all of the teachers implementing the task *Employing the Personal* required that students select from among a set of these texts, essentially proscribing the textual analysis portion of the task to a set of essays about democratic values, even though the intent of the task was to engage students in selecting and analyzing a series of *personal* essays on one of two topics: "man's responsibility to the environment" or "man's responsibility to others", and then writing their own personal essays on the same topic. This resulted in a disconnect between the analytic textual analysis portion of the task and their own personal essays, which had nothing to do with the historical essays analyzed in the first part of the task. In other cases, teachers provided a set of texts that were analyzed together in a whole-class discussion format resulting in the same or very similar student responses to a prompt across a class set or even across sections of the same course.

Both of these issues raise the question of the extent to which teachers and students should be given agency in the selection of texts for the ELA learning tasks. While having choice to select the texts for analysis is more likely to support student engagement in a given task, and having choice to prescribe a text that is already taught in an English course is more likely to support teacher buy-in to the project, the set of texts used in a learning task has a direct bearing on the difficulty of the task (depending on the complexity of the texts). Thus, a single task (e.g., *Constructing the Self*) may be highly variable in its rigor and cognitive demand depending on the teacher's implementation choices and students' selection of texts. This level of variation was deemed by our task design team to be too variable, and this informed the revision of the ELA tasks for the second pilot year. While there were still some elements of choice in the tasks for teachers and students, a greater attempt was made to lend more "standardization" to the tasks by requiring one or more particular texts as the basis for an initial textual analysis. The second part of the task was more open-ended and allowed for greater choice among teachers and students. For example, in the new version of the task *Americans Dreaming*, all students were required to study and analyze three sets of texts: Langston Hughes' poem "Let America Be America Again", a specific movie adaptation of Arthur Miller's play "Death of a Salesman", and Dan Barry's *New York Times* article "At an Age for Music and Dreams". In the first part of the task, students were required to respond to and write a brief 500-word essay analyzing and comparing the ideas about the American Dream captured in Hughes' poem and Miller's play. In the second part of the task "Produce a Portrait of an American Dreamer", students were given more choice in whom they selected to profile and how they would represent the portrait (in a formal essay like Barry's article, or in a multi-media profile).

There were similar issues related to teachers' variation in how they implemented tasks in mathematics. In a few cases, it was apparent from student samples that the teacher over-scaffolded the tasks, resulting in the same answer or solution strategies across an entire class. (This was apparent, for example, when a short-cut formula was used over and over again by all students in a class.) In those cases, teachers effectively changed the cognitive rigor and demands of the task by selecting and prescribing an algorithm in advance for students so that all students had to do to solve a problem was "plug and solve." This was less true after the fall 2009 pilot after additional professional development during scorer training was provided and teachers gained a greater understanding of what it means to implement a performance task well. However, such guidelines cannot guarantee that teachers will not provide too many hints about ways to solve a problem.

In science, there were a number of cases in which teachers' own unfamiliarity with science inquiry, misunderstandings about a task, or lack of science inquiry skill was translated into how their students performed on a task. When teachers themselves did not know how to effectively analyze a dataset, for example, they were unable to provide adequate instruction to their students for how to analyze a dataset, and this often showed up in the students' work. In other cases, an error in the way a teacher demonstrated the computation of a formula showed up in all of the students' lab reports. There was the aforementioned example of a teacher who did not teach the content necessary for completing the *Medical Mysteries* task prior to administering the task, assuming that students would pick up the content knowledge necessary during the task administration. Again, these variations in classroom level implementation meant that the task demands changed and/or students' opportunities to learn and to succeed on the task varied depending on their teachers' implementation decisions. Sometimes, teachers' implementation decisions were related to conditions outside their control, such as access to lab or research materials, class schedules, and interruptions to regular class sessions. In the fall of the second pilot year (fall 2010), OPAPP sought to address some of the content knowledge and science inquiry skill gaps among participating teachers by providing focused professional development on particular skills required by the learning tasks. For example, one session was spent providing teachers with a data set and having them learn the appropriate way to analyze the data, modeling an instructional approach for teaching their own students how to analyze data.

Performance Task Implementation - Lessons Learned. Given the wide variation in teachers' implementation decisions when they are left with the discretion of how to present and administer the tasks, it is clear that the variations in student performance on the tasks cannot be attributed solely to variation in student ability and that the resulting scores also represent teachers' implementation decisions to a great extent. It is, therefore, quite sensible to use the learning tasks primarily as formative learning opportunities (as "instructional tasks") and to include standardized assessment tasks, which are designed to be administered under on-demand, standardized conditions, with the foremost purpose of measurement.

Second, ongoing professional development to help teachers understand what "inquiry-oriented" instruction looks like and to practice and develop those teaching approaches are also necessary if students' are to have a reasonable chance of developing the inquiry, analysis, and problem-solving skills necessary for completing performance tasks of this type. The rigor and nature of students' experiences with the learning tasks has as much to do with the way in which teachers present and implement these tasks as the quality of the tasks themselves. The intent of the learning tasks is that they should provide students opportunity to provide "open-ended" responses rather than constraining or prescribing a "right" answer or approach. In a school environment in which correct answers were previously valued more than students' ability to think through and develop an original response, it will take time for both teachers and students to unlearn this approach to teaching and learning and learn the higher order thinking skills assessed in the Common Core State Standards and necessary for college and career success.

Consequences for Participating Teachers and Students - *Consequential Validity*

The following description of the outcomes for participants is based on comments made by OPAPP science teachers at scorer training sessions following task implementation, during site visits made by coaches and the SCALE science content lead, and in participants' session evaluations.[7] While these outcomes are not directly generalizable across content areas, anecdotal evidence from ODE content leads and coaches who made site visits as well as feedback from students captured in their task reflections (ELA tasks) suggests that these outcomes are common themes across content areas.

Students' perspectives. Science coaches collected feedback from pilot site sites about their experiences with the learning tasks, including feedback from students. Students' reactions to the cognitive demands of the learning tasks were diverse. Some students expressed appreciation that the tasks enabled them to make decisions and to learn for themselves. For example, one student noted "This task was better than what my teacher usually does because we were able to make decisions and figure out how to solve the problem, instead of just mindlessly following his directions." In contrast, other students felt the learning tasks were too open-ended, wanted more structure, and had difficulty with inquiry-based learning as illustrated in this typical quote: "I'll be glad when we go back to our regular assignments, this is too much work. How are we supposed to know how to design an experiment? That's why we have lab books." Despite students' initial reactions to the tasks, science coaches reported that many expressed a feeling of achievement after completing the challenging tasks. A physics student who completed the task Energy Efficient Vehicles commented: "I'm really proud that our group's vehicle traveled as far as it did. It was hard, but it was doable hard, not impossible hard."

[7] We did not have the opportunity to directly conduct a study of the consequences for participants because it was not part of SCALE's contracted scope of work with the Ohio Department of Education.

Teachers' perspectives. Overall, the teachers were impressed with the learning tasks and immediately made connections between the goals of the learning tasks and specific content standards. In most instances, teachers were not accustomed to and expressed discomfort with the idea of allowing their students to investigate on their own. When questioned about their typical teaching strategies, the majority of responses included content delivery through lecture, use of highly-directed lab procedures, and assessing through multiple-choice and constructed-response tests. Some of these teachers indicated that they were more confident about inquiry-based teaching after they experienced the model lessons with the SCALE and ODE coaches. While completing the model lessons during professional development, some teachers displayed significant misconceptions or gaps in their content knowledge.

When asked to comment on the implementation of the tasks, teachers commonly cited the need to adjust their thinking around students' abilities: "I learned that I need to have higher expectations for my students and to do more inquiry based labs. My students exceeded my expectations. They were able to complete the task with very little direction from me." Teachers also noted the level of interest displayed by their students as illustrated by this physics teacher: "I saw students who were very engaged and invested in the designs of their cars. They were having great evidence-based discussions to determine which changes to make on their cars. I asked the kids if they liked what they were doing and they answered a resounding 'yes.' " Teachers reported that students were motivated to learn content in order to complete the tasks as described by this biology teacher, "My students came in voluntarily during lunch and after school to research the normal values for each of the diagnostic tests and what the results meant related to the Medical Mystery task….that is the first time I have seen that in 20 years of teaching."

These positive perceptions about the impact of learning task implementation are tempered by a number of challenges cited by teachers:

- ***Additional time in curriculum needed to prepare students for success on the tasks.*** Teachers needed more time to fully embed the learning tasks into their existing curriculum. Numerous teachers commented that they needed more time to scaffold the content and pre-skills necessary for students to successfully complete the learning tasks.

- ***Difficulty embedding the tasks in existing curriculum.*** Many teachers were not able to alter the sequence of their school-based curriculum that had been developed by course teams and found it very difficult to "fit" the learning tasks into their existing sequence.

- ***Variation in task implementation by teachers.*** Some teachers carefully planned how they would infuse the learning tasks within their instructional practices. Others simply administered the learning tasks like assessments with very little scaffolding or providing the content framework. An extreme example was a Biology teacher who assigned students the Medical Mysteries learning task but had not yet taught the students any of the body systems. Without some prior knowledge of the circulatory or digestive system students had to learn the prerequisite background content as they were working on the Medical Mysteries learning task. These students had no context or previous knowledge with which to brainstorm possible causes of someone's death. They struggled through the task using an uninformed trial and error approach.

- ***Teachers' lack of previous experience using inquiry-based teaching strategies***. Many teachers had previously not incorporated inquiry-based strategies within their current instructional practices. They were introduced to these strategies during their introduction to the tasks and had an opportunity to discuss how they would implement the tasks with their students. However, it takes much more ongoing professional development to support teachers' effective use of inquiry-based strategies. This level of professional development was not within the scope of this project.

Outcomes of Scoring Protocols and Teacher Involvement in Scoring

Scoring student work had a significant impact on teachers. During the scoring sessions teachers reflected on the quality of the students' work, focused on students' current performance levels, and how to support students to achieve at the proficient level. Teachers indicated the scoring process enabled them to assess their students' work using a common standard and forced many teachers to re-examine their expectations of students' performance. The opportunity to engage in this work also created a network for teachers to discuss instructional practices and other content-specific issues with colleagues. Responses on session evaluations revealed teachers viewed this professional development as "the most useful for improving my teaching and understanding of where my students are in comparison to where they should be." Table 2 below displays the feedback on the Spring-Summer 2011 Scorer Training sessions in all three content areas. In addition, participants were able to provide open-ended feedback. One ELA teacher who had participated in the spring 2011 scorer training wrote: "I learned how to better apply rubric expectations – in my instruction – by clarifying what those thing 'look like' within student examples." Another wrote, "I learned just how difficult it is for students to do the reflection piece of the project. I will try to construct some models and activities to help students to better accomplish this task." Two mathematics teachers who had participated in the spring 2011 scorer training noted how they would use information from scoring to inform their instruction: "Students did a lot of good work, but had trouble labeling and explaining. Need to emphasize that more in the classroom." "I will be incorporating communication skills in 'bite-sized' chunks into day-to-day teaching; and incorporating smaller 'rich' problems into homework."

Despite evidence that some teachers were beginning to use results from scoring student samples to inform their instructional decisions and plans for ongoing work with their students, this was certainly not the norm. Other teachers struggled with the scoring process, found the scoring rubrics to be difficult to understand and use, or complained about the feasibility of scoring all of their students' work using this approach, given limited time and resources.

There were also practical challenges related to scoring that did not support formative use of the learning tasks. Since most of the scoring took place many weeks after students completed the tasks it was not effectively used as a means for providing students with feedback on their performance, though the data could be used to guide instruction for the remainder of the year or in subsequent years. However, teachers who learned to score a particular task and calibrated during the training sessions in past years will be able to score and provide immediate feedback to their students in the future, since the tasks are scored using the same rubric (in ELA and science inquiry tasks).

As noted above, it is also a practical challenge to score large numbers of student responses, particularly in science and English language arts. There is a need for a more time and cost efficient process for scoring large quantities of student work samples to enable teachers to provide more immediate feedback to the students. At the high school level, assuming that a teacher instructs approximately 30 students per class and 3-5 sections of the same class, a teacher might have 90-150 learning task samples to score at one time. Some of the teachers complained about the amount of time it took them to score one sample. An online system for collecting, scoring, and providing scores and feedback directly to students would facilitate this process. In addition, the online system would need to substantially reduce the time required to score each sample. If it is important that all samples be scored, the system might include opportunity for teachers to score some of their own student's work, some distributed scoring (scoring by other trained raters within a district or state), and some automated scoring (using artificial intelligence scoring methods). This would reduce the volume of work that an implementing teacher would have to score and still support moderated scoring and use of scores for formative purposes.

Table 2. Participant Feedback on Scorer Training Sessions - Spring and Summer 2011 (ELA, Science, and Mathematics)

	ELA		**Science**		**Mathematics**	
	Mean Rating	Percent Agree	Mean Rating	Percent Agree	Mean Rating	Percent Agree
a. Through this session, I have come to better understand the meaning of the rubric and the score levels.	3.2	83.0%	3.5	100.0%	3.3	96.0%
b. Through this session, I have come to better understand the expectations of the performance assessment task.	3.2	87.0%	3.5	95.0%	n/a	n/a
c. The discussion of scores on the anchor performances helped me gain a better understanding of how to apply the rubrics appropriately.	3.2	91.0%	3.6	96.0%	3.2	93.0%
d. The activity of scoring independently and discussing the scores with a partner was valuable for practicing how to apply the rubrics appropriately.	3.3	93.0%	3.8	99.0%	3.3	96.0%
e. I am confident that I can apply the rubrics to score student work fairly and reliably in the future.	3.2	81.0%	3.2	95.0%	3.2	93.0%
f. Our facilitators provided a clear overview of the purposes of the scorer training.	3.5	100.0%	3.6	91.0%	3.3	96.0%
g. Our facilitators effectively led discussions of the anchor performance work samples & helped me to better understand the meaning of the score levels.	3.4	94.0%	3.5	96.0%	3.2	89.0%
h. Our facilitators effectively led discussions about implications for task implementation and instruction.	3.3	90.0%	3.5	82.0%	3.2	91.0%
i. Our facilitators listened to our ideas and were receptive to our feedback about the task and rubric.	3.5	90.0%	3.6	87.0%	3.3	96.0%
j. The opportunity to discuss experiences completing the spring task with other teachers was useful.	3.5	97.0%	3.7	100.0%	3.5	99.0%

Note: Scale is 4=Strongly agree, 3=Agree, 2=Disagree, 1=Strongly disagree, US=Unsure. "Percent Agree : "Agree" + "Strongly Agree"

Discussion & Conclusion

Supporting a Validity Argument for Formative Use of OPAPP Learning Tasks

As noted in the conceptual framework, there are two primary aspects of validity that are supported by evidence discussed in this paper: 1) Substantive theories, process models, and process engagement; and 2) Consequential validity.

Validity Built Into the Design Process. In the first section of this paper, we described in detail the processes that were used to design and develop the OPAPP learning tasks and rubrics. We can articulate several assessment design principles that support validity that were implemented during the design and development phase of this project:

- Utilization of a backwards planning design process, beginning with content and skill standards as the foundation and identifying constructs to be measured (performance outcomes), scoring criteria, and finally tasks;
- Involvement of Ohio educators, stakeholders, content experts, and higher education faculty in the design process so as to ensure maximum "buy-in" and to reflect core academic content, curriculum used by OPAPP participants, and expectations for college/career readiness;
- A continuous review process that incorporated feedback from Ohio educators and system users to improve on and revise the performance outcomes, rubrics, and learning tasks;
- An iterative design process in which pilot results (observed outcomes, feedback from the participants, and analyses of score data) are used to reflect on and revise the design of assessment instruments.

Consequential Validity. Although most of the evidence for this aspect of validity comes from anecdotal evidence rather than systematic data collection, we find that the impact of participation in the OPAPP pilot had mixed results in terms of both teacher learning and student learning. Some teachers reported positive changes in their instructional practices (at least during implementation of the learning tasks), while others struggled to embed the learning tasks into their curriculum in a seamless way because it was so far beyond their normal classroom practice. Some students reported having positive learning experiences with this new format of assessment, while others struggled with the amount of writing or the expectation for demonstrating reasoning and providing explanations, rather than just correct answers. What we find is that the *conditions of implementation* and *conditions for learning* matter, and that the instruments themselves are insufficient for supporting constructive and productive use of the learning tasks. This has a direct bearing on the validity of the formative uses of the learning tasks.

Can the OPAPP Learning Tasks Support Assessment FOR Learning?

While the OPAPP assessment system was originally conceived as a way to provide alternative measures of student learning that could assess knowledge and skills associated with college and career success, it was also designed with formative assessment principles in mind with the goal of supporting both student and teacher learning, as well as organizational learning. However, from our experience with the first two years of the OPAPP pilot, we can conclude that the design of the assessments is insufficient to support formative use of the performance assessments FOR learning and that implementation factors are critical for the ways in which teachers and students experience the performance assessments. It is possible, as we found from results of the pilot, that some teachers will interpret "performance assessment" as just another "test" and implement the performance tasks like tests, with no scaffolding or support for students completing the tasks, either before or during administration. It is also possible for some teachers to over-scaffold the performance tasks and provide too much structure, removing the cognitive demand of the tasks and pre-determining a set of "correct" answers or approaches to constructing a response. Some of these outcomes stemmed from inadequate preparation and professional development for inquiry-based teaching, and in some cases, they stemmed from external constraints placed on teachers' implementation decisions by local conditions and decisions (e.g., requiring texts or curricula; lack of library, technology, and laboratory resources; lack of time to meet and collaborate with other teachers).

In surveys of participating teachers, the external evaluators found that only 50% of ELA respondents, 45% of mathematics respondents, and 39% of science respondents agreed that they were able to embed the tasks in their existing curriculum units with minimal issues (Woodruff, Zorn, Castañeda-Emenaker, & Sutton, 2010a, p.83). The level of agreement for science teacher respondents that they were able to embed the tasks with minimal issues increased from 2.6 in January 2010 to 4.0 (on a 5 point scale) after task implementation (May 2010), while the levels of agreement declined for ELA and mathematics respondents (Woodruff, Zorn, Castañeda-Emenaker, & Sutton, 2010b, p.76)). We also observed wide variation in the ways in which teachers approached task implementation. These findings reinforce the previous observation about the need for more intensive and ongoing professional development in order for teachers to implement the learning tasks well and to adopt "inquiry-oriented" instruction.

Because we did not purposefully set out to study the consequences of participating in the pilot on teacher and student learning, we do not have a robust body of evidence to support drawing strong conclusions about the impact of the pilot on teachers' practice and student learning. As with many pilots in which teachers and students are trying something new, some portion of participants were enthusiastic about the new tasks and began to adopt or adapt some of the new practices, while others found it difficult to adopt the new practices. In many cases, implementation conditions and local factors may have trumped the role of design in determining individual participants' experiences in the pilot than anything else.

Another important external factor that had a clear impact on how the OPAPP tasks were taken up and used was the way in which the state (ODE) articulated the purposes of the pilot and the role of the OPAPP assessments. From the beginning of the pilot, there were mixed messages about the purposes and potential uses of the OPAPP learning tasks. At the beginning of the project, schools and teachers were told that the OPAPP tasks might be used as part of the state's new system of assessments or as an alternative to existing standardized assessments. This messaging put the OPAPP tasks into a summative framework, and likely led some teachers to treat the OPAPP tasks like standardized tests. However, other teachers welcomed the potential inclusion of alternative measures into their state's system of assessments and understood their potential for capturing a more complex set of measurement constructs while honoring the work that teachers do to support students' performance. They did not necessarily equate summative use with standardized administration of a "test" and understood their role in supporting and scaffolding the performance tasks for students.

After the first pilot year (following the release of the Common Core State Standards and announcement of Ohio's participation in the SMARTER-Balanced and PARCC assessment consortia), there was a shift in the conceptualization of OPAPP. In consultation with SCALE, ODE's leadership decided to begin experimenting with the more standardized performance tasks to be included in the Common Core assessments to be designed. The 1-2 week extended OPAPP performance tasks were newly designated as "learning tasks", and ODE introduced a new assessment element into the OPAPP project - "assessment tasks". The assessment tasks were designed to be 60-90 minute, on-demand performance assessments meant to align with the learning tasks in content and construct. SCALE was commissioned by ODE to develop these assessment tasks and these tasks were successfully piloted in the Spring 2011. OPAPP participants were encouraged to continue implementing the learning tasks from the previous pilot year, and were required to score sample and submit data for learning tasks completed in the fall 2010. In the spring 2011, the focus was on piloting, scoring, and collecting data on the assessment tasks. Figure 1 below displays the relationship of the learning tasks and the assessment tasks in this new " task dyad learning and assessment system" (coined by Terrence Moore of ODE).

While many participants in the pilot were disheartened by the change in this framework and by the way the learning tasks were discounted (as a summative measure), this was an important and astute policy move on the part of ODE. There have been many challenges to the use of rich, curriculum-embedded performance tasks for summative purposes due to a variety of technical challenges (score reliability, breadth of content sampling, and issues of task variability and equivalence, to name a few). (See Wei, Cor, Arshan, and Pecheone, 2012 for information on the technical quality of the OPAPP learning and assessment tasks, including reliability and validity evidence.) The question of "whose work is it?" is always raised when it comes to performance tasks that are administered and scaffolded by teachers under non-standardized conditions. The assessment tasks, which are administered under standardized conditions, eliminates the question of "whose work is it?".

Shifting the role of the learning tasks into a formative framework not only reduces the psychometric demands placed on them (reliability and validity), but is also likely to better support their formative use to support teacher and student learning by reducing the anxiety around the use of their scores. By linking the learning tasks to the summative assessment tasks through alignment of content and constructs measured, there is still a policy imperative driving the effective implementation and use of the learning tasks. The learning tasks provide the opportunity to learn the content and skills necessary for success on the assessment tasks. (Opportunity to learn and alignment with curriculum is an important aspect of validity for any summative assessment.) The hope is that by holding teachers accountable to the results of the assessment tasks (the content of which is unknown to teachers until they receive them for administration), teachers will focus on effectively implementing the learning tasks to support student learning and provide opportunities to learn critical college and career readiness skills. Of course, there is no guarantee that the learning tasks will be used as designed or lead to changes in teachers' assessment practices to better support student learning. Participating teachers will still need opportunities for professional development, ongoing support, and local conditions that support inquiry-based instruction and formative assessment practices.

What would it mean to implement the OPAPP learning tasks in a way that supports teacher and student learning? We can derive several principles of performance assessment use and important implementation conditions relevant to student and teacher learning.

To support student learning, teachers:

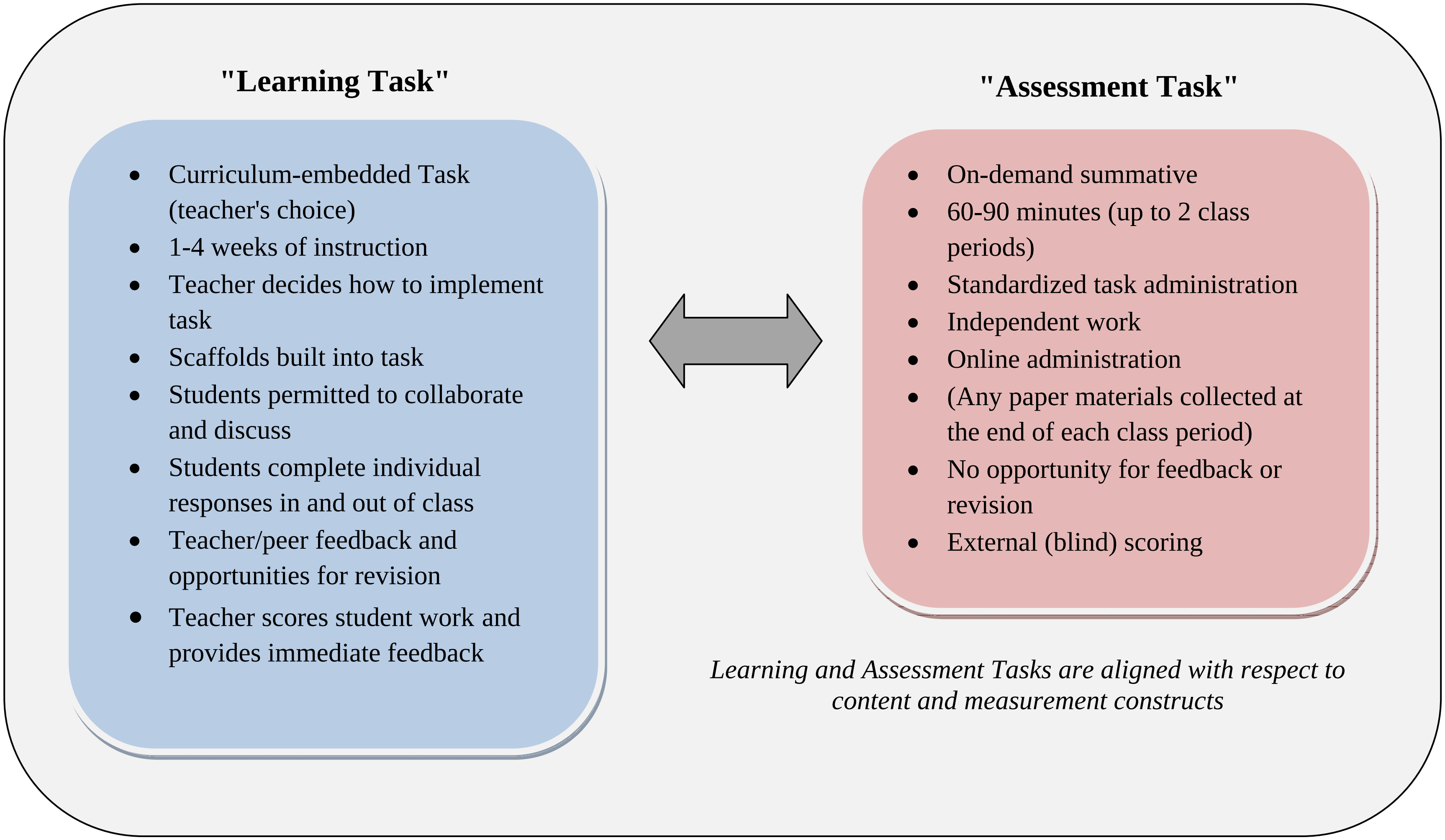
- Provide sufficient instruction before and during implementation of the tasks so that students have opportunity to learn the knowledge and skills necessary for success on the tasks.
- State the learning and skills targets assessed by the performance tasks, as well as the evaluation criteria used to score students' work (scoring rubrics), and revisit them throughout the task implementation to remind students of learning targets.
- Show models of work samples that demonstrate the proficiencies evaluated by the performance task.
- Scaffold the task to support authentic student responses without reducing the demands of the task or pre-determining responses.
- Provide students with feedback (tied to evaluation criteria) on initial drafts or early portions of the task as well as opportunities for revision.
- Present students with opportunities for self and peer-assessment.
- Score student work soon after completion and return scored work to students with scored rubrics and feedback on strengths and weaknesses.
- Analyze the results of students' performance and use this evidence to guide instruction (to revisit content or skills that require additional attention or to inform planning of instruction on future topics).

To support teacher learning, empower teachers with opportunities to:

- Select performance tasks that fit with their curriculum, interests, and characteristics of their students.

- Learn and practice the content and skills necessary to complete the performance tasks to be implemented.
- Gain permission and support from school leaders to make room in their curriculum to complete the performance tasks, including exemption from required reading lists or pacing guides.
- Make decisions about how best to implement the performance tasks within their curriculum.
- Access a community of practice engaged in performance assessment work, and opportunities to collaborate in planning for and reflecting on performance task implementation, including analysis of student work samples and scores.
- Participate in learning to score using a common set of evaluation criteria, engage in calibrating conversations with colleagues within their school, and score at least one class set of their own students' work in order to analyze patterns of performance across students.

Figure 1. Ohio Performance Assessment Pilot Project (OPAPP) Task Dyad Framework



"Learning Task"

- Curriculum-embedded Task (teacher's choice)
- 1-4 weeks of instruction
- Teacher decides how to implement task
- Scaffolds built into task
- Students permitted to collaborate and discuss
- Students complete individual responses in and out of class
- Teacher/peer feedback and opportunities for revision
- Teacher scores student work and provides immediate feedback

"Assessment Task"

- On-demand summative
- 60-90 minutes (up to 2 class periods)
- Standardized task administration
- Independent work
- Online administration
- (Any paper materials collected at the end of each class period)
- No opportunity for feedback or revision
- External (blind) scoring

Learning and Assessment Tasks are aligned with respect to content and measurement constructs

References

Barron, B. & Darling-Hammond, L. (2008). Teaching for meaningful learning: A review of research on inquiry-based and cooperative learning. Retrieved on July 14, 2009 from: http://www.edutopia.org/files/existing/pdfs/edutopia-teaching-for-meaningful-learning.pdf

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Conley, D.T. (2005). *College knowledge: What it really takes for students to succeed and what we can do to get them ready*. San Francisco: Jossey-Bass.

Conley, D.T. (2007). *Redefining college readiness.* Eugene, OR: Educational Policy Improvement Center.

Darling-Hammond, L., Ancess, J., & Ort, S. (2002). Reinventing high school: Outcomes of the Coalition Campus School Project. *American Educational Research Journal, 39*(3), 639-73.

Darling-Hammond, L., Wei, R.C., Andree, A., Richardson, N., & Orphanos, S. (2009). Professional learning in the learning profession: A status report on teacher development in the United States and abroad. Dallas, TX: National Staff Development Council. Retrieved from:

http://www.learningforward.org/news/NSDCstudy2009.pdf

Fadel, C., Honey, M, and Pasnik, S. (2007, May). Assessment in the age of innovation. *Education Week.* May 18, 2007. Retrieved on July 10, 2008 from: http://www.edweek.org/ew/articles/2007/05/23/38fadel.h26.html?print=1

Foote, M. (2005). *The New York Performance Standards Consortium: College performance study*. New York: The New York Performance Standards Consortium. Retrieved on July 14, 2009 from: http://performanceassessment.org/consequences/collegeperformancestudy.pdf

Kane, M.T. 2006. Validation. In *Educational measurement*, 4th ed., ed. R.L. Brennan, 17–64. Westport, CT: American Council on Education/Praeger.

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). *Final Report: Perceived Effects of the Maryland School Performance Assessment Program.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20,* 15-21.

Matthews, B. (1995). *The implementation of performance assessment in Kentucky classrooms*. Louisville, KY: University of Louisville.

Messick, S. (1989). Validity. In R.L. Linn (Ed.) *Educational measurement.* 3rd ed. (pp. 13-103.) New York: Macmillan.

Stevens, D., Sporte, S., Stoelinga, S.R., & Bolz, A. (2008, October). Lessons from high performing small high schools in Chicago. Consortium on Chicago School Research, University of Chicago Urban Education Institute. Research Brief. Retrieved on July 14, 2009 from: http://ccsr.uchicago.edu/publications/08%20Small%20Schools-6.pdf

Stiggins, R. J. (1994). *Student-centered classroom assessment.* New York: Merrill.

Vogler, K. E. (2002). The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. *Education,123(*1), 39.

Wasley, P., Fine, M., Gladden, M., Holland, N.E., King, S.P., Mosak, E., Powell, L.C.. (2000). *Small schools great strides: A study of new small schools in Chicago.* New York: Bank Street College of Education. ED465474.

Wei, R.C., Cor, K., Arshan, N., & Pecheone, R. (2012). Can performance-based assessments be reliable and valid? Findings from a state pilot. Paper presented at the Annual Conference of the American Educational Research Association, Vancouver, B.C., April 2012.

Woodruff, S., Zorn, D., Castañeda-Emenaker, I., & Sutton, J. (2010a). Evaluation of the Ohio performance assessment pilot (OPAP) project phase 1: Final report March 2010. Oxford, OH: Miami University, Ohio's Evaluation & Assessment Center for Mathematics and Science Education.

Woodruff, S., Zorn, D., Castañeda-Emenaker, I., & Sutton, J. (2010b). Evaluation of the Ohio Performance Assessment Pilot (OPAP) Project Implementation Phase - Year 1, October 2010 Oxford, OH: Miami University, Ohio's Evaluation & Assessment Center for Mathematics and Science Education.

APPENDIX A. SAMPLE TASKS AND SCORING CRITERIA

1. OPAPP English Language Arts task: Constructing the Self
(2009-10 pilot version)

2. OPAPP English Language Arts Performance Outcomes
(Scoring Criteria - full rubric not included)

3. OPAPP Mathematics task: Open for Business
(2009-10 pilot version, student task)

4. OPAPP Mathematics Scoring Rubric (Open for Business)

5. OPAPP Science Inquiry task: Chemistry "Got Relieve It?"
(2009-10 pilot version)

6. OPAPP Science Inquiry Performance Outcomes
(Scoring Criteria - full rubric not included)

(Note that Assessment Tasks are not made available as they are considered secure tasks.)

Constructing the Self

This performance assessment begins with the idea that *texts—including, fiction texts—teach*, that we learn lessons about how and who we are to be—for example, as female, Hispanic, athletic, skinny, handsome, American, etc.—from the things we read, look at, listen to, watch, interact with, and discuss. In short, we are shaped not only by biology, but also by the values and expectations that are communicated to us through images and words. This process starts early with family stories, television series, books, and illustrations and it continues through print and media texts, including the magazines, posters and ads that now surround us all our lives long.

In many respects reading, along with the fictions and images it introduces us to, can be a resource that opens up possibilities and loosens the grip of the particular worlds into which we happen to be born. They spell out the history and the richness of who we are and hint at what we might become. Writers such as Alice Walker and Julia Alvarez have drawn our attention to how texts can introduce us to countries, cultures, languages and possibilities we might otherwise never even imagine.

But there are others who have pointed out another, darker side of these "fictions," drawing attention to the potential dangers they create by promoting false stereotypes, expectations, and values. One of these thinkers, the playwright and critic Ariel Dorfman, wrote about what he called the "secret education" these texts provide. He argues that such texts can distort our sense of ourselves, substituting commercial and homogenizing values for more independent ways of thinking:

> *Although these stories are supposed to merely entertain us, they constantly give us a secret education. We are not only taught certain styles of violence, the latest fashions, and sex roles by TV, movies, magazines, and comic strips; we are also taught how to succeed, how to love, how to buy, how to conquer, how to forget the past and suppress the future.*

Becoming reflective about how texts of all kinds influence and shape our self-images allows us, as readers, viewers and listeners, to make choices about the messages we believe and absorb or critique and reject. In this performance assessment, you will have an opportunity to consider and describe the lessons "carried" by a set of popular culture texts that you believe have influenced you, your peers, or shaped the way that others view young adults like yourselves.

Constructing the Self - Revised 2.15.10

The parts of this performance assessment are sequenced in a certain order. Be sure to complete them in order because the work you do in the first parts will help you with the later portions of the assessment. The chart below shows what you will be expected to do and submit at the end of this assessment. The specific prompts for each of the tasks are found in the pages that follow.

TASK OVERVIEW

Task	What You Will Do	What to Submit
Part 1	Read selections by Alvarez and Bordo. Make notes on and compare/contrast the two perspectives represented in these texts.	▪ One page of notes on Alvarez ▪ One page of notes on Bordo ▪ A 1-2 page response in which you compare/ contrast the two perspectives. *Written using ink pen on lined loose-leaf paper (8.5 x11 inch), or typed.*
Part 2	Select, study, and make notes on three texts.	▪ Notes on three texts that address one important aspect of identity. *Written using ink pen on lined loose-leaf paper (8.5 x11 inch), or typed.*
Part 3	Synthesize your perspectives on the three texts that you studied, connecting back to the lenses provided by Alvarez and Bordo.	▪ 1000-1500 word typed essay synthesizing your perspectives on the texts
Part 4	Write a reflective essay on what you learned from completing the performance assessment.	▪ 250-500 word typed Reflection essay

Note: Word count limits are guidelines and are not strict requirements.

NOTE TO TEACHERS: *If there are any terms used that are unfamiliar or unclear to you or your students, please consult the Glossary found in the K-12 English Language Arts Academic Content Standards, available from the Ohio Department of Education (http://education.ohio.gov).*

I. Focus on Perspectives: Two Essays on Fiction and Identity

In this task, you will look at a selection of texts by two different contemporary writers: a chapter from *Something to Declare* by Julia Alvarez and an excerpt from *Unbearable Weight* by Susan Bordo. Each essay presents a perspective on the idea that our identities are formed and informed by what we read, see, and hear. In this task, you will take a close look at each essay in order to understand each author's arguments and to notice how each selects and uses evidence to persuade.

Please take time to review the introductions to Julia Alvarez and Susan Bordo below.

Novelist, essayist, and poet **Julia Alvarez** has written a number of books, including *How the Garcia Girls Lost Their Accents* and *In the Time of the Butterflies*, a book that was eventually made into a movie starring Salma Hayek and Edward James Olmos. She was born in 1950 in New York, but spent the first ten years of her life in the Dominican Republic until, for political reasons, she and her family were forced to leave and return to the United States. Alvarez's experience growing up in and between two cultures shows up in much of her work where she often explores what it is to live between cultures and how that "in-between" experience influences a person's identity and use of language.

Susan Bordo is a Professor of English and Gender and Women's Studies at the University of Kentucky. She is the author of several books, including *Unbearable Weight: Feminism, Western Culture*, and *The Body and Twilight Zones: The Hidden Life of Cultural Images from Plato to O.J*. The excerpt you will read here is from the tenth anniversary edition of *Unbearable Weight*, a book that investigates the way that popular culture artifacts (e.g. television, advertising, and magazines) shape how people think about and see the female body, and connect to disorders such as anorexia nervosa and bulimia.

NOTE: Teachers may select alternate texts (for either the Alvarez or Bordo selections) that make an argument or present a perspective on the idea that our identities are formed and informed by what we read, see, and hear. These substitute texts must be submitted to the Ohio Department of Education prior to use and are subject to approval.

Part 1. Studying Two Perspectives on Fiction and Identities

Complete the following work for EACH essay (Alvarez and Bordo):

- Imagine the author is making an argument about the way texts shape or influence identity. Keep in mind that in an argument a person makes claims about a topic or idea and that they support those claims with evidence of different kinds. What are Bordo's claims? What are Alvarez's claims? What evidence do they draw on to support those claims? Summarize the author's arguments and the evidence used to support those arguments, using a table similar to the one at the bottom of this page (OR in the format your teacher prefers).

Then, in a 1-2 page response, compare and contrast the arguments Alvarez and Bordo make in their essays. In the last part of your response be sure to address the following questions:

- What questions do their arguments raise for you? (Do their arguments make sense? Do you doubt it? Do you see it differently?)
- When have you experienced the weight of the "empire of images" (Bordo)? When have you had a text anchor you, give you hope or open doors for you as the tale of Scheherazade did for Julia Alvarez?

You may work in small groups to study and discuss the texts, but you must complete your notes individually.

Example Table for 1.1

Argument	Evidence that supports the argument
1.	
2.	
3.	
4.	
Etc.	

II. Constructing Selves

Part 2. Studying Texts about How People Should Be

Select a set of three texts from popular culture (e.g. books or other forms of print text, graphic novels, episodes of a television show, movies, an ad campaign, video games, social networking websites, etc.) that influenced your views, whether positive or negative, constructive or destructive, about yourself or people your age (e.g., teen-agers.)

Focus on *one* aspect of your identity, for example, your gender, race, age, or social class. What do these texts teach about this aspect? **At least one text must be a print (written) text, or a written version of a text, for example, a transcript, script, or lyrics). Your teacher must approve your selections.**

Identify three texts that send messages about some aspect of your identity. For EACH text, make a set of notes in response to the following questions:

- What messages does this text send?
- What methods are used in the text to communicate these messages? Be sure to support your answers with references to specific moments in the text.
- How adequate or accurate is this representation? Explain.

In your notes, refer to specific lines or examples in the text to support your ideas. These notes will be submitted to your teacher to be scored as part of this performance task.

You may work in small groups to identify and discuss the texts, but you must complete your notes individually.

Part 3. Synthesizing Perspectives
The goal of this last assignment is to help you synthesize the work you did in the two previous tasks: to make comparisons and judgments about the visions about identity set forth in those texts. Working from the idea that “texts teach,” imagine that each one of the texts you selected in Part 2 contain lessons or arguments about an important aspect of your identity—who you are, as well as how you should relate to other people, what you (and others like you) like, what you *are* (or should be) like, etc.

`

With this perspective in mind, please write a typed essay of 1000-1500 words in which you do the following things:

a. Summarize the arguments about identity contained in each text—in other words, what does each text teach about the aspect of identity you are looking at, and does it teach? Be sure to draw on your notes from Part 2 and to cite lines and examples from the texts to illustrate your answers.
b. Connect your analysis to the arguments made by Alvarez and/or Bordo. For example, you might “talk back to them” in your essay by challenging or extending arguments one or both of them make. You might also use one or both of them as a “lens” you look through to compare, contrast, and critique the arguments about identity contained in the texts you are analyzing. Be sure to draw on your notes from Part One and to cite lines or examples from the texts to illustrate your answers.
c. Pick the text whose perspective you find most compelling in its power to influence or shape one’s identity. Explain this selection and your reasons for making it. Be sure to include in this section your own reflections about how you believe young people should be represented.

You may work in small groups to study and discuss the texts, but you must complete the written task individually. You may also collaborate with other students to revise and refine your writing (e.g., through writer’s workshop).

Part 4. Reflection Task

In a 250-500 word typed essay, reflect on what you have learned from completing this performance assessment. In your response, consider the following questions:

- What did you learn from working with the idea that we learn lessons about who we are and how we are to be from fiction texts of all kinds (as well as non-fiction media)? How did your work with the three texts you chose expand or clarify this thinking?
- What specific activities, processes, or strategies helped you develop and refine your ideas or your writing? Explain how these strategies helped.
- What did you learn about yourself, and/or ways of learning and working that worked or did not work well for you? Do you see ways to apply your learning to your future work or other contexts?
- In what ways could you improve on your work on the Synthesizing Perspectives task, OR if you could do it over, what would you have done differently?

OHIO PERFORMANCE ASSESSMENT PILOT PROJECT
Performance Outcomes – English Language Arts
Inquiry and Communication
2009-2010 Pilot Version

Textual analysis and text production are at the heart of English language arts. We ask students to inquire into texts (defined broadly to include print, digital media, audio/visual media, dramatic performances, and multimedia texts) and to critically examine the ideas presented in a variety of texts for a variety of purposes. Students should develop textual habits of thinking – ways of interacting with and talking about texts that are practiced in post-secondary education, in workplaces, and in other community and institutional settings. Further, we expect students to develop the appropriate skills and understandings necessary to be confident critical readers and thinkers, as well as effective communicators in a global society.

Performance assessments that evaluate students' achievement of these performance outcomes will be tasks designed to engage students in:

- Critical examination and analysis of one or more texts;
- Use of print text, digital media, audio/visual media, dramatic performances, OR multimedia texts, as appropriate, to conduct the inquiry and communicate one's ideas;
- Generation of ideas of their own, based on inquiry, analysis, and synthesis of the ideas in the text(s);
- Production of complex texts of their own;
- Independent and collaborative examination of ideas and communication of those ideas to refine the text;
- Production of multiple drafts or other formative work to show how the student's thinking and quality of the student's text has evolved; and
- Reflection on the process of generating and refining the text

CRITICAL INQUIRY INTO TEXT(S)

Analysis and Interpretation

- Critically examine the ideas presented in one or more texts
- Support interpretations with reasons, examples, and other relevant evidence from text(s)
- Identify, interpret, and analyze literary elements (e.g., figurative language, rhetorical devices) and their impact on meaning
- Demonstrate an understanding of the significance of the texts and how they are situated within their genre, historical/global context, and/or culture
- In research projects, analyze a variety of primary and/or secondary sources, evaluate their accuracy and credibility, and synthesize and present information

Perspective/Position

- Respond to texts with a clear perspective or position that demonstrates engaged reading and critical thinking
- Consider alternative perspectives and ways of thinking and viewing

- Analyze and make connections among multiple perspectives and different points of view from across cultural or global contexts
- Make insightful connections, including connections to one's personal experience, and draw implications and meaningful conclusions as a result of the reading and analysis
- Create/generate new insights, knowledge or information from the inquiry, rather than re-presenting what has been read, viewed, or learned from text(s)

EFFECTIVE COMMUNICATION

What effective communication looks (or sounds) like will depend on the medium used to communicate. However, all types of texts must:

Power of Language

- Effectively use language to communicate one's ideas *to* persuade, convince, or appeal to the audience
- Demonstrate an understanding of how language and images can manipulate responses, shape thinking, and influence judgment
- Communicate with a strong voice and rhetorical techniques that are appropriate to the purpose, context, audience, and medium
- Communicate with clarity and precision

Structure, Organization, and Language Conventions

- Present a clear controlling idea that guides the text's organization
- Effectively organize and connect ideas and information
- Develop ideas and concepts in appropriate depth
- Effectively communicate ideas and information in ways that are appropriate to the specified audience, context, purpose, and medium
- Demonstrate mastery of language conventions and other conventions appropriate to the medium
- Skillfully use print text, digital media, audio/visual media, dramatic performances, or multimedia texts, as appropriate, to communicate one's ideas
- Cite textual evidence accurately and consistently when appropriate to the medium

PROCESS AND REFLECTION

Reflection *DURING* the Process of Textual Production

- Plan, draft, review, revise, and edit one's work to refine ideas and the communication of those ideas
- Independently and collaboratively examine and reconsider one's ideas and communication of those ideas
- Refine one's ideas and the communication of those ideas based on individual reflection on the work and in response to audience responses/feedback

Reflection *AFTER* the Process of Textual Production

- Make thinking visible by reflecting on new understandings or how one's thinking has evolved through the process of textual production
- Go beyond the texts and topics at hand to make connections to other texts/topics/ contexts/disciplines, recognizing the value of integrating diverse approaches to learning and inquiry
- Explain specific thinking strategies used in the process of textual production
- Reflect on the strategies for learning, thinking, and producing text that worked well for the student and what did not work well (meta-cognition)
- Reflect on how the work could be improved in specific and actionable ways, and/or specific strategies or techniques to use in future text production
- Draw on specific evidence from the work to support reflections

GLOSSARY

Critical analysis: A way of reading a text that employs close re-reading and that investigates the relationship of language use to its social/political context and examines how an author uses language to produce meanings and make arguments

Rhetorical techniques: Author's techniques that are designed to persuade or otherwise guide the audience response. Examples include style, voice, text structure, word choice, and tone. Additional examples can be found here: http://writingcenter.tamu.edu/content/view/31/76/

Evaluate accuracy and credibility: Question and analyze a source for its perspective/bias, cross-check a source with empirical evidence or consistency with other sources of evidence, examine what the source says implicitly as well as explicitly, and/or determine whether it is a trustworthy source

Synthesis: Combining ideas/elements into a new whole to predict, invent, redesign, and imagine a new way of looking at something.

Open for Business

Malena is a student who wants to raise $5,000 to tour South America next summer. To raise the money, she decides to open her own business on eBay.

The owner of an electronics shop offers to sell Malena some of his products at the wholesale price. She needs to decide which items to sell and how to price those items in order to maximize her profit.

She does some market research and finds the information provided in the table below about some of the items she is considering selling. Her research results include the cost to buy these items from the wholesale supplier, the retail price at which different items were sold at different times, and the number of items sold at these different prices during the month.

Item	Wholesale Price	Jan. Price	Number Sold	March Price	Number Sold
iPod	150	225	27	200	35
X-Box 360	250	300	41	275	53
Laptop	700	900	15	950	12
Stereo	125	150	21	131	35
Calculator	65	85	31	75	45

Malena also does some research on eBay. She learns that on each item sold, eBay will charge her 8.75% of the initial $25 of the selling price, plus 3.50% of the remaining selling price.

Task Description

Your task is to help Malena decide which items to sell and how to price them to maximize her profit.

She wants to sell some combination of items, and she wants to reach her goal of $5000 profit within a month.

Decide which of the items from the table above will be sold, and what their retail prices will be. Be sure to find the prices that maximize Malena's profit.

Prepare graphs, equations, and a detailed explanation of the calculations you performed to find each price.

Be clear about how you found the price that maximizes profit for each item, and identify how many of each item Malena needs to sell in order to reach her profit goal of $5000.

Helpful Guidelines

You may assume that all shipping costs will be paid by Malena's customers.

On each type of item, Marlena's profit will be the difference between the total revenue (amount received from retail sales of that item) and total cost (amount paid to the wholesale supplier and to eBay). In other words, Profit = Revenue – Cost.

You may also assume that the demand for an item is equal to the number of that item sold.

The demand for an item is related to the price of the item, and you may assume this relation is linear.

A linear demand function has the form $q = mp + b$, where q is the demand (quantity of items sold) and p is the price per item.

Internet Resources

An explanation of linear demand functions is available at http://www.zweigmedia.com/RealWorld/tutorialsf0/frames1_4B.html

An explanation of the relationships among profit, revenue, and cost is available at http://www.zweigmedia.com/RealWorld/tutorialsf0/frames1_4.html

An example of a "revenue problem" is available at http://www.uncwil.edu/courses/mat111hb/Izs/linear/linear.html - sec3

2010 Draft Rubric – Ohio Spring Performance

<table>
<tr><th colspan="2">Open for Business</th><th>Rubric</th></tr>
<tr><td>The core elements of performance required by this task are:
Modeling a supply and demand situation to maximize profits. Using tables and graphs to find a maximum value. Using proportional reasoning and generalization.
Based on these, credit for specific aspects of performance should be assigned as follows</td><td>points</td><td>section points</td></tr>
<tr><td>1. <b><i>Mathematics of the Task</i></b>
Determines the linear relationship of demand and price for Laptop, Xbox and Ipod.
Laptop: <b>q = (-3/50)p + 69</b> Xbox<b>: q = (-12/25)p + 185</b> Ipod: <b>q = (-8/25)p + 99</b>
Stereo: <b>q =(-14/19)p</b> +132 Calculator: <b>q = (-14/10)p + 150</b>
<i>Partial Credit</i>: 1 to 4 errors
<i>Partial Credit</i>: 3 slopes correct

Finds the optimum number, selling price & maximum profit for Laptop, Xbox & Ipod.
<table>
<tr><th>Item</th><th>Optimum No. Sold</th><th>Selling Price</th><th>Maximum Profit</th></tr>
<tr><td>Laptop</td><td>13</td><td>$933</td><td>$2,592</td></tr>
<tr><td>Xbox</td><td>30</td><td>$323</td><td>$1,810</td></tr>
<tr><td>Ipod</td><td>24</td><td>$234</td><td>$1,797</td></tr>
</table>

Shows that the Stereo and Calculator will produce smaller profits.

Determines she can sell <b>13</b> laptops, <b>30</b> XBoxes and <b>8</b> iPods to raise $5,000</td><td>5

(3)
(1)

3x1
3x1
3x1

1

2</td><td>17</td></tr>
<tr><td>2. <b><i>Mathematical Reasoning</i></b>
Derives a linear relationship between <i>demand</i> and <i>price</i>. (q = mp+b)
<i>Partial Credit:</i> correctly derives the slope.

Uses the relationship that <i>revenue</i> equals <i>demand</i> times <i>price</i> (r=qp).

The profit made by selling different items once the demand for those items is modeled appropriately and costs are subtracted (quadratic equation or polynomial relationship).

The amount Malena earns by selling products at different prices, accounting for the percentage taken by eBay (accept earning equation or eBay overhead expression).

Construct graphs of the earnings function

Interprets optimum number sold as a whole number

Reasons how to optimize earnings to raise $5,000</td><td>2
(1)

1

2

2

2

1

1ft</td><td>11</td></tr>
<tr><td>3. <b><i>Approach</i></b>
<i>Maximizing Equation method:</i> Finds an equation that maximizes earnings. Uses a method (perhaps derivative or vertex of the parabola) to determine the maximum.
<i>Partial Credit</i>
<i>Graph method:</i> Determine relationship of price to earnings taking into account the demand, revenue and cost. Creates a graph to estimate the maximum earnings. (trace function)
<i>Table, spreadsheet or a repeated calculation method</i>
<i>Errors</i>: Uses either approach but misses a significant step or steps in the process.</td><td>5

(4)
(3)
(2)</td><td>5</td></tr>
<tr><td>4. <b><i>Communication</i></b>
Clear explanation of the process
Shows table(s)
Shows correct profit or earning equation(s)
Shows graph(s) to illustrate earnings
Summary explanation
<i>Partial credit</i>: maximum earnings that sum close to $5,000 or maximum of top 3</td><td>2
1
1
1
2
(1)ft</td><td>7</td></tr>
<tr><td><b>Total Points</b></td><td></td><td><b>40</b></td></tr>
</table>

Equating Analytical Point Scoring Rubric to Holistic General Rubric

Open for Business

Dimension	*Total Points Available*	*Equating Points to Performance Level*	
Approach	**5 points is the total** • **All 5 pts from part 3**	**Perf. Level**	**Points**
		Level 1	**0**
		Level 2	**2 - 3**
		Level 3	**4**
		Level 4	**5**
Mathematics	**17 points is the total** • **All 17 pts from part 1**	**Perf. Level**	**Points**
		Level 1	**0 - 4**
		Level 2	**5 - 11**
		Level 3	**12 - 15**
		Level 4	**16 - 17**
Mathematical Reasoning	**11 points is the total** • **All 11 pts from part 2**	**Perf. Level**	**Points**
		Level 1	**0 - 3**
		Level 2	**4 - 7**
		Level 3	**8 - 9**
		Level 4	**10 - 11**
Communications	**7 points is the total** • **All 7 pts from part 4**	**Perf. Level**	**Points**
		Level 1	**0 - 2**
		Level 2	**3 - 4**
		Level 3	**5**
		Level 4	**6 - 7**

Ohio Performance Assessment Pilot Project (2009-2010)
Science Inquiry – Chemistry Performance Task

Got Relieve IT?
Student Materials

Your Task

You are an employee for a chemical company called Achoo-B-Gone and your team has been working for the past year to create a new drug that will instantly relieve cold symptoms. The new product, "Relieve IT", is in the final testing stages before being sent to the Food and Drug Administration (FDA) for human trials. Part of the FDA approval requires your team to share your current knowledge about acids and bases and to provide all of your experimental data on "Relieve IT".

As a part of the approval process, the FDA conducted a preliminary test on the pH of Relieve IT and reported some concerns about potential negative human side effects. Unfortunately, the report did not indicate whether the product is too acidic or too basic. The FDA wants to know what you are going to do to "fix" the product before beginning human trials. Your team will synthesize your current knowledge about acids, bases, and neutralization. You will design and conduct an experiment to determine the pH of the product and to determine which solution (A, B, C,) or combination of solutions can be used to neutralize any excess acid or base. You will prepare an individual formal lab report (your teacher will provide the format) including recommendations for "fixing" the pH levels of "Relieve IT."

Task Overview

Task Part	What You Need To Do	Product
1	Prepare an introduction for the FDA application (lab report)	Lab Report
2	Design an experiment to determine the pH of Relieve IT	
3	Conduct the experiment	
4	Analyze and interpret your findings	
5	Draw your conclusions	
6	Reflect on the Findings	
7	Prepare final application to FDA	
8	Reflect on Learning	Essay
9	Group Presentation	Optional *

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

Part 1: <u>Research and</u> <u>prepare an introduction to your lab report sharing what you have learned about acids and bases and the process of neutralization</u>. (**Individual Activity**). Collect, analyze, and synthesize information from at least four credible and reliable sources. For each source remember to indicate any potential sources of bias and take into account the perspective of the author. Based on your research, your introduction should:

- Explain the significance of acids and bases;
- Describe what you learned about acids and bases and what it means to neutralize an acid from conducting the lab.

Part 2: Plan the design of your experiment. **(Lab Partner)**. Prepare detailed procedures for testing the pH levels of the product based on what you have learned about acids, bases, and the neutralization process. Then proposed a procedure to determine which solution or combination of solutions might help your team to neutralize "Relieve IT". **In your individual lab report:**

- State the problem or question. In your own words, state the problem or question you are going to investigate;
- State a hypothesis with an explanation of your current thinking. Write a hypothesis using an "If … then … because …" statement that describes what you expect to find and why;
- Clearly identify all the variables to be studied (independent and dependent variables including controls if applicable);
- Plan out your experiment. Your experimental design should match the statement of the problem and should be clearly described with sufficient detail so that someone else could easily replicate your experiment. Include in your design the:
  - materials to be used
  - specific procedures including exact quantities of substances
  - appropriate tools and techniques to be used to gather data,
  - appropriate sampling and number of trials
  - need for safety precautions when conducting your experiment
- Show your design and lab procedures to your teacher to check for any safety issues. Once you have your teacher's safety approval record this information into you lab report.

Part 3: Conduct your Experiment. (**Lab Partner**). While conducting your experiment, take notes of any changes you make to your procedure, record all relevant data, and indicate the number of trials performed during the experiment. **Record this information in your individual lab report.**

Part 4: Analyze and Interpret your Findings. (**Individual Activity**). This is an essential part of your experiment. You need to careful examine the data you have collected and determine what you can say about the results of the experiment based on the evidence. Include the following steps in your analysis:

- Perform calculations and/or make estimates to understand your data (i.e., converting units, taking an average, etc.) when appropriate;
- Organize the data into charts, tables, and/or graphs where appropriate. Remember to properly label everything and provide a key/legend when applicable;
- Describe and explain any patterns and/or trends that you notice when examining the data;
- State in your own words the results from the experiment and specifically cite evidence from the data to support your explanation'
- **Record this information in your individual lab report.**

Part 5: Draw Your Conclusions. (**Individual Activity**). Review your analysis and interpretations of the data and write the conclusion section of the lab report. In the conclusion be sure to:

- List your findings using data to support your statements;
- Discuss any potential sources of error and explain how that error might be eliminated or reduced in future investigations;
- Identify the limitations of the findings and explains how those limitations might be addressed in future experiments;
- Develop a scientific explanation that is fully supported by your data and addresses your hypothesis. Make connections between your findings and the appropriate scientific content;
- **Record this information in your individual lab report**.

Part 6: Reflect on the Findings. (**Individual Activity**). Based on your conclusions, reflect and comment on:

- Potential implications of your findings (applications, policy decisions, and implications of your investigation -remember to address the concerns for FDA);
- Generate a list of other scientific explanations and explain whether they are supported and/or refuted by the data;
- New questions or unanswered questions that were generated during this study that you would like to explore in future investigations;
- Next steps to answer those question by either modifying your investigation or developing a new design (be specific about your ideas);
- **Record this information in your individual lab report.**

Part 7: Prepare final lab report to the FDA panel . (**Individual Activity**). You will need to submit a final lab report to the FDA panel before they will continue with the approval process. The formal lab report (get format from teacher) and must include:

- Research Question with appropriate background information (Part 1);
- Hypothesis with explanation of why this is your current thinking (Part 2);
- Design of Experiment including a list of all variables, materials, and detailed procedures (Part 2);
- Presentation of data -Tables, Graphs, Visuals (Part 4);
- Analysis and Interpretations (Part 4);
- Conclusions including possible error, limitations, and future investigations (Part 5 and 6);
- Address concerns of FDA citing evidence from your investigation to justify why your product will not be harmful to humans (Part 6);

- Check any written materials and visuals to ensure that you have used proper chemical formulas and proper scientific convention ;
- Cite all of your references using the APA format or the format selected by your teacher.

Part 8: Reflect on Your Learning. (**Individual Activity**). Write an essay reflecting on your learning, specifically address what you:

- Learned about the properties and relationship between acids and bases;
- Discovered about your ideas and how those ideas evolved over the course of completing this performance assessment;
- Used as strategies for learning, thinking, and producing work that were effective and those that did not work so well;
- Leaned about investigative skills and/or your understanding of scientific inquiry;
- 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 Group Activity**). You will be asked to make an oral presentation to an FDA panel sharing what you learned from your investigations and making recommendations on which solution can be used to address the FDA's concern about the pH level of Relieve IT. When preparing your presentation:

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

OHIO PERFORMANCE ASSESSMENT PILOT PROJECT
Performance Outcomes – Science Inquiry
2009-2010 Pilot

The performance task will engage students in a scientific inquiry or investigation. The task will require students to research a specific science topic including the relevant standards-based science content and to apply that content knowledge to perform an investigation. The investigation may ask students to design and conduct an experiment, carry out some portion of an experiment, and/or analyze and interpret data from external sources (e.g.,NSF data) using appropriate quantitative or qualitative reasoning skills. Students will summarize their findings, reflect on the process and on their learning, and communicate their explanations effectively.

Collect and Connect Content Knowledge

- Identify and evaluate the significance of a topic (problem, issue, phenomenon, and/or technology)
- Compare and synthesize information from a variety of sources to investigate and explain the scientific content relevant to the topic
- Examine the credibility and accuracy (reliability) of the information by indicating any potential bias, when appropriate
- Demonstrate logical connections between the scientific concepts, the purpose of the investigation, and the design of the experiment

Design and Conduct Investigation *(NOTE: Depending upon the type of investigation, not all the sections may be applicable.)*

- Pose an appropriate and testable question
- State a hypothesis including a clearly stated rationale
- Identify variables that will be measured, including independent, dependent, and those held constant (controlled variables) as appropriate
- Plan and follow appropriate scientific procedures (use appropriate tools and techniques, collect relevant data including the appropriate units of measure, and conduct multiple trials when possible)

Analyze and Interpret Results

- Apply appropriate computation and estimation skills necessary for analyzing data
- Organize scientific information (data) using appropriate tables, charts, graphs, etc.
- Identify and describe patterns and relationships between variables using qualitative reasoning and appropriate quantitative procedures, including mathematics
- Interpret findings based on data analysis

Draw Conclusions

- Summarize the findings
- Identify potential sources of error and limitations of the data
- Formulate a cohesive scientific argument or explanation based on evidence from the investigation
- Connect findings to relevant scientific content and other key interdisciplinary concepts to demonstrate a broader understanding of the finding, when appropriate

OHIO PERFORMANCE ASSESSMENT PILOT PROJECT
Performance Outcomes – Science Inquiry
2009-2010 Pilot

Reflect on the Findings

- Discuss implications of the findings (i.e., applications, policies, solutions, social considerations) when appropriate
- Identify alternative scientific arguments or explanations and explain how they are supported or refuted by the data, when possible
- Generate new questions and next steps based on investigation results

Communicate and Present Findings

- Communicate reasoning and findings clearly using the appropriate vocabulary, symbols, and conventions of science
- Design visual aids that clearly presents relevant information highlighting the key components of the investigation
- Cite sources of information properly
- Present findings in a clear, concise, engaging, and coherent manner appropriate to the audience

Reflect on the Learning Process

- Reflect on personal growth in terms of content knowledge and learning process throughout the investigation
- Discuss how this project has impacted personal investigative skills and understanding of scientific inquiry
- Identify your teams' and/or partners' strengths and recommend areas for improving while working on the collaborative portions of the task