Running Head: RELIABILITY AND VALIDITY OF PERFORMANCE-BASED ASSESSMENTS

Can Performance-Based Assessments be Reliable and Valid? Findings From a State Pilot

Ruth Chung Wei

Ken Cor

Nicole Arshan

and Raymond Pecheone

Stanford University

Can Performance-Based Assessments be Reliable and Valid? Findings From a State Pilot

Introduction

Over a three year period (2008 - 2011), the Ohio Department of Education (ODE), in collaboration with Stanford University[1], undertook a pilot project across the state 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. The state'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. Another goal of the project is to promote changes in instructional practice and enrich students' learning experiences in ways that lead to higher levels of student success in college and beyond.

During the academic years of 2009-10 and 2010-11, the state completed pilots of curriculum embedded performance assessments in English language arts, mathematics, and science with approximately 140 teachers across 31 schools, and completed research to explore the validity and reliability of these performance tasks as a possible set of measuresin a multiple-measures assessment system. In light of the recent release and adoption of the Common Core State Standards across 45 states, and the agreements among two state consortia (Smarter Balanced Assessment Consortium and the Partnership for Assessing Readiness for College and Careers) to develop and field test high quality student assessment systems aligned to the

[1] The Stanford Center for Assessment, Learning, & Equity (SCALE) was the research lab associated with this project.

Common Core, a number of states have inquired into this performance assessment pilot in Ohioand lessons learned.

The purpose of this paper is to present findings regarding the reliability and validity of the performance based assessments used in the state pilot, including the results of three sets of studies: 1) Generalizability studies; 2) Regression analyses to ascertain the relationships between performance assessment scores and other measures of student achievement (high school GPAs, exit exam scores, ACT scores); and 3) Content alignment and content validity reviews.[2]

This state performance assessment pilot and the results of the studies described herein provide preliminary evidence that performance assessments can be scored reliably by teachers (with sufficient training and using a blind scoring approach). This paper also builds an argument around the generalizability and transfer of the performance assessments based on their design features. Evidence from the pilot suggests that these performance assessments provide new information about student achievement (in particular, measurement of features of college readiness and 21st century skills) that are not currently available in scores from Ohio's high school exit exams and course grades. However, there is a high level of variability in both score profiles and relationship to other academic indicators across performance tasks. The results of these studies are also used to propose principles of design and use that could improve the validity and reliability of the performance assessments and scoring instruments.

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),

[2] While the early draft of this paper (uploaded to AERA by March 20) does not include full results from the higher education content validity review, these results will be added to the final conference paper and uploaded before the conference.

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, well designedperformance-based tasks have the potential to measure these cognitive abilities more directly. 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). Research on thinking and learning processes also shows that performance-based assessments propel education systems in a direction that corresponds with how individuals actually learn (Herman, 1992).

Despite the theoretical potential of performance based assessments and empirical evidence of positive outcomes for student learning and achievement in small scale settings, there remains a wide-spread public skepticism that performance assessments can validly and reliably assess student learning on a large scale with students in regular public schools, especially when teachers themselves are involved in scoring these performances. There are also a number of technical issues that continue to raise questions about the validity of performance-based assessments. One challenge relates to the generalizability of performance on one task and

comparability of performance tasks. Shavelson, Baxter, and Pine (1990) investigated the performance of students on different hands-on science tasks and found that performance depended on the task completed. In another study,Shavelson, Ruiz-Primo, and Wiley (1999) found that while certain types of tasks - direct observation, notebook and computer simulation methods - appeared to be comparable,they were not equivalent due to the volatility of student performances across tasks and occasions.In other words, the tasks students complete matter because a student's relative performance on one task could be inconsistent with their relative performance on another task.

Another criticism about performance based assessments is that they sacrifice content breadth in favor of focused measurement of a few constructs. Concerns about classroom -based performance assessments also arise, including the question of "whose work is it?" when student collaboration and independent work outside of the classroom are permitted, and when teachers support completion of the performance assessments through instructional scaffolding. These technical challenges, along with concerns about cost effectiveness and feasibility of scale-up, continue to plague attempts to use performance-based assessments as part of any large-scale assessment system that plays any role in measuring student achievement.

Evaluating Validity

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 ona few of these commonly defined criteria for supporting a validity argument, based on the availability of data -- *content relevance and representativeness; scoring models as reflectiveof task and domain Structure; generalizability and the boundaries of score meaning; convergent and discriminant correlations with external variables; and fairness.* In another paper presented at this conference (Wei, Schultz, and Pecheone, "Performance Assessments for Learning: The Next Generation of State Assessments"), two other criteria for validity cited by Messick (1989) are addressed: consequential validity; and substantive theories, process models, and process engagement.

Evaluating Reliability

*You cannot have validity without reliability.*Alternative frameworks for the evaluation of reliability in performance assessmentshave challenged traditional approaches to evaluating reliability with important implications for building a validity argument (see for example Moss, 1994). However, there is also a policy imperative to quantify the reliability of scoring when performance assessment scores are used for consequential decisions, high-stakes or otherwise. Therefore, we draw on classical test theory methods (Generalizability theory) to evaluate the sources of score variation attributable to students, tasks, raters, and the interactions among these

facets of variation. The results of these analyses are critical to understanding the efficacy of the scoring models, the construct validity of the scoring domains, and for inferring how the performance task scores may be used validly for decision-making.

The Ohio Performance Assessment Pilot Project (OPAPP) Tasks

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 by 147 teachers (approximately 40-50 in each content area) from those schools[3].Learning tasks were designed with the input of each content-specific design teams, comprised of 20 teacher representatives, 4 coaches, 1-2 higher education advisors, 2-3 content specialists from the ODE, and representatives of exceptional students - special education, English learners, and gifted students (See Appendix A for a sample of a few of these learning tasks).

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 reported in this paper. A second round of piloting was completed in 2010-11, with the addition of a new component of the OPAPP system -- "assessment tasks". These 60-90 minute on-demand assessment tasks - part of the "task-dyad learning and assessment system" (a term coined by Terrence Moore of ODE) -- were developed and piloted in the spring of 2011. Assessment tasks are on-demand performance-based tasks intended to be used as a summative measure of students' performance on specific constructs. The assessment tasks are aligned with the content and constructs measured in the learning tasks. The relationship between the "learning tasks" and "assessment tasks" is depicted in **Figure 1** below. This paper reports results of analyses of score data from the spring 2011 "assessment

[3]Four English language arts tasks, seven mathematics tasks, and eight science tasks were designed in the spring of 2009 by expert task designers employed by the Stanford Center for Assessment, Learning, and Equity (SCALE).

tasks" - including G-studies as well as correlationsbetween the assessment task scores and the learning task scores (using available data from one task-dyad).

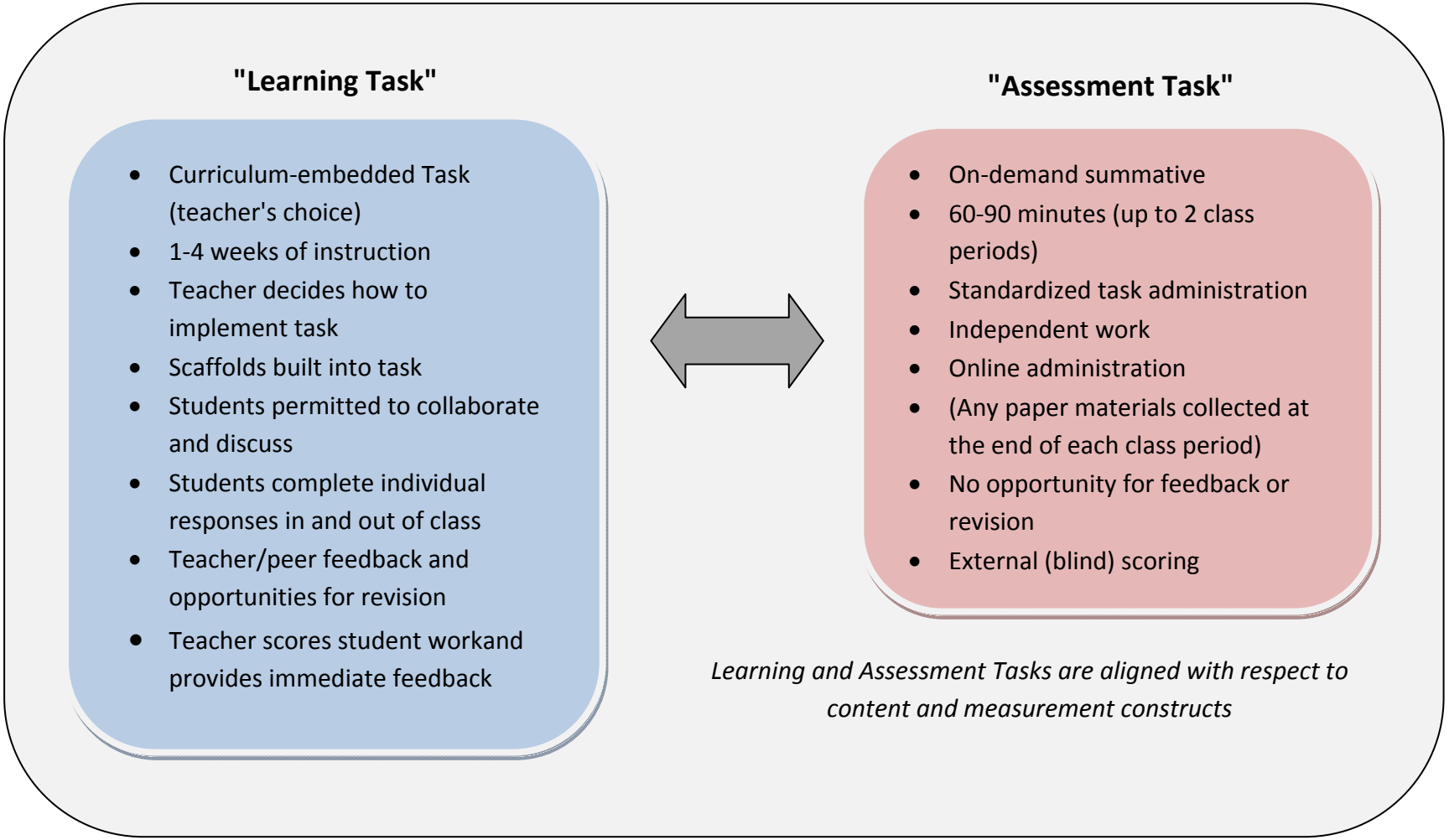


Figure 1. Ohio Performance Assessment Pilot Project (OPAPP) Task Dyad Framework

Methods

Data Sources

This paper reports on findings of several studies conducted based on score data collected from the OPAPP pilot of curriculum-embedded performance assessments completed by 11th and 12th graders in English language arts, mathematics, and science. During the spring of 2010, teachers who participated in the pilot were trained to score the particular tasks that they had piloted that spring. ELA and science teachers in the pilot then scored at least one class set of their own students' work and submitted the scores and samples. Mathematics teachers were able to score most of the submitted samples during the two-day scoring session through a blind, distributed scoring method. These student-level scores were linked using unique student identifiers with data on other academic indicators such as students' high school GPAs, their Ohio Graduation Test scores, and ACT scores.[4]

G-study Data for Learning Tasks. During the spring-summer of 2010 scorer training sessions (a two-day session), a subset of calibrated teachers were randomly selected to participate in scoring a common set of work samples for the G-studies. A small subsample (15-20 samples) of work samples of at least two tasks in each content area was randomly selected from across the class sets submitted by teachers. For the G-studies, 4-5 raters for each task were randomly selected from among OPAPP teachers who had previously participated in scorer training for a particular task and were judged to be calibrated during scorer training. Raters selected for these G-studies were re-calibrated ifthe scoring session occurred long after the

[4]High school GPAs and ACT scores were obtained from students' transcripts provided by their schools, while the Ohio Graduate Test scores and demographic data were obtained from the Ohio Department of Education through a data-sharing agreement.

training session.[5] For English language arts and science, two of thefour piloted tasks in each content area were used as the focus of the G-studies due to limited resources and time. In English language arts, the tasks *Americans Dreaming* and *Constructing the Self* were selected for the G-studies. In science, the Chemistry task *Got Relieve It?*and the Physics task *How It Works* were selected for the G-studies. In mathematics, a subsample of work from all three of the tasks piloted in Spring 2010 - *Maximum Volume*, *Open for Business,* and *Wheelchair Ramp* - were scored in the G-studies.

G-Study Data for Assessment Tasks. In the Spring of 2011, teachers administered the assessment tasks designed to be aligned with the learning tasks piloted that same spring, and scores were collected during the scorer training and distributed scoring sessions planned and executed by ODE in April and May 2011.[6] Those scores were used for the G-studies described below. Scorers were OPAPP teachers, coaches, and ODE staff with content expertise who were trained and calibrated to score by SCALE content leads. Following a training module that was about 2 hours in length, trained raters completed scoring almost allassessment task work samples submitted during the remainder of the day. Approximately 1750 math assessment task samples were double scored scored by 33 mathematics raters in a single day, while approximately 550 science assessment task samples were double scored by 19 science raters in a single day. The distributed scoring method was designed to produce two sets of scores for every work sample and to match every rater with every other rater multiple times. The datasets that resulted were mined to extract multiple sets of scores in which the same pairs of raters rated a common set of samples. These smaller subsets of scores were used for the G-studies.

[5]Re-calibration involves having raters independently score a work sample, comparing their scores and evidence, and coming to consensus on the final scores for that sample.
[6]Only the mathematics and science assessment tasks were piloted and scored that spring because the assessment tasks for English language arts were developed during the summer of 2011.

Generalizability and DecisionStudies

In order to investigate the reliability of teacher raters on the mathematics, English, and science rubrics designed and used by the OPAPP teachers to score student work, techniques from Generalizability Theory (See Cronbach, Gleser, Nanda, and Rajaratnam, 1972 or Brennan, 2001) are employed. Generalizability theory facilitates the investigation of how different sources of variation in the measurement design contribute to the reliability of scores produced from the rubrics. For the purposes of our analysis, we consider whether randomly selected raters (from among calibrated teachers) contribute variance to scores over and above that which is attributable to measured differences in student ability. Results from these studies should be generalizable to teacher raters who are trained using the same methods of training to be certified as a calibrated rater.

For each task and its associated rubric, separate Generalizability studies (G-studies) and Decision studies (D-studies) are performed. G-studies produce estimates of the variance contributions of a single anatomical unit of each factor included in the design for each item (row) in the rubric. In each analysis, variation from students, raters, and error is considered. In our design of the G-studies for the learning tasks, a conscious decision was made to treat item as a fixed facet because each task was designed to measure a particular set of constructs of performance and to be scored using a pre-defined (finite) set of evaluation criteria. That is, the items (the rows of the scoring rubrics) were viewed as the entire universe of items that could be used to assess student performance for the given dimensions and as a result there is no sampling error due to the item facet. Therefore, when calculating composite reliabilities for different combinations of items, the only source of random error used is the combination of estimated error components (sxr,e) for the items that make up each dimensional and total score.

In contrast, the G-studies of the assessment tasks utilize a three-facet student by item by rater (s x I x R) univariate measurement model because the assessment tasks were designed differently. That is, for the learning tasks an effort was made to identify a fixed set of sub-dimensions that make up the underlying constructs of interest. For example, the science tasks are designed to measure students' overall science inquiry ability such that it is made up of of six componentsthat represent all aspects of the science inquiry process and are generalizable across all OPAPP science inquiry learning tasks. In contrast, the assessment tasks were designed to measure a subset of science inquiry ability (e.g., evaluating the design of an experiment, evaluating the soundness of conclusions drawn, making measurement calculations, identifying potential sources of error), with reference to a particular task (and the scoring rubrics are unique for each science inquiry assessment task). As a result, the items included on the assessment tasks can be viewed as a sample of the universe of possible items that are needed to fully assess science inquiry. This fact results in potential item sampling error from task to task. In other words, the items selected for one task could be more difficult than the items selected for another task designed to measure the same construct. Therefore, the measurement model used as a basis to conduct the G-studies for the assessment tasks include estimates of the amount of variance each of the following factors contributes to observed score variance: student, item, rater, student by item, student by rater, item by rater, and error.

Next, as a part of the G-study results,dissattenuated correlation matrices for the items in each rubric are also produced to facilitate the consideration of the relationships among the items after accounting for unreliability due to measurement error. In rubrics with multiple items measuring separate dimensions, the disattenuated correlation matrix provides evidence about whether items relate to each other as they have been specified in the design of the rubrics.

Next, D-studies are performed. The results from each D-Study are derived from the separate G-studies and produce estimates of the expected reliability of scores for different measurement designs. D-studies facilitate the modeling of how the reliability of scores change as a function of the number of levels of different factors in a G-study. For example, it is easy to imagine how the reliability of average scores for students on each item of a rubric increases as a function of the number of raters evaluating each performance. For our analysis, D-studies permit the production of reliability curves for individual items as well as for groups of items as a function of the number of raters used to evaluate each task.

Concurrent validity

Score data from the spring 2010 pilot were linked to other measures of achievement from students' high school transcripts (including cumulative GPAs and ACT scores) and achievement on the state graduation tests. Through a data sharing agreement made at the beginning of the pilot, the Ohio Graduation Test scores and demographic information (e.g,. gender, race/ethnicity, LEP status, disabilities, economically disadvantaged) were obtained from the ODE, while cumulative high school GPAs and ACT scores were obtained from high school transcripts collected from each pilot site. The State Student ID (SSID) was used to identify and link these data for each individual pupil to the OPAPP performance assessment scores. Unmatched data (due to mismatches in SSIDs or incomplete transcript/data collection) led to a decline in the total sample size of performance assessment scores that could be analyzed. Data about schools (e.g., size, location, percentage of ethnic minority students, percentage of free and reduced lunch eligible students) was gathered from the Common Core Data (available from the U.S. Department of Education) or from individual schools when those data were not available from either the Common Core Data system or the state's online data systems.

For each set of performance task scores, multiple regression models were used (due to the lack of sufficient sample sizes to allow for fixed effects multi-level modeling) to assess the contribution of these other measures to performance task scores, controlling for student demographic characteristics such as gender, non-white minority status, free and reduced lunch status, and special education status. These analyses allow us to assess the degree to which performance assessment scores were distinct from or highly correlated with other measures of student achievement, allowing us to evaluate the degree to which the performance assessment scores provide new information about student learning that is not provided by other measures. These analyses also allow us to ascertain differences in scores across demographic groups.

Content Validity

We report on expert reviews of the performance tasks and scoring rubrics that were conducted by an external evaluator contracted by the Ohio Department of Education to assess their alignment with valued standards (the Common Core State Standards, the state's content standards, 21st century skills). In addition, we conducted a separate review of the performance tasks, scoring rubrics, and student work samples produced during the pilot using higher education faculty as reviewers to assess the relationship between the performance assessments and entry level expectations for college students in content-specific courses (English language arts, science courses, mathematics courses).

Results

Generalizability and Decision Studies

For each learning task included in the study, scores generated from four or five raters are analyzed within the framework of a fully crossed student by rater multivariate design. In other words, each rater rates all students on all items and the items are treated as separate variables.

Using this design, the results for each task are presented in the same format. First, the source table for each variance component and the disattenuated correlation matrix from a G-study are presented. Second, the results from separate D-studies are summarized in 1) a table showing the expected reliability of scores for each item, each sub-dimension (if there are sub-dimensions), and the entire rubric and 2) reliability curves that show how the reliability of different components of each rubric change as a function of the number of raters. The mathematics, English, and science tasks are discussed separately.

Mathematics Learning Tasks. The scores collected for the G-studies for the mathematics learning tasks were embedded in the regular scorer training and scoring session conducted in the spring of 2010. Following scorer training, four teachers who met calibration standards for scoring a particular task by the trainers were randomly selected to score the samples selected for that task. G-studies were conducted for three mathematics tasks - *Maximum Volume*, *Open for Business*, and *Wheelchair Ramp*. For each task, four raters independently scored the same set of 20 work samples representing work selected from across submitting teachers. In this paper, we report on the results for the task *Open for Business*which is assessed on 19 items across four scoring dimensions. We briefly compare results for this task with those obtained with the other two mathematics tasks.

Open for Business (mathematics task). This task has four sub-dimensions (Mathematics, Mathematics Reasoning, Approach and Communication) measured by 6, 7, 1, and 5 items, respectively. (Note that the four dimensions of performance were predefined as evaluation criteria for scoring and guided the design of the tasks, but the task-specific scoring "items" for mathematics tasks were designed post-hoc and then back-mapped to the four dimensions of performance. In other words, the items were not designed specifically to evaluate

the four dimensions in a systematic way - evidenced by the varying number of items used to score each of the four dimensions.) Table 1 shows the absolute estimated variance contributions from each factor for each item as determined by the G-Study. Table 2 shows the same information but as proportions.

The absolute variances reported in Table 1 show that students' scores vary on all items of the *Open for Business* task with the exception item 1f. This item appears not to be producing meaningful variation in scores. The remaining items have absolute variance contributions from the student factor that range between .03 and 3.75. The rater row of Table 1 suggests that there is very little absolute variation due to rater. In other words, raters appear to be equally stringent on the items. Finally, all items produce error variance as indicated by the non-zero contributions in the error row. This suggests that items are susceptible to either a rater by student interaction or other sources of error not captured in the measurement design or both.

Table 2 shows the same results but as proportions. Items 1a through 1d, 2e and 4d are likely to be the most reliable items given the large relative contributions from the student factor. Taken together, these results suggest 1) there is measureable variation in students' scores on 18 of the 19 items, and 2) scores from all items contain undocumented error. We now turn to results that speak to how the item scores relate within and across dimensions.

Table 3 shows the disattenuated correlations for the 19 items of the *Open for Business* task. The grey cells show the distinction between dimensions (with the items allocated within dimensions) assessed in the *Open for Business* rubric.

Table 1

Source Table for student• x rater• Multivariate Item Design Based on Four Raters for the Open for Business Task

	Item																	
Source	**1a**	**1b**	**1c**	**1d**	**1e**	**1f**	**2a**	**2b**	**2c**	**2d**	**2e**	**2f**	**2g**	**3**	**4a**	**4b**	**4c**	**4d**	**4e**
Student	3.75	1.59	1.67	1.47	.13	.00	.53	.03	.33	.36	.77	.10	.16	.89	.59	.19	.09	.20	.24
Rater	.05	.03	.01	.00	.01	.02	.03	.00	.01	.00	.01	.02	.01	.00	.00	.00	.02	.00	.01
Error	1.11	.31	.23	.39	.10	.06	.28	.10	.51	.37	.11	.14	.12	2.48	.29	.07	.15	.03	.09

Table 2

Source Table for student• x rater• Multivariate Item Design Based on Four Raters for the Open for Business Task as Proportions

	Item																	
Source	**1a**	**1b**	**1c**	**1d**	**1e**	**1f**	**2a**	**2b**	**2c**	**2d**	**2e**	**2f**	**2g**	**3**	**4a**	**4b**	**4c**	**4d**	**4e**
Student	.76	.82	.88	.79	.53	.00	.63	.22	.39	.49	.87	.38	.56	.26	.67	.71	.34	.88	.70
Rater	.01	.02	.00	.00	.05	.28	.04	.00	.01	.00	.01	.09	.03	.00	.00	.01	.07	.00	.03
Error	.23	.16	.12	.21	.41	.72	.33	.78	.61	.51	.13	.53	.42	.74	.33	.28	.59	.11	.27

Table 3

Disattenuated Item Correlations (Diagonal Elements are Variances) - Open for Business Task

	1a	**1b**	**1c**	**1d**	**1e**	**1f**	**2a**	**2b**	**2c**	**2d**	**2e**	**2f**	**2g**	**3**	**4a**	**4b**	**4c**	**4d**	**4e**
1a	3.75	.57	.55	.56	.59	.00	.66	.63	.88	.81	.59	.05	-.05	-.14	.48	.44	.78	.60	-.11
1b		1.59	.87	.86	.54	.00	.58	.44	.67	.51	.85	.41	.29	.24	.65	.59	.74	.86	.22
1c			1.67	.99	.70	.00	.41	.30	.57	.37	.88	.43	.32	.48	.62	.58	.82	.79	.35
1d				1.47	.73	.00	.43	.31	.59	.38	.84	.44	.33	.49	.62	.60	.78	.75	.33
1e					.13	.00	.33	.87	.58	.61	.62	.59	.68	.40	.84	.78	.84	.46	.59
1f						.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
2a							.53	.48	1.00	.96	.46	.32	.12	.20	.56	.51	.60	.51	.17
2b								.03	.90	1.00	.48	.26	.54	.38	.66	.35	.90	.48	.45
2c									.33	1.00	.62	.20	.02	.42	.52	.35	.87	.68	.11
2d										.36	.45	.33	.17	.29	.54	.42	.84	.48	.27
2e											.77	.39	.40	.06	.72	.53	.84	.98	.35
2f												.10	1.00	.42	.85	.90	-.01	.39	1.00
2g													.16	.39	.85	.87	.10	.32	.96
3														.89	.23	.32	.39	.02	.55
4a															.59	1.00	.58	.66	.72
4b																.19	.42	.48	.69
4c																	.09	.75	.07
4d																		.20	.29
4e																			.24

Based on the patterns of correlations within the grey cells of Table 3, it appears that for the most part the items are measuring a single underlying dimension. That is, in general, the items within the grey cells correlate moderately well to strongly with each other while also correlating moderately well to strongly with the remaining items. That being said, items 2f, 2g, and 3 could be measuring something unique given the lower overall correlations with most of the other items. The correlational evidence does not appear to support these groupings of items for the purpose of measuring specific constructs. Instead, there is more evidence to support treating the scores as measuring the same underlying construct. When these results were shared with the designer of the task-specific scoring rubrics, he was not surprised by this outcome and suggested that the task-specific scoring items altogether are designed to measure one particular construct (mathematical problem solving) rather than a set of differentiated constructs.

The indexes of reliability from a series of D-Studies are summarized next.A four-rater design is used to estimate the reliability of each item in the rubric, the reliability of the scores for each dimension, and the reliability of scores produced for the entire rubric (Table 4).

Table 4

Generalizability Coefficients Based on a 4 Rater D-Study For the Open for Business Task

		G-Coefficients	
Dimension	Item	Item Level	Composite
Math	**1a**	.93	
	1b	.95	
	1c	.97	
	1d	.94	
	1e	.84	
	1f	.00	.97
Math Reasoning	**2a**	.88	
	2b	.53	
	2c	.72	
	2d	.80	
	2e	.96	
	2f	.74	
	2g	.84	.93
Approach	**3**	.59	NA
Communication	**4a**	.89	
	4b	.91	
	4c	.70	
	4d	.97	
	4e	.91	.96
Entire Rubric			.97

From Table 4 we see that the reliability index for item 1f cannot be calculated because of a lack of student variation, items 2b and 3 are the least reliable, and scores based on the average of four raters on the rest of the items are highly reliable (G-coefficients range between .70 and .97).

The dimensional analysis (the composite column of Table 4) shows that with a four-rater design, scores that represent the sum of the average scores from the four raters for the relevant items have estimated reliabilities above .93. Finally, scores produced for each student that represent the sum of all the averages for the four raters for each item are estimated to have a reliability of .97.

Next, using a series of D-studies we consider how changes in the number of raters affects the reliability of the dimensional and entire rubric scores. Figure 2 shows the expected reliability as a function of the number of raters.

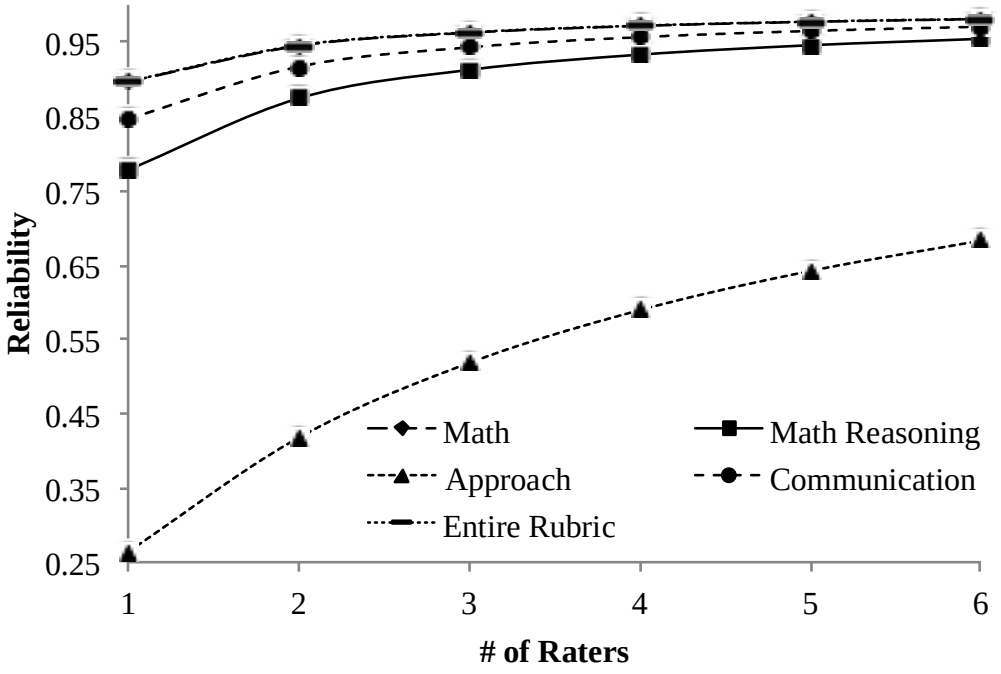


Figure 2. Estimated Reliability of Scores Produced from the Open for Business Rubric as a Function of the Number of Raters

Based on the curves in Figure 2, the Math dimension and the Entire Rubric score (the sum of scores) are consistently the most reliable. To achieve a reliability of at least .80 on scores representing the sum of the averages for all items on the rubric (the Entire Rubric curve), the Math dimension items, and the Communication dimension items, only a single rater would be

required. For the Math Reasoning dimension, at least two raters are required. Finally, the Approach dimension is the least reliable and not even six raters would produce a reliability index of .8. This is not surprising give that the approach dimension is only measured by a single item.

Taken together, the *Open for Business* rubric has been shown to produce reliable scores for the entire rubric (sum of scores). While the sub-dimension scores have been shown to be reliable, whether or not the sub-scores can be considered valid cannot be determined by these results. However, given the dissattenuated correlations reported in Table 3, there is some support for a single underlying dimension, meaning that sub-dimension scores will be highly correlated and reporting them separately is probably not warranted for this task.

The results for this task are similar to the results obtained from G-studies and D-studies of the other two tasks. The levels of Reliability based on the four-rater design were less robust for *Maximum Volume* and especially for *Wheelchair Ramp,* so while total scores had sufficiently high reliability coefficients (.80 or higher) based on a one or two-rater model, most of the dimension-level scores were not sufficiently reliable with only one or two raters. This was true for the *Open for Business* task as well.

Since it would be practically unfeasible to have work samples scored by four raters given the costs associated with hiring and training raters, it is important to consider the implications of how scores might be used. The results suggest that total scores are sufficiently reliable (using one or two raters) to report total scores, while the dimension level (and sometimes item-level) scores are not sufficiently reliable to warrant any summative use of those scores. Even from a formative standpoint - with the goal of providing accurate feedback to students and teachers on specific strengths and weaknesses within a work sample - some of the dimension and item-level

scores may not be sufficiently reliable to support their valid use to inform learning and teaching efforts.

English Language Arts (ELA) Tasks.A special calibration session was conducted during the Fall 2010 for teachers who had been pre-selected from among previously trained and calibrated scorers (in Spring 2010) on two tasks: *Americans Dreaming* and *Constructing the Self*. Four teachers per task were convened by a lead scorer to re-calibrate and refresh their familiarity with the scoring rubric and process. They spent approximately 1.5 hours reading a new sample, scoring the sample independently, discussing their scores, and coming to consensus on the scores. These discussions were led by a lead scorer in each of two task-specific groups. Following this re-calibration session, teachers were sent home with a packet of 16 work samples to score on their own. The 16 samples were randomly selected from across the class sets submitted by teachers, with the goal of representing every teacher who submitted samples.[7]They were given one month to complete the scoring and to send in their scores. Each task contains the same six dimensional items (Analysis, Perspective, Power of Language, Structure, Reflection, and Overall[8]). Unlike the Mathematics tasks, there are no sub-dimensional items in the ELA tasks. Hereafter, the ELA dimensions are called "items". The G-study and D-study results for one of the ELA tasks *Constructing the Self* are reported in this paper.

Constructing the Self - English language arts task. This task is scored using the same genre-specific rubric as the *Americans Dreaming* task and contains the same six items. Table 5 shows the absolute estimated variance contributions from each factor for each item as determined by the G-Study. Table 6 shows the same information but as proportions.

[7] It did not make sense to have the teachers score onsite, as it would have taken approximately another full day to complete the scoring of all 16 samples - and teachers were reluctant to spend that much time out of the classroom, having already spent two days in professional development.
[8] The Overall score is a holistic rating based on the five other dimensional scores.

The absolute variances reported in Table 5 show that students' scores vary on all items of the *Constructing the Self* task with absolute variance contributions ranging between .28 and .76. The rater row of Table 5 suggests that unlike the previous ELA task, items do not appear to be susceptible to a rater effect—raters are equally stringent. Finally, all items produce error variance as indicated by the non-zero absolute contributions in the error row of Table 5. Once again, the items are susceptible to either a rater by student interaction or other sources of error not captured in the measurement design or both.

Table 5

Source Table for student• x rater• Multivariate Item Design Based on Four Raters for the Constructing the Self Task

	Item					
Source	Overall	Analysis	Perspective	Power of Language	Structure	Reflection
Student	.54	.44	.38	.61	.76	.28
Rater	.02	.06	.06	.04	.00	.02
Error	.14	.13	.20	.15	.12	.29

Table 6 shows the same results but as proportions. There appears to be some variation across the items in terms of the proportional contributions from the student factor. Further, this task has produced a relatively small proportional contribution from the error factor. These results suggest these items, as a group, are producing some meaningful variation. Further, these results suggest 1) all items produce measureable variation in students' scores, and 2) scores from all items contain some undocumented error. We now turn to results that speak to how the item scores relate to one another.

Table 6

Source Table for student• x rater• Multivariate Item Design Based on Four Raters for the Constructing the Self Task as Proportions

	Item					
Source	Overall	Analysis	Perspective	Power of Language	Structure	Reflection
Student	.77	.70	.59	.75	.86	.47
Rater	.03	.10	.10	.06	.00	.03
Error	.20	.20	.31	.19	.14	.50

Table 7 shows the disattenuated correlations for the *Constructing the Self* items. Based on the patterns of correlations, the items appear to measuring the same underlying constructs. That is, all items correlate highly with one another. The indexes of reliability from a series of D-Studies are summarized next.

Table 7

Disattenuated Item Correlations (Diagonal Elements are Variances) - Constructing the Self task

	Overall	Analysis	Perspective	Power of Language	Structure	Reflection
Overall	.54	1.00	.93	1.00	1.00	.87
Analysis		.44	.99	1.00	.99	.96
Perspective			.38	.95	.93	.65
Power of Language				.61	.98	.94
Structure					.76	.88
Reflection						.28

A four-rater design is used to estimate the reliability of each item in the rubric and the reliability of scores produced for the entire rubric (Table 8).

Table 8

Generalizability Coefficients Based on a 4 Rater D-Study For the Constructing the Self Task

Item	G-coefficient
Analysis	.93
Perspective	.88
Power of Language	.94
Structure	.96
Reflection	.79
Over All	.94
Entire Rubric	.97

From Table 8 we see that, once again, all of the items produce quite strong reliability indexes when rated by 4 raters except for the Reflection item (reliability index = .79). Finally, scores produced for each student that represent the sum of all the averages for the four raters for each item are estimated to have a reliability of .97.

Next, using a series of D-studies we consider how changes in the number of raters affects the reliability of the item and entire rubric scores. Figure 3 shows the expected reliability as a function of the number of raters.

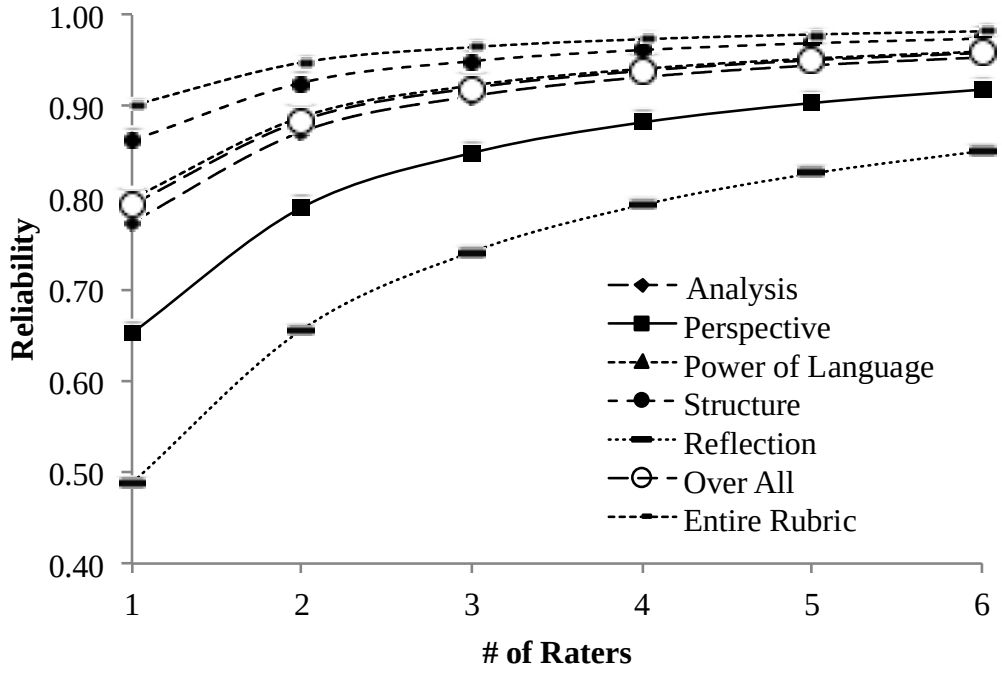


Figure 3. Estimated Reliability of Scores Produced from the Constructing the Self Rubric as a Function of the Number of Raters

Based on the curves in Figure 3, the Entire Rubric score (the sum of scores) is consistently the most reliable. To achieve a reliability of at least .80 on scores representing the sum of the averages for all items on the rubric (the Entire Rubric curve), the Power of Language scores, the Structure scores, and the Overall scores, only a single rater would be required. The Analysis, Perspective, and Reflection items require 2, 3, and 4 raters respectively to achieve a reliability of .80. The Reflection item is consistently the least reliable.

Taken together, *the Constructing the Self* rubric has been shown to produce reliable scores for the entire rubric and a few of the individual items. Similarly, some of the individual items have been shown to be very reliable. It should be noted, however, that the *Constructing*

the Self task is the better of the two ELA tasks based on the reviewed metrics of reliability.[9] For both tasks, there were high correlations between items. This might suggest that there is redundancy in the items (that they are measuring a unitary construct). However, it might also suggest that the score ranges are too constricted to provide sufficient differentiation in performance to allow for a greater range and variability in scores. Descriptive statistics for the ELA tasks show that most students receive a score of "2", with a much smaller percentage of students performing at the "1" and "3" levels. These findings might suggest a need for a greater number of levels to differentiate more subtle differences in performance across dimensions, or it may suggest a need to more clearly differentiate the constructs measured across the dimensions and to improve the scorer training protocol to support raters' ability to differentiate across the dimensions. Again, similar to the mathematics tasks, the total scores appear to have sufficient levels of reliability(with one or two raters), while more than two raters would be required to achieve a reliability of .80 on most of the dimensional scores.

Science Tasks. Special scoring sessions were convened to collect scores for the G-studies in Chemistry and Physics. These sessions occurredin August of 2010, prior to the beginning of the academic year. Because there had been an insufficient number of Physics task samples submitted the previous spring, the scorer training for that task had been delayed until late summer. Thus, calibration and scoring for the G-study was embedded within the regular scorer training session for the Physics group.

Six physics teachers scored the same set of 16 physics work samples, randomly selected from the class sets submitted by all teachers. The teachers completed the training and scoring for the G-study on-site in one and a half days. Four chemistry teachers were recruited from among

[9]Most of the dimension-level scores for the Americans Dreaming task did not produce sufficient levels of reliability to warrant the reporting of the dimensional scores.

those who had been trained and met calibration standards during the scorer training session held in the spring of 2010. They were convened for a re-calibration activity in August 2010 and completed scoring for the G-study on-site in one day. The four chemistry teachers scored the same set of 16 chemistry task samples that were also randomly selected from the class sets submitted by all teachers. Both the Chemistry and Physics tasks were scored using the same rubric for science inquiry. The Chemistry task (*Got Relieve It?*) was scored using seven items and the Physics task (*How Things Work*) was scored using six items (Design of the Investigation was omitted for that task). There are no sub-dimensional items for the Science tasks. G-study and D-study results for the Chemistry task (*Got Relieve it?*) are presented here.

Chemistry - Science Task. The seven items scored in the Chemistry task are Collect and Connect Content Knowledge, Design of the Investigation, Analysis of Data, Draw Conclusions, Communicate and Present Findings, Reflect on the Learning Process, and Overall. (Again, the Overall score is a holistic rating based on ratings of the dimensions.) Table 9 shows the absolute estimated variance contributions from each factor for each item as determined by the G-Study. Table 10 shows the same information but as proportions.

The absolute variances reported in Table 9 show that students' scores vary on all items of the Chemistry task with absolute variance contributions ranging between .13 and .44. The rater row of Table 9 shows that variation due to rater is low on all items. Finally, all items produce error variance as indicated by the non-zero absolute contributions in the error row of Table 9 suggesting that items are susceptible to either a rater by student interaction or other sources of error not captured in the measurement design or both.

Table 9

Source Table for student• x rater• Multivariate Item Design Based on Four Raters for the Chemistry Task

	Item						
Source	Overall	Cllct. & Connect	Design	Analysis	Conclu.	Comm.	Reflect
Student	.18	.13	.18	.25	.12	.13	.44
Rater	.01	.02	.04	.01	.00	.01	.03
Error	.09	.14	.15	.19	.11	.16	.15

Table 10 shows the same results but as proportions. The items produce similar proportional contributions from all the factors. These results suggest these items are functioning more similarly. Further, these results suggest 1) there is measureable variation in students' scores on all the items, and 2) scores from all items contain some undocumented error. We now turn to results that speak to how the item scores relate within and across dimensions.

Table 10

Source Table for student• x rater• Multivariate Item Design Based on Four Raters for the Chemistry Task as Proportions

	Item						
Source	Overall	Cllct& Connect	Design	Analysis	Concl.	Comm.	Reflect
Student	.65	.44	.48	.56	.53	.43	.71
Rater	.02	.08	.11	.02	.00	.05	.05
Error	.32	.48	.41	.42	.47	.52	.24

Table 11 shows the disattenuated correlations for the Chemistry items. Based on the patterns of correlations, the items appear to measuring the same underlying constructs. That is, all items correlate highly with one another. The indexes of reliability from a series of D-Studies are summarized next.

Table 11

Disattenuated Item Correlations (Diagonal Elements are Variances) - Chemistry Task

	Overall	Cllct. & Connect	Design	Analysis	Cnclns.	Comm.	Reflect
Overall	.18	.86	.76	.89	.99	.95	.79
Cllct. & Connect		.13	.95	.97	.81	.99	.79
Design			.18	.95	.42	.90	.88
Analysis				.25	.84	1.00	.87
Cnclns.					.12	.83	.83
Comm.						.13	.85
Reflect							.44

Once again, a four-rater design is used to estimate the reliability of each item in the rubric and the reliability of scores produced for the entire rubric (Table 12).

Table 12

Generalizability Coefficients Based on a 4 Rater D-Study - Chemistry Task

Item	G-coefficient
Collect	.78
Design	.83
Analysis	.84
Conclusions	.82
Communication	.77
Reflect	.92
Overall	.89
Entire Rubric	.95

From Table 11 we see that all of the items produce quite strong reliability indexes. The weakest items are the Communication and Collect items with a reliability index of .77 and .78, respectively. Finally, scores produced for each student that represent the sum of all the averages for the four raters for each item are estimated to have a reliability of .95.

Next, using a series of D-studies we consider how changes in the number of raters affects the reliability of the item and entire rubric scores. Figure 4 shows the expected reliability as a function of the number of raters.

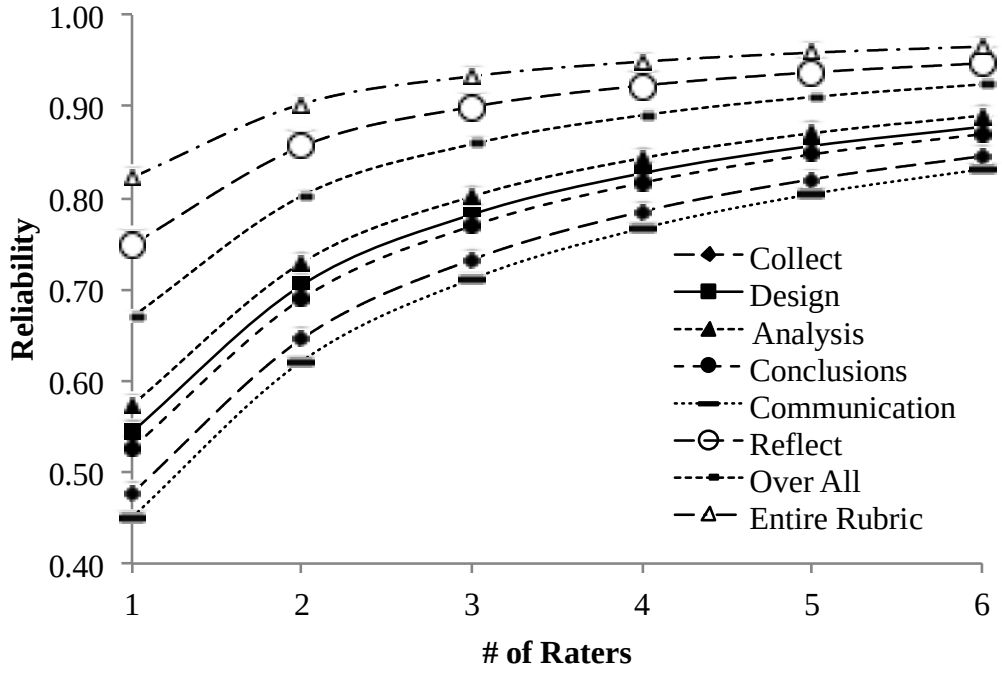


Figure 4. Estimated Reliability of Scores Produced from the Chemistry Rubric as a Function of the Number of Raters

Based on the curves in Figure 4, the Entire Rubric score is consistently the most reliable. Only a single rater is required to achieve a reliability of at least .80 on scores representing the sum of the averages for all items on the rubric (the Entire Rubric curve). In terms of the reliability of individual item scores, the Reflect, Overall, Analysis, Design, Conclusions, Collect, and Communication items require 2, 2, 3, 4, 4, 5, and 5 raters, respectively to achieve a reliability of .80.

Taken together, the Chemistry rubric has been shown to produce reliable scores for the entire rubric. However, the individual items have varying levels of reliability and some items require a large number of raters to achieve sufficient reliability. Further, as was found for the ELA rubrics, the correlations between items suggest a need to revisit the design of the rubrics to provide greater differentiation across levels of the rubrics and across dimensions.

In addition, the results of the Physics task *How Things Work*were significantly worse than those for the Chemistry task. While the error due to raters was low, there were also high proportions of unexplained error. The application of the Science Inquiry rubric (the same rubric used for the Chemistry task) to the Physics task *How Things Work* resulted in relatively unreliable scores for both the entire rubric and on individual items. In sum, the scoring rubric for Science Inquiry seems to be more robustly applied on samples from the Chemistry task. The results of these analyses are consistent with feedback provided by raters who scored the Physics task. They reported finding it difficult to apply the common scoring rubric for science inquiry to the Physics task *How Things Work*. This feedback combined with the low levels of reliability and high error variance strongly suggest a poor alignment between the rubrics and the task. Ultimately, the science team decided to discard this particular Physics task as it did not adequately measure the constructs assessed in the scoring rubrics.

Summary of Results of G-studies and D-studies

Overall, results from the generalizability and dependability studies showed that for most of the performance tasks, there were sufficiently high levels of reliability (0.8 or higher), but were particularly high for mathematics, which were scored using task-specific, analytic score scales that include more "items". Reliability at the "Entire Rubric" level was still sufficiently high for tasks in the other content areas (ELA, science) that were scored using generic four-level rubrics, but less so for a physics task that was already determined during the pilot to produce weak evidence in relation to the evaluative criteria of the scoring rubrics. However, the inter-correlation levels between scoring elements were quite high across most of the rubrics, suggesting that the individual rubric dimensions were not distinguishing adequately between the different characteristics of performance that they were intended to measure.

Discussion

The results of the G-studies and D-studies described above have implications for two aspects of validity, as described in the conceptual framework section of this paper: 1) *Scoring Models as Reflective of Task and Domain Structure; 2) Generalizability and the Boundaries of Score Meaning.* First, the structure of the scoring rubrics used to score the OPAPP performance tasks do not appear to be supported as unique domains that differentiate student performance. The results suggest, instead, that each set of rubric domains are really supporting the evaluation of one underlying construct, whether that be mathematics problem solving, textual analysis, or science inquiry. While results of our studies indicate that the scores are providing reliable differentiation in student performance at the "entire score" level (which is a composite of the individual dimension/item scores), the dimensional or item scores differ in whether they 1) provide meaningful variation or 2) reliably differentiate student performance.

Given SCALE's desire to construct scoring tools that do provide differentiation within a student's performance across different dimensions, results suggest that multiple occasions (using different and distinct tasks) for collecting and scoring students' performance would be needed to achieve this goal. For example, in addition to having students complete an integrated science inquiry task that includes all aspects of inquiry from start to finish, a series of more discrete tasks that evaluates students on more aspects of the inquiry process could also be administered (e.g., a task in which students use content knowledge to design an investigation but do not conduct the investigation; a task in which students analyze a set of data and draw conclusions; a task in which students are asked to evaluate the analysis or conclusions drawn by someone else based on a set of data and identify potential sources of error).

Second, the generalizability results have implications for the intended purposes and uses of the scores. The high levels of reliability of the scores at the "Entire Rubric" level suggests that the learning tasks could be used for summative purposes where the goal is to be able to reliably rank order students based on their performance. Alternatively, the lower levels of reliability at the dimension/item level suggest that these scores should not be used for summative purposes. In fact, even for formative uses - to support more fine-grained decisions about ways to improve instruction and learning - higher levels of reliability on the dimensional scores would be required.

In addition, there are several implications for performance assessment design that can be gleaned from these analyses. In particular, there are lessons about the design of the scoring rubrics and the design of tasks that are aligned with the rubric constructs. With regard to the design of scoring rubrics:

1. A greater number of items within dimensions, as with the mathematics task scoring rubrics, is likely to support greater levels of scoring reliability. Since it is unlikely that a performance based assessment system is likely to have sufficient resources to support more than a small percentage of double-scoring or scoring a single sample more than twice, a design that supports greater reliability would be more efficient. This would suggest that the ELA and science common scoring rubrics should be broken down into more discrete items within dimensions, as they are in the mathematics scoring rubrics. Currently, each ELA and science scoring dimension includes multiple indicators and evidence collected across these indicators is used to arrive at a holistic score for the dimension, using professional judgment. Breaking down these dimensions along the lines of these different indicators may be a logical way to add items

within dimensions and would simplify the judgments used by raters to arrive at an individual item score, further supporting reliability.

2. The grouping of items within dimensions should be based on empirical evidence of their correlation with other items as well as theoretical assumptions about their connection to particular scoring dimensions. It is apparent that for the mathematics tasks, the correlations between individual items did not support their current grouping scheme. In fact, the items designed for the three mathematics tasks were not constructed to specifically assess the constructs assessed by each dimension. Rather, the task developers constructed the scoring items, and then "back-mapped" the items to the larger scoring dimensions. If the composite dimension scores are to have meaning, empirical data about how the items work together should be used to determine the composition of dimensions using individual items.

3. Rubric dimensions should be designed to more distinctly differentiate between aspects of students' skills and different levels of performance. High correlations in scores between dimensions for both the ELA and science tasks suggest that the rubrics and scoring protocols are not producing sufficient range or variation in scores across the dimensions and across students. Descriptive statistics show that the vast majority of students end up being scored at the "2" level, and qualitative examination of work samples scored at that level show that there are differences among performances scored at that level. In addition, raters often struggle to assign scores, wanting to give "1+", "2-", "2+" or "3-" scores. This suggests a need for a score scale that has more score levels -- perhaps 5-7 score levels. These data also suggest a need for improvements in the rubric descriptors to improve clarity about the different constructs measured in each dimension, and/or improvements in the scorer training protocols to support raters' ability to score each dimension more distinctly and to avoid the "halo effect".

4. Need for clearer differences across score levels. From the G-study and D-study results for the science and ELA learning tasks, we find that many of the dimensional scores do not have sufficient reliability to stand on their own (although the Entire Rubric Score does have a sufficient level or reliability with one rater, in most cases). These results suggest that individual dimensional scores cannot yet be used to track progress in student performance on these dimensions over time and across tasks, although their overall performance based on the Entire Rubric Score can be used in this way. However, given that these are "learning" tasks, it is important that the dimensional scores have meaning and can be used for formative purposes to provide feedback to teachers and students about their relative strengths and weaknesses. We suggest that the wording of the science and ELA rubrics (in addition to breaking down the dimensions into more discrete items) be improved to more clearly distinguish between different levels of performance.

5. Task-specific vs. common scoring rubrics. While the task-specific rubrics used with the mathematics tasks were easier to score and were probably more reliably scored due to the less complex set of judgments made by teachers, there are also drawbacks - the lack of alignment of the task-specific items to common dimensions that allow for comparability of student performance across tasks. On the other hand, the common scoring rubrics used for the ELA and science tasks were more difficult to score and less reliable, requiring more complex judgments, but would allow for comparison of student performance across tasks with sufficient reliability. A hybrid approach in which raters score a common set of constructs and dimensions, but includes task-specific "look fors" to supplement the information needed to score each common scoring dimension may provide the best features of both types of rubrics. Another approach is to design task-specific rubrics in which some "items" are specifically aligned to common scoring

dimensions (e.g., mathematical practices in the Common Core State Standards). In this case, some of the score data may be used to track progress over time on these common scoring dimensions, while other parts of the score data would be unique to the particular task. A last approach is to develop task-specific rubrics and to convert the raw rubric scores into standardized scores. While these standardized scores are less transparent to users (teachers, students) and do not provide analytic information/feedback to teachers or students about how they might improve their performance along specific dimensions, they allow for comparability of scores across tasks (by rank ordering students in comparison with the norm). This would be useful primarily for research purposes rather than for the purpose of reporting/feedback to schools, teachers, and parents/students.

The results of the G-studies also have implications for the design of tasks:

The tasks should be designed using task shells to ensure alignment with the constructs evaluated by the rubrics. The Physics task *How Things Work* was clearly not aligned well to the scoring criteria resulting in large error variances and a lack of student-attributable variance in scores. A task shell or other task specifications should be used to design any tasks that are to be scored using the common scoring rubric. This would also support the comparability of tasks. The mathematics tasks were clearly difficult to compare and were likely not "equivalent" tasks due to differences in the nature and content of the tasks and other task features unique to each task. A task shell that is used for specific genres of tasks would support task comparability within task genres.

Reliability of Assessment Task Scores

A similar set of G-studes and D-Studies were conducted on the score data that were available from distributed scoring sessions following scoring training of teachers and project

staff. Because the assessment tasks were designed to be completed in an on-demand, standardized format, the rubrics that were designed to score them were task-specific (in mathematics and science tasks). However, there were also differences in the design of the rubrics.

In the mathematics rubrics, the majority of the scoring rubrics evaluated students' abilities to produce correct responses, with additional credit being awarded in a very few cases for elaborated and correct explanations in one or two cases. In contrast, items on all of the science rubrics were scored on a greater range of scores (0-4), with a requirement that students produce both correct responses AND evidence of reasoning/explanations to even achieve a level 1 score. So even though the rubrics were task specific, on both tasks, the scoring scheme was meant to capture variation in students' ability to explain their reasoning and demonstrate conceptual understanding.

One common design aspect between the math and science tasks was that the rubric items were not organized into common scoring dimensions - rather each scoring section simply evaluates a different part of the task. Therefore, the format of the scores do not lend themselves to dimensional analysis or comparison of performance across tasks.

In the following sections, we present results of the G-studies and D-studies for two tasks - the chemistry assessment task *Neutralizing Relieve It* (designed to be aligned with the chemistry learning task *Got Relieve It?*) and the mathematics task - *The Phone Plan* (designed to be aligned with the mathematics tasks *Open for Business).* Results for the other two assessment tasks piloted in the spring of 2011 (the physics task *Energy Efficient Rubberband Vehicles* and the mathematics task *Tipping the Tank*) are also noted and compared.

Chemistry - *Neutralizing Relieve It.* The Chemistry rubric contains 10 items. Data from two sub-samples (n = 10, n = 25) of student work that were double scored by raters 50 & 78, and 78 & 92, respectively, are used to conduct the analyses.[10] Table 13 shows the G-study variance estimates of the different tractable factors based on the measurement design for the two individual samples and the weighted average across the two samples for the Chemistry task.

Table 13

Chemistry Task Absolute and Proportional Estimates of Variance Contributions for the Different Factors in the Measurement Design

	Absolute Variance		
Source	Raters 50 & 78	Raters 78 & 92	Weighted Average
Student	.13	.26	.19
Item	.00	.03	.01
Rater	.00	.00	.00
student x item	.12	.23	.17
student x rater	.00	.01	.01
item x rater	.12	.01	.07
Error	.20	.14	.17
	Proportion of Total Variance		
Student	.23	.38	.31
Item	.00	.04	.02
Rater	.00	.00	.00
student x item	.21	.33	.28
student x rater	.00	.02	.01
item x rater	.22	.02	.11
Error	.34	.21	.27

[10] Rater IDs are the last two digits of the raters longer ID set up in the original data file

Based on these results, 31% of the observed total score variance can be attributed to actual differences in student performance while 28%, 1%, and 27% can be attributed to student by item, student by rater, and error, respectively (lower right cells in Table 13). While not shown here, the Physics task resultsproduce a very similar pattern . That is, there is very little in the way of error due to rater inconsistency and as a result, increasing the total number of raters will not improve the reliability. *Further, a large proportion of the observed score variance is due to a student by item interaction and as a result, increasing the number of items will produce a more reliable total score.* Figure 5 illustrates these points graphically by plotting the reliability of the total score as a function of the number raters and the number of items based on the weighted average absolute variance estimates from Table 13.

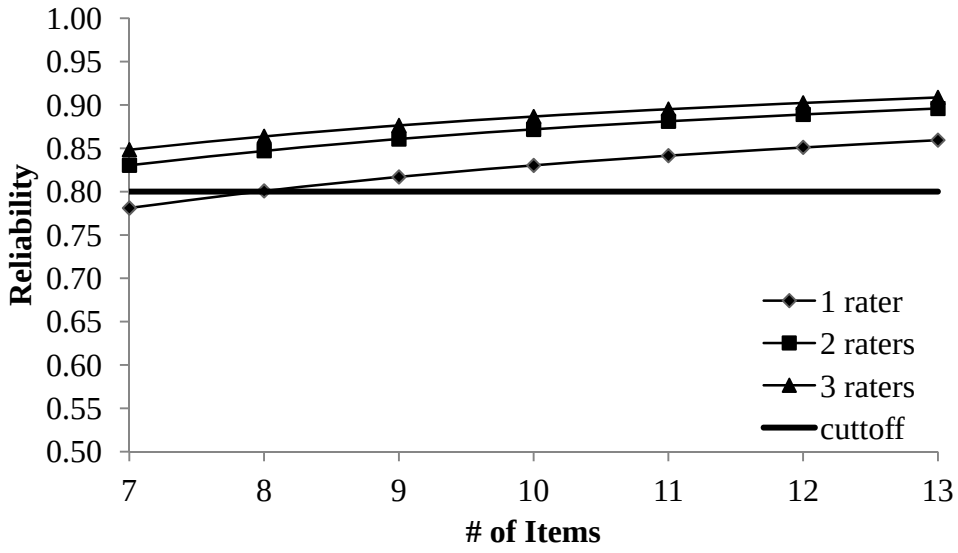


Figure 5

Reliability of the total score on the Chemistry task as a function of the number of raters and items

In its current state (10 items), total scores produced from the Chemistry task based on ratings from at least 2 trained raters are estimated to have a reliability of .89. Compared to the Physics task, the presence of more student variation and slightly less student by item variation

results in a generally more reliable total score for the Chemistry task. In fact, based on these results, options such as employing fewer raters or using fewer items can be considered because they will not likely result in much lower reliability.

Mathematics - *The Phone Plan*. The Phone Plan rubric contained 10 items. Data from a single sample (n = 17) of student work that were double scored by raters ODE1 and ODE2 are used to conduct the analyses. Table 14 shows the G-study variance estimates of the different tractable factors based on the measurement design for the two individual samples and the weighted average across the two samples for the Phone Plan task.

Table 14

The Phone Plan Task Absolute and Proportional Estimates of Variance Contributions for the Different Factors in the Measurement Design

	Absolute Variance
Source	Raters ODE1 & ODE2
Student	.03
Item	.12
Rater	.00
student x item	.08
student x rater	.00
item x rater	.02
Error	.04
	Proportion of Total
Student	.11
Item	.41
Rater	.01
student x item	.29
student x rater	.00
item x rater	.05
Error	.14

Findings show that only 11% of the observed total score variance can be attributed to actual differences in student performance while 29%, 0%, and 14% can be attributed to student by item, student by rater, and error, respectively (lower right cells in Table 14). While score variance attributable to student performance is not as low as on the *Tipping the Tank* task, *these findings suggest that the total score is not picking up much meaningful variation in student scores.* Once again, increasing the number of items is more important than increasing the number

of raters in order to produce a more reliable total score. Figure 6 illustrates the findings graphically by plotting the reliability of the total score as a function of the number raters and the number of items based on the absolute variance estimates from Table 14.

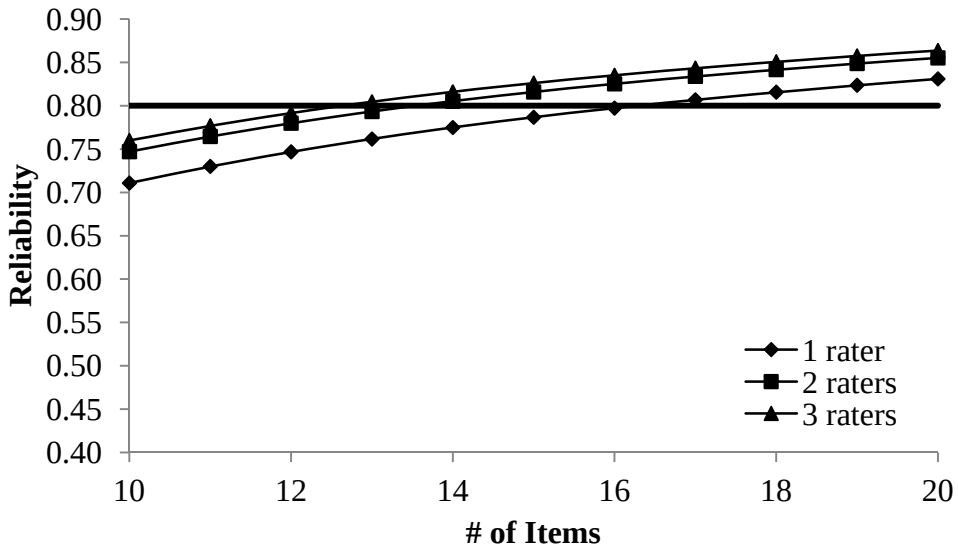


Figure 6

Reliability of the total score on the Phone Plan task as a function of the number of raters and items

In its current state (10 items), total scores produced from *The Phone Plan* task based on ratings from at least 2 trained raters are estimated to have a reliability of .75. *The Phone Plan* task is the more reliable of the two math tasks but reliability can be improved by adding more items or improving the quality of the items in terms of their ability to produce more meaningful variation in scores.

Implications for Task and Rubric Design. Results from the G-study analyses indicate that rater reliability is not so problematic for the OPAPP assessment tasks given the design of the point-based scoring rubrics. There was a very low percentage of error attributable to raters and the rater by student interaction. However, the results of the G-studies underscore the lack of variation in student performance produced by some of the scoring items, especially in the

mathematics tasks. This could be a problem with the task design itself. That is, the tasks are designed so that almost everyone can succeed on the first part of the tasks which rely on fairly basic mathematical skills and thatprovide entry into the problem. Alternatively, almost no one is able to arrive at correct solutions on the more complex parts of tasks. This suggests that a task that includes more intermediate level items, rather than very easy or very difficult items might produce more variation in scores.

Another source of the problem may be the dichotomous scoring rubric. Because on almost all items, the only possible score values are 0 or 1, with no partial credit given for attempting the problem, getting an answer partially correct, or explaining one's reasoning, there is less opportunity for students to vary in their performance on the items. This is less problematic in the science assessment task rubrics, which also has a point-based scoring scheme while allowing for partial credit to be awarded on a score scale between 0 and 3 on most items. Simply increasing the number of dichotomously scored items in mathematics may prove less valuable for picking up variation in performance than increasing the range of possible scores achievable by students on the individual items and moderating the ease or difficulty of the task prompts.

Implications for Summative Use of Assessment Task scores. While there is evidence that rater reliability is not a problem and that raters are able to score the assessment task rubrics with a high level of consistency, evidence from the G-studies indicate that not all of the tasks as currently designed produce a sufficient level of variation in scores based on the design of the tasks and rubrics themselves. The strongest task and rubric design was found in the Chemistry assessment task *Neutralizaing Relieve It,* which had a high level of reliability given the existing number of items even with a single rater. On the physics assessment task *Energy Efficient*

Rubberband Vehicles, adding a few more items in the scoring rubrics may be sufficient to improve the reliability of the instrument (as well as ensuring that all samples are double-scored). With minor tweaks, the science tasks appear to be robust enough in technical quality to be used as a summative measure to evaluate student performance.

However, in the mathematics assessment tasks that were piloted in spring 2011, while the raters were highly reliable (based on a 0-1 scoring scheme for almost 100% of items), the tasks and scoring rubrics themselves were not picking up much variation in student performance. There was such little variation in student performance that even with the addition of many more items of the same type, the level of reliability would still be insufficiently high enough to warrant making high-stakes summative claims about students' performance. These results indicate that the mathematics tasks and/or the scoring rubrics used to evaluate them should be revised before considering their use for high-stakes summative purposes.

A closer look at the score patterns on the mathematics assessment tasks indicates that total score distributions are skewed toward the lower end of the scale, with average performance being less than 50% of the maximum point values. Item distributions[11] indicate that on both tasks, students perform well on the initial items of the task, with high percentages of students getting correct answers on the first few items. In *The Phone Plan*, most students are able to provide correct answers and correct explanations of their answers (Items 1-3A). However, beginning with Items 3B, 4 and 5, which require that students use data to produce equations or functions, most students are unable to provide correct equations or do not attempt to answer. Since Items 6 and 7 depend on correct equations in Items 4 and 5, this means that they are unable to succeed on these items if they give incorrect answers on Items 4 and 5. According to the

[11] Item distributions are available from the authors and in the final technical report for the Ohio Performance Assessment Pilot Project, available from the Stanford Center for Assessment, Learning, & Equity.

mathematics task designer, the mathematics assessment tasks were deliberately designed in this way, using a "ramping" strategy that allows all students to initially access the tasks by successfully completing "entry level" questions, and gradually ramping up the difficulty of the items within a task.

The low average levels of performance in general, and the almost universal poor performance on the latter items in the task raises implications about task design and about the validity of the items on which there is little student variation, either due to constrained score ranges (almost all are scored 0/1), or because of opportunity-to-learn issues (few students having had prior experiences with or instruction on the kinds of higher level skills expected in the more complex sections of the assessment tasks). The mathematics task designer clarifies that the assessment tasks were not designed to produce much variation within individual items, but that the task as a whole should be considered the unit of analysis, and that composite scores are likely to produce more variation. Thus, the psychometric framework in which sampling error due to items is evaluated may not apply well to the performance task score because these "items" may simply be different facets of performance on a single task or "item". On the other hand, the latter items of the mathematics assessment tasks, on which most students get "0" scores do seem to provide some diagnostic information about the limits of individual students' conceptual understandings and skills and what they can and cannot do. And it is not surprising that in the first pilot of these tasks within Ohio that students unaccustomed to this task format and its more challenging cognitive demands would do poorly on the more demanding parts of the tasks. It may be that over time, as teachers and students have more experience with this kind of assessment task format, that average performance on these tasks will improve leading to greater variation in scores. However, until greater variation is achieved, as designed, the assessment

tasks and scoring rubrics in mathematics lack sufficient reliability to be used for high-stakes purposes.

TheRelationship between Learning Task Scores and Assessment Task Scores. Another hypothesis that is exploredis whether the learning tasks produce score profiles that are related to the score profiles obtained from the assessment tasks for individual students, even though the scoring scales used are different (point-based scoring vs. generic rubrics) and the type of performance assessment format and context are different (timed and on-demand, completed individually vs. embedded in an instructional unit and completed with support of peers and teachers). As noted earlier, the Learning and Assessment Task Dyads were designed so that both types of tasks should measure the same constructs.

Unfortunately, three out of the four learning tasks that were part of these Task Dyads were "spring" tasks, and no score data were collected for those learning tasks because they were optional in that semester and no scorer training was provided for the spring learning tasks in 2011. However, the physics learning task that was piloted in the fall of 2010 (*Energy Efficient Vehicles*) was aligned with the physics assessment task (*Energy Efficient Rubberband Vehicles*), and scores were collected from teachers who had used that learning task following scorer training in January 2011. (No samples from this task were double scored, so we do not have any estimates of reliability for those learning task scores.) When the datasets for the physics learning task (N=106) and the physics assessment task (N=203) were merged using the student ID, 49 matches resulted. Despite the small sample size, the Pearson bi-variate correlation between the Total Scores for both tasks was moderately high (.60) and statistically significant (at the .001 level).

Though these results are limited by small sample size on only one of the task dyads (out of four task dyads piloted in the spring of 2011), they suggest that the two assessment formats have a fairly high degree ofoverlap, but they are also measuring something different. It is not surprising that this would be true, given the difference in formats. Generally speaking, students are likely to perform better on classroom assignments with peer and teacher assistance, access to their notes, textbooks, and other resources, and opportunities for revision, than on a timed, on-demand test with no access to outside resources.

Concurrent Validity Studies

One of the premises for integrating performance based assessments into a multiple-measures assessment system is that performance based assessments are able to measure college and career readiness skills more directly than standardized tests, and therefore, the score data they produce should tell us something about student performance *distinct from their performance on standardized measures.* Another tenet is that performance based assessments may provide different kinds of opportunities to demonstrate learning for students who may not typically be successful in demonstrating their learning on standardized tests. The analyses that follow examine the relationships among the performance assessment scores obtained from the Spring 2010 pilot of the learning tasks, Ohio Graduation Test (OGT) scores, demographic information about students, high school GPAs, and ACT scores.

Overview of Analyses. The research questions guiding the analyses include the following:

1) What is the relationship between students' performance assessment scores and other measures of their achievement? (e.g., GPAs, Ohio Graduation Test scores, ACT scores)

2) To what extent is this relationship mediated by content area, task, and student/school demographic characteristics?

The hypotheses or predicted relationships based on the underlying premise that performance based assessments measure student abilities that are distinct from those measured by standardized tests and regular classroom grades is that there will be moderate (.4 – .6) levels of correlation between performance task scores and other measures of achievement. Correlations higher than that would suggest that the performance tasks are measuring a similar set of

constructs as those measured by the other measures, while correlations lower would suggest a disconnect from other estimates of student achievement.

While the original intent was to use multi-level regression modeling to control for the influence of school and teacher level factors (as fixed effects) on these relationships, there were insufficient samples sizes within tasks for each performance task and multiple linear regression models were employed instead. (Aggregation of score data across tasks even within content fields did not make sense given differences in the tasks and differences in the score scales by task in mathematics.) In addition, while the intent was to analyze the extent to which individual pupil's race/ethnicity, language background, disabilities, and free and reduced lunch status mediated the relationships, low sample sizes again limited our ability to include some of these factors in the regression models. To supplement the lack of adequate sample sizes, school-level data from the Common Core Data on the demographic characteristics of students were used as controls in some cases. To adequately address the question of the role of school/classroom level factors or students' individual demographic backgrounds in mediating the relationship between performance assessment scores and other achievement measures, a much larger sample size *within task*sis required. We recommend that for the next pilot phase, sites with much more diverse student populations should be included in the project and that the choice of tasks to be piloted be controlled to a greater extent to improve the sample sizes within tasks.

Data and Descriptive Statistics.Score data from the Spring 2010 pilot of the Learning Tasks in English language arts, mathematics, and science were used as the basis for these analyses. When multiple scores were available for a single work sample, two strategies were used to select the set of scores: 1) If a sample was scored more than twice, the "consensus score" was used - meaning that the most common score assigned a given rubric item was selected; 2) If

a sample was scored twice, one set of scores was selected at random. (Given the high levels of inter-rater reliability that we observed in the G-studies, this seems to be a defensible choice. The alternative, which is to average the scores, would change the structure of the score scales making interpretation of the data more complex.) The sample consists of 1,363 student observations from 25 schools and 76 classrooms. These students were each given one task (some students are observed twice, as they were in more than one classroom). There are ten different tasks across three different subject areas: English language arts, mathematics and science. The sample is described first followed by the outcome variables and the most important control variables.

The Sample. The sample sizes of different groups of interest are displayed in Table15, grouped by the learning task completed. Migrant status is not displayed as no students are classified as migrants. Most students in this sample began high school in 2007 (500 students -- 37%) or 2008 (763 students -- 57%).The majority of students in the dataset are white (1,161 students -- 87%). The largest minority group is African American students (104 students -- 7.75%). Since the sample sizes are too small to analyze racial or ethnic groups separately, we created an indicator variable equal to one if a student is either African American or Hispanic, as these two groups had significantly lower achievement scores than the other groups. All other groups (white, Asian, Native American or multi-racial) will serve as a reference category in further analysis.

Ten students (.75% of the full sample) are immigrants. English is the native language for 2 of the 10 immigrant students. A slightly larger number (42 students – 3%) have a native language other than English. The majority of students (1317) speak English at home, including 17 of the students for whom English is not their Native language. Of students who speak a language other than English at home, the largest group speaks Arabic (7 students) followed by

Spanish (4 students). For our purposes, we only classify students as Limited English Proficient if they are classified as such by their school, which accounts for less than 2% of the sample (24 students).

Students are classified as Special Education if they have a disability or are taken outside of the class for Special Education services. The majority of students in the dataset are not disabled (1,257 students – 94%). Of those with a disability, the largest is a category called "specific learning disability" (56 students – 4%). Sixteen students (1%) are listed as having a "minor" health impairment. Five or fewer students are listed in the other categories: emotional disturbance; visual, speech or language impairment; autism and cognitive disability. Only 19 students (1%) have a 504 plan; 18 of these are listed as not having a learning disability and are therefore not classified as Special Education. Overall, 87 students in the sample are classified as Special Education (6%).

About 21% of the samples (308 students in all) were classified in the demographic information provided by ODE as "economically disadvantaged." This is a designation defined by ODE as "the student who meets the definition of economic and/or academic disadvantagement." It is not clear what criteria were used to define economic/academic disadvantagement. The number of students meeting this definition for each task can be found under the "Low SES" category in Table 15 below.

Roughly a quarter of students are classified as "gifted" in one of five categories: thinking ability, math, reading, science or "super cognitive." However, this gifted label is not equally distributed among students; only 7% of African American or Hispanic students are classified as gifted, as opposed to 29% of other students.

Table 15

Sample Sizes by Assessment Task, Demographic Groups and School Units

Assessment	Total Sample	By Gender		By Race/ Ethnicity						Other Demographics Groups			How Grouped	
		Male	Female	White	Asian	Hispanic	African American	Native American	Multi-racial/ ethnic	Low SES	Limited English Proficient	Special Education	Schools	Class-rooms
English Language Arts														
Americans Dreaming	302	153	146	260	2	2	25	0	10	67	3	35	8	16
Constructing the Self	105	46	59	91	1	4	8	0	1	23	2	8	4	6
Employing the Personal	61	29	30	51	0	5	3	0	0	32	0	3	2	3
Hero's Journey	137	65	70	120	3	2	9	0	1	30	11	24	4	7
Science														
Chemistry: Got Relieve IT	95	42	51	83	0	1	5	1	3	14	0	1	6	9
Physics: How Things Work	27	14	9	18	1	0	3	0	1	4	0	1	8	8
Biology: Medical Mysteries	53	18	32	36	4	2	5	1	2	7	0	0	3	3
Mathematics														
Maximum Volume	295	144	149	265	1	5	18	0	4	69	1	10	9	14
Open for Business	221	97	121	177	3	6	26	0	6	57	6	4	7	8
Wheelchair Access	67	27	40	60	0	3	2	0	2	17	1	1	3	4

Outcome variables. The outcome variable of interest is thestudent's score on one of ten learning tasks in English language arts, science or math. Each task consists of several subscores, which were averaged to obtain a mean score. This mean score is used as the outcome variable. Mathematics tasks are assigned points using task-specific point scoring systems which allow for partial credit, summed to obtain a total score. Since the maximum possible score varies for each math task,the percentage of the total score points is used as the outcome variable to allow for some comparability across math tasks. The four ELA and three science tasks were scored on a four-point scale:

Level 1 = Little to no evidence of proficiency

Level 2 = Developing proficiency

Level 3 = Proficient / college ready

Level 4 = Exceptionally advanced

For ELA and science tasks, the overall mean scores can therefore be interpreted as a student's mean proficiency across the subscores.

Average scores for each task (and their standard deviations) are listed in Table 16. Overall, most of the ELA and science scores hover above 2 (developing proficiency to proficient). Math scores tend to hover around 50% points earned. Standard deviations indicate that there are scores throughout the available range. On average and in most tasks, women tend to score higher than men, African American and Hispanic students tend to score lower than other students, and low SES and Special Education students tend to score below the overall mean in most tasks.

Other Academic Indicators. The two main indicators of a student's academic performance prior to the assessment tasks are a student's cumulative grade point average and

scores on the Ohio Graduation Test. Scores for each of these indicators are broken up by demographics in Table 17.

The Ohio Graduation Test (OGT) tests five subjects: reading, writing, science, math and citizenship. The OGT is originally given in the 10th grade, though the scores in our data are from whenever students most recently took the exam. We therefore use 10th grade scores for 10th graders and students who passed the exam the first time. For students in the 11th or 12th grade we may have their 10th grade score or, if they did not pass the OGT in the 10th grade we have their most recent score. We have no indicator of when the exam was taken. Scores have been standardized to a mean of 0 and standard deviation of 1 within each test in the sample. Because students are given multiple opportunities to take and pass the OGT before they graduate, the range of scores is constrained, and likely does not reflect the full range of performance that would be obtained for an assessment that is administered once.

GPA is a cumulative measure from roughly 0 (F) though 4 (A). We are unclear as to whether or not GPAs are weighted. The top GPA at one school is in excess of 4.3 (which is generally given for an A+), despite not indicating that they weight their GPAs on the transcripts. Transcript data for 157 students (12% of the sample) are missing.

On average, female students have slightly higher OGT scores in reading and writing, as well as higher GPAs. Male students tend to score higher on the math, science and citizenship OGTs. African American and Hispanic students tend to score about a half a standard deviation lower than other students on each of the OGTs. African American and Hispanic students also tend to have lower GPAs than other students, though this gap is slightly smaller than the gap in OGT test scores.

Table 16

Means and Standard Deviations for Assessment Scores

Task	Overall	Male	Female	White, Asian, Nat. Am. Or Multi	Black or Hispanic	Low SES	Special Education
Americans Dreaming	2.2	2.1	2.4	2.3	1.6	1.8	1.8
	(0.7)	(0.7)	(0.6)	(0.7)	(0.4)	(0.7)	(0.6)
Constructing the Self	2.4	2.3	2.5	2.4	1.9	2.1	1.9
	(0.6)	(0.7)	(0.5)	(0.6)	(0.3)	(0.6)	(0.3)
Employing the Personal	2.0	1.8	2.2	2.0	1.9	2.0	2.3
	(0.6)	(0.5)	(0.5)	(0.6)	(0.4)	(0.5)	(0.5)
Hero's Journey	2.2	1.9	2.4	2.2	1.8	2.0	1.5
	(0.7)	(0.7)	(0.7)	(0.7)	(0.6)	(0.7)	(0.6)
Chemistry: Got Relieve IT	2.0	1.9	2.1	2.0	1.8	2.1	1.8
	(0.7)	(0.5)	(0.7)	(0.7)	(0.4)	(0.7)	--
Physics: How Things Work	1.8	1.7	2.1	1.8	2.1	1.7	1.6
	(0.4)	(0.4)	(0.3)	(0.4)	(0.6)	(0.6)	--
Biology: Medical Mysteries	2.4	2.3	2.4	2.3	2.0	1.8	--
	(0.6)	(0.7)	(0.6)	(0.5)	(0.7)	(0.3)	--
Maximum Volume	48%	48%	49%	49%	41%	40%	38%
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.1)
Open for Business	40%	37%	41%	38%	49%	40%	33%
	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)
Wheelchair Access	49%	52%	47%	49%	47%	40%	23%
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	--

Table 17

Means and Standard Deviations for Prior Academic Achievement

Task	Overall	Male	Female	White, Asian, Nat. Am. Or Multi	Black or Hispanic	Low SES	Special Education
OGT: Reading	0.0	-0.1	0.1	0.1	-0.6	-0.5	-1.0
	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
OGT: Writing	0.0	-0.2	0.2	0.0	-0.5	-0.4	-1.1
	(1.0)	(1.0)	(1.0)	(1.0)	(0.8)	(0.9)	(0.9)
OGT: Science	0.0	0.1	-0.1	0.1	-0.6	-0.4	-0.9
	(1.0)	(1.0)	(0.9)	(1.0)	(0.9)	(1.0)	(1.1)
OGT: Math	0.0	0.1	-0.1	0.1	-0.6	-0.4	-0.9
	(1.0)	(1.1)	(0.9)	(1.0)	(0.8)	(1.0)	(0.9)
OGT: Citizenship	0.0	0.1	-0.1	0.1	-0.6	-0.4	-0.9
	(1.0)	(1.1)	(0.9)	(1.0)	(0.9)	(1.0)	(1.0)
Cumulative GPA	3.0	2.9	3.2	3.1	2.6	2.7	2.5
	(0.7)	(0.8)	(0.7)	(0.7)	(0.8)	(0.8)	(0.7)

The Relationship Between Learning TaskScores and Other Academic Indicators. A series of correlation tables is presented below. The first, Table 18, shows the relationship between each of the OGT scores and the ELA, science or math performance task mean scores (for those who took them). The correlations of most interest are presented in bold.

The correlation between the ELA performance task scores and the other academic indicators is high, despite the restricted rangewith which it was scored. In fact, the correlation between the ELA performance task scores and a student's cumulative GPA (.58) is higher than the correlation between GPA and any of the OGT tests. The ELA performance task scoreshave a moderately high correlation with each of the OGT scores, hovering around .5 for each of the OGT subscores, with the highest correlation for the OGT Writing.

The science performance task scores have a slightly weaker relationship with other academic indicators. Although science task scores have a fairly strong positive and statistically

significant correlation with GPA (.42), this correlation is weaker than the relationship between GPA and any of the OGT scores. The science performance task scores had weak relationships with the OGT Reading, Writing, and Math scores (.34, .31, and .32 respectively), and there were no statistically significant relationships with the OGT Science scores. This suggests that the science performance tasks are measuring something distinct from the science OGTs.

The math performance task scores have the weakest relationships with other academic indicators. The correlation between the math performance task scores and GPA is moderately low (.36), as is the correlation with the OGT Math scores (.27) and even lower for the other OGT subject area tests.

The relationship between the performance task scores and the other academic indicators may be more complicated than a simple correlation would demonstrate, however. For example, different tasks may have different relationships with the academic indicators. In addition, problems with the academic indicators (lack of certainty around weighted or unweighted GPAs and the grade in which students took an OGT exam) could bias the relationship between the performance task scores and the academic indicators. We examine the first concern by consideringcorrelations between individual performance task scores and the other academic assessments, displayed in Tables 18a-18j. Note that in these tables, correlations are only made for students given the relevant task. Sample sizes indicated are for the correlation using GPA, which is a smaller sample given the missing GPA data. All correlations use a Sidack correction to adjust for multiple hypotheses.

Among the four ELA performance tasks, the Hero's Journey task scores are the most highly correlated with the OGT Reading scores (.64) and the OGT Writing scores (.70). On the other hand, *Employing the Personal* stands out as the least predictive of other academic

indicators. The correlation between the performance task score and GPA is positive but not statistically significant, and the correlation between the performance task score and the writing OGT is negative (though also not statistically significant). This unexpected finding may be because the task involves a genre of writing not typically included in most writing assessments given to students. The *Employing the Personal* task, while including a formal essay in which texts are analyzed, also includes a personal essay in which students use a personal experience to make a larger statement about something of importance to them. On the other hand, the relationship between GPA and OGT scores is also weaker in this sample (in some cases, these point estimates are also negative), so this may be an odd sample. The small sample size makes this a difficult question to answer with any certainty. On the other hand, the scores for *Employing the Personal* task had no statistically significant relationships with either the OGT tests or GPA, though this may be related to low sample sizes. These results could suggest that the task provides an opportunity for students who don't normally perform well on traditional measures (classroom grades, standardized tests) to demonstrate their competencies. Additional evidence is needed to ascertain how each task functions among different student populations (e.g., economically disadvantaged, LEP students, and disabled students).

Small sample sizes are a problem for all three science performance tasks. In science, all three performance task scores have the expected positive correlations between GPA and the task score (.4 for Chemistry and .44 for Biology). The strong correlation between task score and GPA for the Physics exam is not statistically significant, though,once again, this is likely due to the small sample size (N=10). In terms of the relationship between learning task scores and the OGT scores, the signs are positive as expected for all correlations though non-significant. In addition, the correlation between the biology task and the OGT Science score is surprisingly low (.14).

The only significant relationship was between the physics learning task score and the OGT Math test. Again, these results suggest that the science learning tasks are measuring something distinct from the OGT tests.

The math tasks seem to vary in their relationship with the other academic indicators. The math performance task scores and the OGT Math scores have a moderate and statistically significant relationship for *Maximum Volume* (.43) and a low but statistically significant correlation for *Wheelchair Access* (.37), while the relationship for Open for Business is non-significant. In *Wheelchair Access*, the correlation between the performance task score and GPA is low and positive (.35) though not statistically significant. However, the relationship between the performance task score and the OGT scores is stronger for this task, with all five correlations positive and three (including math) statistically significant.

Table 18

Correlations Between Performance Task Scores and Academic Indicators

	OGT: Reading	OGT: Writing	OGT: Science	OGT: Math	OGT: Citizenship	GPA
OGT: Writing	0.63 *					
OGT: Science	0.68 *	0.53 *				
OGT: Math	0.62 *	0.54 *	0.69 *			
OGT: Citizenship	0.71 *	0.56 *	0.74 *	0.65 *		
GPA	0.52 *	0.53 *	0.53 *	0.56 *	0.52 *	
ELA Performance Tasks	0.52 *	***0.54 ****	0.49 *	0.50 *	0.52 *	***0.58 ****
Science Performance Tasks	0.34 *	0.31 *	***0.21***	0.32 *	0.25	***0.42 ****
Math Performance Tasks	0.15 *	0.18 *	0.19 *	***0.27 ****	0.16 *	***0.36 ****

*NOTE: * p<.05. All correlation significance tests use a Sidak correction to adjust for multiple hypothesis tests. Correlations use all data available for pairwise correlation; sample sizes are therefore larger when correlating the OGT and GPA scores for students than when correlating performance task scores. Bolded Correlation coefficents represent the correlations of greatest interest for the purpose of our study.*

Table 18a

Correlations for Students Completing the Americans Dreaming Task (N=293)

	OGT: Reading	OGT: Writing	OGT: Science	OGT: Math	OGT: Citizenship	GPA
OGT: Writing	0.69 *					
OGT: Science	0.73 *	0.59 *				
OGT: Math	0.68 *	0.59 *	0.73 *			
OGT: Citizenship	0.79 *	0.65 *	0.80 *	0.74 *		
GPA	0.56 *	0.55 *	0.62 *	0.62 *	0.64 *	
ELA Performance Task	0.54 *	0.58 *	0.56 *	0.54 *	0.54 *	0.60 *

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment. Sample size given in the title is for the correlations using GPA.

Table 18b

Correlations for Students Completing the Constructing the Self Task (N=77)

	OGT: Reading	OGT: Writing	OGT: Science	OGT: Math	OGT: Citizenship	GPA
OGT: Writing	0.58 *					
OGT: Science	0.68 *	0.44 *				
OGT: Math	0.68 *	0.54 *	0.77 *			
OGT: Citizenship	0.70 *	0.50 *	0.73 *	0.70 *		
GPA	0.50 *	0.40 *	0.50 *	0.62 *	0.46 *	
ELA Performance Task	0.50 *	0.49 *	0.43 *	0.55 *	0.53 *	0.59 *

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment. Sample size given in the title is for the correlations using GPA.

Table 18c

Correlations for Students Completing the Employing the Personal Task (N=59)

	OGT: Reading	OGT: Writing	OGT: Science	OGT: Math	OGT: Citizenship	GPA
OGT: Writing	0.54 *					
OGT: Science	0.52 *	0.34				
OGT: Math	0.42 *	0.26	0.48 *			
OGT: Citizenship	0.60 *	0.19	0.62 *	0.24		
GPA	0.09	-0.10	-0.07	0.15	-0.01	
ELA Performance Task	-0.10	-0.19	-0.11	-0.22	-0.04	0.34

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment. Sample size given in the title is for the correlations using GPA.

Table 18d

Correlations for Students Completing the Hero's Journey Task (N=131)

	OGT: Reading	OGT: Writing	OGT: Science	OGT: Math	OGT: Citizenship	GPA
OGT: Writing	0.73 *					
OGT: Science	0.70 *	0.69 *				
OGT: Math	0.57 *	0.64 *	0.71 *			
OGT: Citizenship	0.79 *	0.70 *	0.75 *	0.62 *		
GPA	0.56 *	0.65 *	0.50 *	0.55 *	0.55 *	
ELA Performance Task	0.64 *	0.70 *	0.49 *	0.54 *	0.59 *	0.57 *

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment. Sample size given in the title is for the correlations using GPA.

Table 18e

Correlations for Students Completing the Chemistry Task (N=70)

	OGT: Reading	OGT: Writing	OGT: Science	OGT: Math	OGT: Citizenship	GPA
OGT: Writing	0.46 *					
OGT: Science	0.56 *	0.14				
OGT: Math	0.46 *	0.41 *	0.52 *			
OGT: Citizenship	0.52 *	0.37 *	0.58 *	0.55 *		
GPA	0.42 *	0.28	0.38 *	0.31	0.32	
Science Performance Task	0.31	0.27	0.30	0.34	0.24	0.40 *

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment. Sample size given in the title is for the correlations using GPA.

Table 18f

Correlations for Students Completing the Physics Task (N=10)

	OGT: Reading	OGT: Writing	OGT: Science	OGT: Math	OGT: Citizenship	GPA
OGT: Writing	0.51					
OGT: Science	0.42	0.42				
OGT: Math	0.71 *	0.50	0.67 *			
OGT: Citizenship	0.47	0.34	0.66 *	0.82 *		
GPA	0.38	0.69	0.57	0.48	0.35	
Science Performance Task	0.56	0.40	0.59	0.63 *	0.45	0.68

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment. Sample size given in the title is for the correlations using GPA.

Table 18g

Correlations for Students Completing the Biology Task (N=44)

	OGT: Reading	OGT: Writing	OGT: Science	OGT: Math	OGT: Citizenship	GPA
OGT: Writing	0.61 *					
OGT: Science	0.51 *	0.38				
OGT: Math	0.59 *	0.34	0.67 *			
OGT: Citizenship	0.51 *	0.47 *	0.62 *	0.45 *		
GPA	0.66 *	0.62 *	0.45 *	0.54 *	0.45 *	
Science Performance Task	0.38	0.33	0.14	0.31	0.29	0.44 *

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment. Sample size given in the title is for the correlations using GPA.

Table 18h

Correlations for Students Completing the Maximum Volume Task (N=247)

	OGT: Reading	OGT: Writing	OGT: Science	OGT: Math	OGT: Citizenship	GPA
OGT: Writing	0.56 *					
OGT: Science	0.63 *	0.48 *				
OGT: Math	0.54 *	0.45 *	0.68 *			
OGT: Citizenship	0.65 *	0.50 *	0.67 *	0.54 *		
GPA	0.51 *	0.53 *	0.51 *	0.55 *	0.51 *	
Math Performance Task	0.35	0.30 *	0.35 *	0.43 *	0.31 *	0.41 *

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment. Sample size given in the title is for the correlations using GPA.

Table 18i

Correlations for Students Completing the Open for Business Task (N=179)

	OGT: Reading	OGT: Writing	OGT: Science	OGT: Math	OGT: Citizenship	GPA
OGT: Writing	0.51 *					
OGT: Science	0.64 *	0.40 *				
OGT: Math	0.56 *	0.46 *	0.56 *			
OGT: Citizenship	0.63 *	0.44 *	0.69 *	0.55 *		
GPA	0.41 *	0.41 *	0.43 *	0.53 *	0.38 *	
Math Performance Task	-0.02	0.05	0.02	0.13	0.02	0.40 *

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment. Sample size given in the title is for the correlations using GPA.

Table 18j

Correlations for Students Completing the Wheelchair Access Task (N=62)

	OGT: Reading	OGT: Writing	OGT: Science	OGT: Math	OGT: Citizenship	GPA
OGT: Writing	0.61 *					
OGT: Science	0.73 *	0.55 *				
OGT: Math	0.61 *	0.60 *	0.60 *			
OGT: Citizenship	0.70 *	0.56 *	0.73 *	0.69 *		
GPA	0.56 *	0.70 *	0.51 *	0.55 *	0.51 *	
Math Performance Task	0.33	0.37 *	0.41 *	0.37 *	0.28	0.35

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment. Sample size given in the title is for the correlations using GPA.

ACT Scores and Performance Task Scores. ACT scores were reported in high school transcripts for approximately 400 students in the total sample. The link between ACT scores and first-year college GPA and attainment in college has been found to be empirically supported, though high school grades (GPA) have been found to be equally predictive for some aspects of college performance and attainment (See ACT, 2008 for a summary of empirical studies). Because not all of the students with performance assessment scores in the sample took the ACT and due to the nature of the sample of students who are more likely to take the ACT (college-bound juniors and seniors), these analyses are biased toward the upper end of the performance task score scales. Table 20 below shows that on average, ACT test-takers had higher scores on the performance tasks across content areas, and this difference is statistically significant ($p<.001$). So the generalizations that can be made based on these analyses are limited to college-bound ACT test takers who are likely those with higher academic performance overall. Nonetheless, there is still a range of scores on the performance tasks even within this sample.

Table 19

Descriptives - ACT Test Takers and Average Subtest Scores

	N	Min	Max	Mean	Std. Deviation
ACT_English	404	9	35	21.63	5.367
ACT_Math	402	14	36	22.41	4.691
ACT_Reading	403	4	36	22.90	5.816
ACT_Science	403	11	36	22.95	4.330
ACT_Composite Score	403	13	35	22.62	4.441
ACT_English/Writing Combined Score	304	11	33	21.50	7.252
ACT_Writing	308	4	25	7.24	1.788

Table 20

Performance Task Scores - ACT Takers and Non-Takers

		ELA Total Score	ELA Avg Score	Science Total Score	Science Avg Score	Math Total Score Percent
ACT Non-Takers	Mean	10.38	2.09	10.35	1.81	42.30%
	Std. Dev.	3.281	.653	3.245	.591	22.40%
	N	436	436	186	186	444
ACT Takers	Mean	12.65	2.55	11.94	2.13	51.70%
	Std. Dev.	3.344	.661	3.510	.650	21.30%
	N	169	169	85	85	154
Total	Mean	11.02	2.22	10.85	1.91	44.70%
	Std. Dev.	3.450	.686	3.405	.627	22.40%
	N	605	605	271	271	598

Analyses of the correlation between performance task scores and ACT scores is conducted by task because of the apparent differences in the score patterns across tasks (as demonstrated in the previous sections). Tables 21a-21j show the results of bi-variate correlations for each task and ACT scores. While low sample sizes limit the power of these analyses across several tasks, a few of the tasks had sufficient sample sizes to assess the relationship between performance task scores and ACT scores(ACT scores are not included in the multiple regression models that follow because they substantially reduce the power of the models by decreasing the number of cases in which performance task scores are matched with demographic data, transcript data, and OGT test score data). The *Americans Dreaming* (ELA) task scores had low to moderate correlations with the ACT Writing subtest and the Composite score ("ACT Comp"). Even more interesting is that the performance task scores for this task had consistently moderate correlations with the ACT Science subtest. The Constructing the Self (ELA) task scores had consistently moderate to high correlations with the ACT English/Writing combined score ("ACT Eng/Wr") and low to moderate correlations with the ACT English subtest (which primarily measures grammar and usage). The Employing the Personal (ELA) task scores had no significant correlations with the ACT scores, which suggests again that there is something anomalous about this ELA task. On the other hand, the correlations for the Hero's Journey (ELA) task scores, though limited by small sample sizes, were high and significant (up to .83 for the correlation between the average task score with the ACT English score and .80 for the ACT Composite score). These correlations across the ELA performance tasks suggest that in general, the ELA tasks are measuring some similar constructs as are measured by the ACT, though the correlations are low enough to suggest that the ELA tasks are also measuring something different

from the ACT, and there is wide variation in the relationships with ACT scores across performance tasks.

Table 21a

Correlations to ACT Scores - Americans Dreaming Task

	ACT English	ACT Math	ACT Reading	ACT Science	ACT Comp	ACT Eng/Wr	ACT Writing
Performance Task Scores							
Avg Score	**0.33** *	**0.28** *	**0.25** *	**0.49** *	**0.37** *	0.13	0.27
Analysis & Interpretation	**0.27** *	**0.26** *	0.19	**0.41** *	**0.30** *	0.08	0.27
Perspective	**0.40** *	**0.35** *	**0.36** *	**0.55** *	**0.45** *	0.12	0.22
Power of Language	**0.42** *	**0.29** *	**0.36** *	**0.48** *	**0.43** *	0.18	0.20
Structure, Org, & Conventions	**0.24** *	0.23	0.17	**0.37** *	**0.27** *	0.11	0.13
Reflection	**0.11** *	0.08	0.04	**0.29** *	0.15	0.08	**0.33** *
N	68	68	68	68	68	51	51

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment.

Table 21b

Correlations to ACT Scores - Constructing the Self Task

	ACT English	ACT Math	ACT Reading	ACT Science	ACT Comp	ACT Eng/Wr	ACT Writing
Performance Task Scores							
Avg Score	**.47** *	**.35** *	**.26** *	**.35** *	**.41** *	**.60** *	**.27** *
Analysis & Interpretation	**.47** *	**.30** *	**.30** *	**.38** *	**.42** *	**.56** *	**.29** *
Perspective	**.37** *	.19	.17	.17	**.27** *	**.51** *	.20
Power of Language	**.47** *	**.40** *	**.34** *	**.40** *	**.48** *	**.61** *	.25
Structure, Org, & Conventions	**.41** *	**.30** *	.17	.12	**.30** *	**.51** *	.20
Reflection	.12	.17	.03	.09	.12	.23	.15
N	62	62	62	62	62	52	52

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment.

Table 21c

Correlations to ACT Scores - Employing the Personal Task

	ACT English	ACT Math	ACT Reading	ACT Science	ACT Comp	ACT Eng/Wr	ACT Writing
Performance Task Scores							
Avg Score	.38	-.13	.21	.15	.21	.26	.29
Analysis & Interpretation	.29	-.16	.22	.13	.17	.21	.27
Perspective	.38	-.10	.29	.25	.27	.43	.33
Power of Language	.38	-.14	.19	.07	.17	.27	.43
Structure, Org, & Conventions	.27	-.07	.11	.11	.14	.21	.27
Reflection	.36	-.12	.12	.11	.18	.08	.00
N	26	26	26	26	26	12	12

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment.

Table 21d

Correlations to ACT Scores - Hero's Journey Task

	ACT English	ACT Math	ACT Reading	ACT Science	ACT Comp	ACT Eng/Wr	ACT Writing
Performance Task Scores							
Avg Score	**.83** *	**.68** *	.51	.46	**.80** *	.62	.59
Analysis & Interpretation	**.63** *	**.62** *	.43	.29	**.66** *	**.65** *	**.66** *
Perspective	**.89** *	.53	**.64** *	.45	**.82** *	.48	.46
Power of Language	**.74** *	**.64** *	.30	.46	**.67** *	.62	.59
Structure, Org, & Conventions	**.85** *	.52	.25	.23	**.58** *	.30	.26
Reflection	**.64** *	**.70** *	**.62** *	**.58** *	**.83** *	**.72** *	**.69** *
N	13	13	13	13	13	10	10

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment.

The relationships between the science performance task scores and ACT scores were even more variable across tasks. In the chemistry task *Got Relieve It?*the only significant correlations (moderate) were for ACT math scores, suggesting that this particular task requires proficiency in mathematical skills or mathematical reasoning. In the physics task *How Things Work*, it appears that the only significant correlation (which was limited by very low sample sizes) was for the dimension "Communicate and Present Findings", which was highly correlated

with the ACT English, Math, Composite, and combined English/Writing scores. The correlations were high for other dimensions of the scoring rubric but none were significant due to low sample sizes. This task has already been eliminated from the pilot, but greater samples sizes would have been helpful for assessing the relationship of this task and ACT scores. In addition, it should be noted that the high (though non-significant) correlations may be a function of the sample of students who typically take physics courses.

The biology task *Medical Mysteries* had the strongest correlations with ACT scores, with the average score moderately correlated with the ACT English, Math, Reading, Composite, and English/Writing scores, although not with the ACT Science subscore. The "Collect & Connect Content Knowledge" dimension, however, was moderately correlated with the ACT Science score. These findingssuggest the ACT Science subtest is more of a test of content knowledge while the performance tasks assess different aspects of the inquiry process and students' ability to apply their content knowledge. This may explain the overall low correlation levels between the average scores on the science performance tasks and the ACT Science subscore.

Table 21e

Correlations to ACT Scores - Chemistry Task

	ACT English	ACT Math	ACT Reading	ACT Science	ACT Comp	ACT Eng/Wr	ACT Writing
Performance Task Scores							
Avg Score	.24	**.41** *	.05	.23	.25	.18	.15
Collect & Connect Content	.15	**.38** *	.09	.23	.24	.09	.15
Design of Investigation	.23	**.30** *	.06	.16	.19	.15	.23
Analyze & Interpret Data	.28	**.38** *	.07	.18	.24	.23	.11
Draw Conclusions	.17	**.39** *	.02	.24	.23	.13	.07
Communicate Findings	.26	**.39** *	.04	.17	.23	.24	.05
Reflect on Learning	.14	.26	-.03	.17	.13	.08	.15
N	48	47	48	48	48	40	41

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment.

Table 21f

Correlations to ACT Scores - Physics Task

	ACT English	ACT Math	ACT Reading	ACT Science	ACT Comp	ACT Eng/Wr	ACT Writing
Performance Task Scores							
Avg Score	.59	.54	.27	.76	.54	.53	.54
Collect & Connect Content	.65	.73	.27	.67	.57	.75	.79
Design of Investigation							
Analyze & Interpret Data	.67	.62	.28	.61	.56	.82	.80
Draw Conclusions	.68	.73	.53	.76	.67	.64	.76
Communicate Findings	**.84** *	**.93** *	.60	.81	**.82** *	**.91** *	.85
Reflect on Learning	.19	.18	.04	.43	.18	-.03	.13
N	6	6	6	6	6	5	5

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment.

Table 21g

Correlations to ACT Scores - Biology Task

	ACT English	ACT Math	ACT Reading	ACT Science	ACT Comp	ACT Eng/Wr	ACT Writing
Performance Task Scores							
Avg Score	**.47** *	**.40** *	**.53** *	.30	**.53** *	**.45** *	.30
Collect & Connect Content	**.59** *	**.40** *	**.62** *	**.43** *	**.64** *	**.49** *	.11
Design of Investigation							
Analyze & Interpret Data	.24	.21	.34	.06	.28	.22	.17
Draw Conclusions	.50	.45	.48	.41	.55	.47	.31
Communicate Findings	**.49** *	.35	**.55** *	.26	**.51** *	**.49** *	.37
Reflect on Learning	.16	.28	.27	.11	.25	.22	.33
N	30	30	30	30	30	26	27

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment.

The correlations for the mathematics task scores with ACT scores was also highly variable by task. The total percentage scores for the task *Maximum Volume* had consistently low correlations (.35 or less) with the ACT subtest scores, including ACT Math. On the other hand the total percentage scores for the *Open for Business* task had moderately high correlations (from .43 to .60) with the ACT subtest scores, including a correlation of .56 with the ACT Math scores. The scores for *Wheelchair Access* had low, non-significant correlations due to low samples sizes, with the exception of the Communication scoring dimension, which had a .55 correlation with the ACT Reading scores. Overall, it appears that there is little overlap between the mathematics performance tasks and the ACT Math subtest, with the exception of the *Open for Business* task.

This contrasts with the consistently strong correlations between the OGT test scores and the ACT scores, which hover around the .60 range, but ranging from .3 to .75. There are also moderately strong correlations between cumulative GPAs and the ACT scores (ranging from .26 for writing to .60 for English and other subtests.)

In sum, there appear to be low to moderate relationships with OGT test scores as well as the ACT scores, suggesting that the performance tasks are measuring constructs that are substantially different from those measured by these other standardized measures. Again, these relationships vary by task, suggesting a need for greater standardization in the design of the tasks within content area genres (e.g., math problem-solving) so that there is greater consistency in the constructs measured and comparability in the rigor and difficulty of each task.

Table 21h

Correlations to ACT Scores - Maximum Volume Task

	ACT English	ACT Math	ACT Reading	ACT Science	ACT Comp	ACT Eng/Wr	ACT Writing
Performance Task Scores							
Percentage Total Score	**.35** *	**.35** *	.13	**.28** *	**.33** *	**.30** *	.21
Mathematics	.01	-.01	-.02	-.02	.00	-.10	-.08
Mathematical Reasoning	**.31** *	**.30** *	.09	**.23**	**.27** *	**.29** *	.18
Communication	**.36** *	**.39** *	**.21** *	**.35** *	**.39** *	**.37** *	**.31** *
Approach							
N	112	112	112	112	112	78	77

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment.

Table 21i

Correlations to ACT Scores - Open For Business Task

	ACT English	ACT Math	ACT Reading	ACT Science	ACT Comp	ACT Eng/Wr	ACT Writing
Performance Task Scores							
Percentage Total Score	**.43** *	**.56** *	.37	**.60** *	**.53** *	.58	.26
Mathematics	.46	.51	.39	.53	.52	.53	.34
Mathematical Reasoning	.20	.36	.16	.41	.30	.29	.01
Communication	.36	**.47** *	.40	**.65** *	**.51** *	**.59** *	.28
Approach	.40	**.52** *	.32	**.45** *	**.48** *	**.65** *	.25
N	22	21	21	21	21	12	15

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment.

Table 21j

Correlations to ACT Scores - Wheelchair Access Task

	ACT English	ACT Math	ACT Reading	ACT Science	ACT Comp	ACT Eng/Wr	ACT Writing
Performance Task Scores							
Percentage Total Score	.25	.42	.38	.31	.39	.29	.08
Mathematics [a]	.16	.30	.08	.11	.18	.24	.19
Mathematical Reasoning	.29	.43	.43	.41	.45	.17	-.18
Communication Approach	.20	.36	**.55** *	.33	.42	.33	.19
N	17	17	17	17	17	16	16

a For this task, Mathematics is the sum of scores for the Details and Guidelines dimensions.

* p<.05. All correlations use all available observations for the sample of students given the relevant assessment.

Multivariate Relationships (Regression analysis). As noted earlier, the relationship between the performance assessment tasks and the other academic indicators may be more complicated than a simple correlation would demonstrate. First, different tasks within a subject may have different relationships with the academic indicators. Second, the relationship between academic indicators and the performance task scores could be a result of strong students performing well in both measures. Finally, problems with the academic indicators (lack of certainty around weighted or unweighted GPAs and the grade in which students took an OGT exam) could bias the relationship between the assessments and the academic indicators. From the bivariate correlations, we saw that several tasks varied in terms of the strength – and even the direction – of their relationship to the other academic indicators. We were unable to parse out to what extent the learning tasks measured the specific academic skills they were designed to measure and to what extent they measured the student's general academic ability. To answer this second question, we utilize a multivariate regression analysis.

Specifically, we estimate the following base equation three times for each subject *s*:

$$Y_i^S = \beta_0 + \boldsymbol{OGT_i\beta_1} + \boldsymbol{Task_i\beta_2} + (\boldsymbol{Task_i xOGT_i^s})\boldsymbol{\beta_3} + \beta_4 GPA_i + (\boldsymbol{Task_i xGPA_i})\boldsymbol{\beta_5} + \mu_i$$

where, the outcome Y_i^S is the learning task score for student *i* in subject *s*. The first time we estimate the coefficients in the equation, we regress the outcome variable, Y_i^S, on all five OGT scores (which have been standardized within the sample) and the student's cumulative GPA. We also include a vector of indicator variables for the task completed and an interaction term between both the relevant OGT subject and GPA and the task taken. This interaction term allows us to test whether each task has a different relationship between the most important academic indicators.

The second time we estimate this equation, we include a vector of indicator variables for individual student demographic characteristics. The vector includes the following variables: gender (=1 if female), race (=1 if African American or Hispanic), years in high school (reference group began in 2008), LEP status (=1 if student is LEP), low SES (=1 if low SES), gifted (=1 if gifted) or special education (=1 if receives special education). The coefficients on these demographic variables should be interpreted with caution. There is error in a number of the regressors (GPA may be weighted or not, OGT scores may be from 10th-12th grade, not to mention the measurement error that exists in the OGT scores) which means that the coefficients on the demographic variables will likely be biased towards the unconditional group differences. We include them primarily as control variables. Including these controls allows for us to ask whether the student's individual characteristics drive the regression model findings.

The third time we estimate this equation, we include a vector of controls for school level characteristics. These control variables include the number of students in the school, the percent of the school that is African American or Hispanic, the percent of the school eligible for free or reduced price lunch, the student-teacher ratio, and a vector of dummy variables for the school's

urbanicity (reference group being a large suburb). Again, we should expect that these variables act more as controls for the student's ability and school quality than as any coefficient of interest in their own right, especially given the small number of schools in each sample.

Results can be seen in Tables 22a (ELA), 22b (Science) and 22c (Math). For ELA, the *Americans Dreaming* task serves as the reference, the Chemistry task *Got Relieve It* serves as the reference for the science assessments and *Maximum Volume*as the reference for the math assessments. The coefficients of most interest are highlighted on each table. The indicator variables for the tasks tell us if some tasks had overall higher or lower scores than the others. Ideally, these coefficients should be zero, indicating that students of the same GPA and OGT scores would be judged equally as proficient, regardless of the task their teacher assigned. We are also interested in the relationship between the OGT in the relevant subject and the student's GPA and the assessment score. Since we include interaction terms, the "main effect" can be interpreted as the relationship between the academic indicator (OGT or GPA) and the performance task score for students who took the performance task in the reference group (*Americans Dreaming*, *Chemistry-Got Relieve It* or *Maximum Volume*). The interaction terms are the difference between this relationship for the reference task and the interacting task. Again, ideally these interaction terms would all be zero, indicating that the relationship between a student's academic ability and their performance task score is the same, regardless of which task they were administered.

In Table 22a, we can see that, at least for the reference task, *Americans Dreaming*, there is a strong positive relationship between the performance task score and both the writing OGT and GPA. The main effects and interaction terms for the *Constructing the Self* task indicate that students perform similarly on both tasks when controlling for their academic ability and

measures of their individual and school characteristics. *Hero's Journey* appears to have a stronger relationship with the OGT writing and weaker relationship with GPA than these other two tasks. The point estimate for the main effect for *Hero's Journey* is fairly high, but due to very large standard errors we cannot conclude that students of the same ability tend to get a higher score on the *Hero's Journey* task than the *Americans Dreaming* task. The *Employing the Personal* task, on the other hand, appears to have a weaker relationship to other measures of academic achievement.

Overall, students of the same ability tend to score lower on the *Employing the Personal* task, but standard errors make this difference statistically non-significant in the third model. In all three models, the interaction term indicates that the *Employing the Personal* has a negative relationship with the OGT writing. The relationship between this task and GPA seems to be roughly the same as for *Americans Dreaming*. Overall, therefore, there is evidence that the *Americans Dreaming* and *Constructing the Self* tasks have roughly equal and strong relationships with other measures of student's academic ability, while the other two tasks, *Hero's Journey* and *Employing the Personal*, have weaker relationships (*The Hero's Journey* task was dropped in subsequent pilot cycles because of the small number of teachers interested in using the task, but also because of its mis-alignment with high school level themes - it was reported that the hero theme is typically covered in middle school.)

In Table 22b, we reduce our sample size substantially – from 560 in the ELA sample to 124 in the science sample. This reduction of sample size – especially for physics, for which we only have 10 students, means that the point estimates – especially for the interaction terms-- will be less reliable. It does appear as though the Chemistry performance task is much more strongly related to GPA than the science OGT. The coefficients for the science OGT scores are positive,

though less than a standard error away from zero. Point estimates suggest that the biology task *Medical Mysteries* has a negative relationship with the science OGT and potentially an even stronger relationship with GPA than the chemistry task.

The sample size for the mathematics learning task analyses is more substantial (N=488). When interpreting the results from the math tasks in Table 22c, we must take care to interpret the coefficients differently because of the use of a different scale (the math performance task scores were computed as percentages since the score scale varies for each task). We therefore could say that the reference task, *Maximum Volume*, has a fairly weak relationship with the math OGT exam. A student who scores a full standard deviation higher on the math OGT would be predicted to score only five percentage points higher on the math performance task (holding all other controls in the model constant). This relationship looks fairly similar for all three tasks. The relationship between GPA and the math learning task scores looks more as we might expect. Holding all other controls in the model constant, we expect that a student with a full letter grade higher GPA would score about 8 percentage points higher on the math performance task. Since 10 points usually indicates a full letter grade, this appears to be a strong relationship. The relationship between GPA and math learning task score appears to be similar for the *Maximum Volume* and the *Wheelchair Access* task, though there may be a stronger relationship between the *Open for Business* task and GPA. Students with the same OGT scores and GPA who complete the *Open for Business* task should expect a substantially lower score than they would on the other math performance tasks.

Discussion

The studies presented above relate to two aspects of validity discussed in the conceptual framework of this paper: *Convergent and Discriminant Correlations with External Variables;*

and Fairness. There are several findings that should be highlighted. First, in almost all cases across all content areas, correlation and regression analyses indicate that there is a stronger relationship between performance task scores with GPA than with OGT scores. This is consistent with hypotheses about what the OPAPP tasks are measuring. Because the tasks are "curriculum-embedded" and are completed as class work or course assignments by students, it is not surprising that they would have a stronger relationship to grades than to test scores. Second, the regression analyses indicate that different tasks within the same content area have relationships to GPA and OGT scores of different magnitude and significance. This suggests that the performance tasks are functioning differently than others in the same content area - in other words, you cannot directly compare the results of one task to another because the tasks appear not to be comparable.[12]This variability by task was also evident in the analyses of the relationship of the performance task scores to the ACT subtest scores. The level of correlation between the performance task scores and the relevant ACT subtest scores were for some tasks moderately high and significant, and for other tasks, low and/or non-significant.

Furthermore, the variability in tasks as measures is evident even without the use of score data analysis. At face value, the content of each task differs, as do the competencies assessed in each task. The approach to task development used for the OPAPP learning tasks was not sufficiently standardized to result in comparable tasks. *This is one of the important lessons learned from this pilot - the need for a common "task shell" or sets of task specifications that pre-define the constructs to be measured so that the design of each task within the same task genre (e.g., science inquiry) is more standardized and comparable.*However, given the evolving

[12] A task equating study, in which the same students complete two or more tasks one after another to reduce learning effects, would be necessary to determine the comparability of tasks. That study is practically unfeasible for ELA and science because of the curriculum embedded nature of the tasks, particularly for science, when the content covered varies by time and sequence in the curriculum. This study would be more easily conducted in mathematics.

context in which the OPAPP learning tasks are to be used (as formative, instructional tasks, rather than as summative assessments), it is less critical that the learning tasks are standardized or completely comparable. The performance tasks that *are* designed to be used as summative measures (OPAPP assessment tasks) should be developed using task shells to ensure greater standardization and comparability.

Fairness. The second regression model represented in Tables 22a, 22b, and 22c shows the extent to which demographic factors contribute to learning task scores. The third regression model shows the extent to which school-level demographic factors contribute to learning task scores. We find that there are differences across demographic groups in certain cases across the different content area tasks. In the English langauge arts task *Americans Dreaming*, females had a slight but significant edge over male students (.17 point higher on a four point scale), students attending career/vocational education centers that enroll students in the 11th grade scored .27 point lower than those enrolled in traditional four-year high schools, and students attending schools in suburban towns were likely to score higher than average than students attending schools in large urban centers (.39 point higher). On the mathematics task *Maximum Volume*, "low SES" students ("economically disadvantaged" as defined by ODE) had very slightly lower scores than students not in "low SES" schools (.06 point on a 25 point scale), although in model 3, students attending "low SES" schools scored significantly higher (1.18 points) than students not enrolled in "low SES" schools. in addition, students attending career/vocational education academies (that start in 11th grade) scored slightly higher than students enrolled in traditional four-year high schools (.32 point higher), students enrolled in schools with a higher percentage of black or Hispanic students scored slightly higher (4.71 points out of 25 points) than students in schools with lower percentage of minority student enrollment, and students enrolled in

suburban schools ("fringe town") scored slightly higher (1.85 point) than students enrolled in schools located in urban centers. There were no significant differences in the Chemistry task scores for students across demographic groups. While the slightly lower scores for students enrolled in urban schools and for students in vocational schools (ELA only) are troubling, most of the other differences are very slight and do not raise as much concern. Slightly higher scores on the mathematics task *Maximum Volume* for students in schools with a higher percentage of black/Hispanic students and for students in low-SES schools is actually quite encouraging.

Table 22a. Multivariate Regression on ELATask Score

(Reference: Americans Dreaming task)

	(1)		(2)		(3)	
Constructing the Self	-0.10		0.00		0.10	
	(0.33)		(0.34)		(0.36)	
Employing the Personal	-0.30		-0.18		-0.18	
	(0.35)		(0.35)		(0.37)	
Hero's Journey	0.36		0.39		0.38	
	(0.27)		(0.27)		(0.29)	
OGT: Reading	0.04		0.02		0.03	
	(0.04)		(0.04)		(0.04)	
OGT: Writing (for Americans Dreaming)	0.17	***	0.11	**	0.11	**
	(0.04)		(0.04)		(0.04)	
Constructing the Self x OGT: Writing	-0.08		-0.03		-0.01	
	(0.08)		(0.08)		(0.08)	
Employing the Personal x OGT: Writing	-0.32	***	-0.31	***	-0.28	**
	(0.09)		(0.09)		(0.09)	
Hero's Journey x OGT: Writing	0.16	*	0.16	*	0.13	
	(0.06)		(0.07)		(0.07)	
OGT: Science	0.01		0.02		0.00	
	(0.04)		(0.04)		(0.04)	
OGT: Math	0.03		0.07	*	0.06	
	(0.03)		(0.03)		(0.03)	
OGT: Citizenship	0.04		0.07		0.07	
	(0.04)		(0.04)		(0.04)	
Cumulative GPA (for Americans Dreaming)	0.29	***	0.24	***	0.25	***
	(0.05)		(0.05)		(0.05)	
Constructing the Self x Cumulative GPA	0.09		0.04		-0.02	
	(0.11)		(0.11)		(0.11)	
Employing the Personal x Cumulative GPA	0.03		-0.02		-0.08	
	(0.13)		(0.13)		(0.13)	
Hero's Journey x OGT: Cumulative GPA	-0.13		-0.13		-0.10	
	(0.09)		(0.09)		(0.09)	
Female			0.19	***	0.17	**
			(0.05)		(0.05)	
Black or Hispanic			-0.06		-0.05	
			(0.08)		(0.08)	
Low SES			-0.07		-0.06	
			(0.06)		(0.06)	
LEP			-0.14		-0.28	
			(0.16)		(0.17)	
Special Education			-0.06		-0.02	
			(0.08)		(0.08)	
Gifted			-0.04		-0.07	
			(0.07)		(0.07)	
School Begins in 11th Grade					-0.27	**
					(0.09)	
Number of Students in School					0.00	
					(0.00)	
% of School Black or Hispanic					1.20	
					(1.22)	
% of School Low SES					0.20	
					(0.19)	
Student Teacher Ratio					-0.03	
					(0.03)	
Mid-sized City					0.00	
					(0.00)	
Small City					-0.72	
					(0.42)	
Fringe Town					0.39	*
					(0.19)	
Distant Town					0.00	
					(0.00)	
Rural Fringe					0.06	
					(0.16)	
Distant Rural					-0.20	
					(0.21)	
Constant	1.42	***	1.48	***	2.05	**
	(0.14)		(0.15)		(0.66)	
N	560		560		560	
Adjusted R-Squared	0.45		0.46		0.47	

*NOTE: *p<.05, **p<.01 ***p<.001. Models 2 and 3 also include controls for the year the student began school, which are omitted from the table to save space.*

Table 22b. Multivariate Regression on Science Task Score

(Reference: Chemistry task)

	(1)	(2)	(3)
Physics	-0.85	-0.40	-2.07
	(1.26)	(1.35)	(2.16)
Biology	0.22	-0.18	-1.59
	(0.70)	(0.77)	(1.20)
OGT: Reading	0.00	-0.01	-0.02
	(0.08)	(0.09)	(0.09)
OGT: Writing	0.05	0.02	0.06
	(0.07)	(0.08)	(0.08)
OGT: Science	0.07	0.06	0.08
	(0.11)	(0.12)	(0.13)
Physics x OGT Science	-0.07	0.00	0.02
	(0.30)	(0.32)	(0.33)
Biology x OGT Science	-0.25	-0.28	-0.30
	(0.14)	(0.15)	(0.16)
OGT: Math	0.13	0.18	0.20 *
	(0.09)	(0.10)	(0.10)
OGT: Citizenship	0.03	0.08	0.05
	(0.07)	(0.08)	(0.08)
Cumulative GPA	0.39 *	0.36 *	0.34
	(0.15)	(0.16)	(0.18)
Physics x Cumulative GPA	0.12	-0.01	-0.09
	(0.39)	(0.42)	(0.42)
Biology x OGT: Cumulative GPA	0.06	0.20	0.22
	(0.22)	(0.24)	(0.25)
Female		0.06	0.07
		(0.12)	(0.12)
Black or Hispanic		0.12	0.19
		(0.25)	(0.25)
Low SES		0.01	-0.12
		(0.19)	(0.21)
LEP		0.00	0.00
		0.00	0.00
Special Education		0.00	0.00
		0.00	0.00
Gifted		-0.15	-0.18
		(0.12)	(0.13)
School Begins in 11th Grade			0.00
			0.00
Number of Students in School			(0.00)
			(0.00)
% of School Black or Hispanic			0.00
			(0.00)
% of School Low SES			0.00
			(0.00)
Student Teacher Ratio			-0.18
			(0.15)
Mid-sized City			-1.94
			(1.54)
Small City			0.00
			(0.00)
Fringe Town			-0.50
			(0.79)
Distant Town			0.00
			(0.00)
Rural Fringe			0.00
			(0.00)
Distant Rural			-2.79
			(2.41)
Constant	0.66	0.79	7.04
	(0.47)	(0.50)	(4.83)
N	124	124	124
Adjusted R-Squared	0.29	0.28	0.29

*NOTE: *p<.05, **p<.01 ***p<.001. Models 2 and 3 also include controls for the year the student began school, which are omitted from the table to save space*

Table 22c. Multivariate Regression on Math Task Score

(Reference: Maximum Volume task)

	(1)		(2)		(3)	
Open for Business	-0.36	***	-0.37	***	-0.31	**
	(0.11)		(0.10)		(0.11)	
Wheelchair Access	-0.10		-0.10		-0.06	
	(0.20)		(0.19)		(0.18)	
OGT: Reading	0.01		0.00		0.00	
	(0.01)		(0.01)		(0.01)	
OGT: Writing	-0.01		0.00		0.01	
	(0.01)		(0.01)		(0.01)	
OGT: Science	0.02		0.02		0.03	*
	(0.01)		(0.01)		(0.01)	
OGT: Math	0.03		0.04	*	0.05	**
	(0.02)		(0.02)		(0.02)	
Open for Business x	-0.02		-0.02		-0.04	
OGT Math	(0.03)		(0.03)		(0.03)	
Wheelchair Access x	0.01		0.00		-0.01	
OGT Math	(0.03)		(0.03)		(0.03)	
OGT: Citizenship	-0.03		-0.02		-0.02	
	(0.01)		(0.01)		(0.01)	
Cumulative GPA	0.08	***	0.08	***	0.08	***
	(0.02)		(0.02)		(0.02)	
Open for Business x	0.07	*	0.06		0.06	
Cumulative GPA	(0.03)		(0.03)		(0.03)	
Wheelchair Access x	0.02		0.02		0.00	
Cumulative GPA	(0.06)		(0.06)		(0.05)	
Female			0.00		0.00	
			(0.02)		(0.02)	
Black or Hispanic			0.06		0.05	
			(0.03)		(0.03)	
Low SES			-0.06	*	-0.06	*
			(0.02)		(0.02)	
LEP			0.18		0.13	
			(0.11)		(0.12)	
Special Education			-0.04		-0.04	
			(0.06)		(0.05)	
Gifted			-0.01		-0.01	
			(0.02)		(0.02)	
School Begins in 11th Grade					0.32	***
					(0.08)	
Number of Students in School					0.00	**
					(0.00)	
% of School Black or Hispanic					4.71	**
					(1.81)	
% of School Low SES					(1.18)	***
					(0.34)	
Student Teacher Ratio					-0.53	**
					(0.17)	
Mid-sized City					0.00	
					(0.00)	
Small City					-3.23	**
					(1.20)	
Fringe Town					1.85	***
					(0.55)	
Distant Town					-1.98	**
					(0.64)	
Rural Fringe					0.67	***
					(0.20)	
Distant Rural					0.55	***
					(0.16)	
Constant	0.26	***	0.23	**	10.76	**
	(0.07)		(0.07)		(3.40)	
N	488		488		488	
Adjusted R-Squared	0.24		0.30		0.36	

*NOTE: *p<.05, **p<.01 ***p<.001. Models 2 and 3 also include controls for the year the student began school, which are omitted from the table to save space*

Content Validity

External evaluators (the Ohio Evaluation & Assessment Center out of Miami University) conducted a standards alignment review with external reviewers comparing the content and task demands of the OPAPP tasks and rubrics to the Common Core State Standardsas well as the 21st Century Skill Standards.[13] The results of this comparison are summarized next.

Although the OPAPP learning tasks were not designed with the Common Core State Standards in mind (since the standards were released *after* the initial development of the OPAPP instruments), it appears that there were moderate to high levels of alignment between the mathematics and ELA tasks/rubrics and the CCSS in terms of both breadth and depth (although there was variation across tasks). In fact, the ELA tasks go beyond the CCSS in terms of the task demands and the types of analytical thinking and writing that it requires, including comparing/contrasting ideas, persuasive writing, and reflective writing.

Six out of eight science tasks were deemed to have a moderate to high level of alignment with the Ohio science content standards (with *Got Clean Water, Energy Efficient Vehicles*, and *How Things Work* having the highest ratings of alignment), while the other two tasks (*Medical Mysteries* and *Is It Physical or Chemical*?) had marginal alignment.

In terms of alignment to 21st Century Skills (as defined by the Partnership for 21st Century Skills, of which Ohio is a member state), the math tasks were found to be moderately to highly aligned to the following skills: creativity and innovation, critical thinking and problem solving, communication and collaboration, information media, and technology, life and career, and leadership and responsibility. The ELA tasks were found to be highly aligned to the following skills: critical thinking and problem solving; information, media, and technology; life

[13] It is not clear from their description of methodology how many reviewers were asked to examine the alignment of each set of content area tasks. The science tasks and rubrics were evaluated against the new Ohio content standards for science.

and career skills; productivity and accountability. Last, the science tasks were found to be highly aligned to critical thinking and problem solving skills as well as life and career skills. For more details on this content alignment review, please see pages 18-21 of Woodruff, Zorn, Castañeda-Emenaker, & Sutton, J. (2010b) as well as the Supplementary Documents that accompany that report.

Higher Education Validity Review. In this study**,** higher education faculty teaching entry level courses (excluding remedial courses) in English language arts, chemistry, physics, and mathematics were recruited from Ohio universities, including research universities, private and public universities, and local colleges. Four faculty members who taught entry level English/writing coureses were recruited to review the four English language arts tasks, four faculty members who taught entry level mathematics courses were recruited to review the six of the seven mathematics tasks[14], and two faculty members teaching entry level chemistry and two faculty members teaching entry level physics were recruited to review the two chemistry and two physics tasks respectively. This sample of volunteer reviewers is one of convienience built based on word of mouth references from initial contacts made at a range of universities.Each reviewer was compensated for their time with a modest stipend.

Each of the reviewers were sent the relevant learning tasks, one benchmark work sample that represented the level of proficiency required to be "college and career ready" for each task[15], and the relevant set of scoring rubrics. Reviewers were given a series of questions and rating scales on which to rate each learning task, work sample, and scoring rubric (in that sequence). After they completed several background questions about their own teaching history and the

[14] One task *Body Surface Area* was omitted because it was not widely piloted during the OPAPP pilot, did not have any representative work samples, or a task-specific rubric developed to score it.
[15] These "benchmarks" were preselected and scored by expert scorers for the purpose of selecting anchor papers for scorer training. The "Level 3" benchmarks in almost all cases were the samples that were selected for review because this level of work was defined as demonstrating "college and career readiness".

courses they teach, the faculty members were asked to rate the content of the learning tasks on the following items:

Please read each performance task, student work sample, and scoring rubric in the order of the questions below, and answer the following questions in the order in which they are posed. *Please do NOT read ahead or look at materials that are not referenced in each set of questions.*

A. MATH TASK 1: "GRAZING AREA"

1. What is the relevance of the content knowledge needed to complete Math Task 1 to your entry level (first year) math course: (CHECK ALL THAT APPLY)

_____Prerequisite to the course
_____Reviewed in the course
_____Covered or taught in the course
_____More advanced than the course
_____Not relevant to the course

Briefly explain:_____

2. What is the relevance of the skills or work habits needed to complete Math Task 1 to your entry level (first year) math course: (CHECK ALL THAT APPLY)

_____Prerequisite to the course
_____Reviewed in the course
_____Covered or taught in the course
_____More advanced than the course
_____Not relevant to the course

Briefly explain:_____

Following the review of the task, faculty members were then asked to review the benchmark student work sample that was scored at the "level 3" or "proficient" score level on the following questions:

STUDENT SAMPLE 1: RESPONSE TO THE TASK "GRAZING AREA"

3. Based on Student Sample 1, to what extent would you say that the high school student who produced this sample is prepared for your entry-level (first year) course: (CHECK ONE ONLY)

_____Unlikely to succeed in the course (would not pass the course)
_____Somewhat likely to succeed in the course (is likely to need remedial coursework)
_____Likely to succeed in the course (will not need remediation)
_____Very likely to succeed in the course (will pass the course with an A or B)

Briefly explain:_____

Finally, reviewers were asked to rate the relevant scoring rubrics used to score student work on the following item. They were asked to rate EACH scoring dimension of the rubrics:

SCORING RUBRIC 1: "GRAZING AREA"

4. Examine the task-specific criteria that were used in the OPAPP pilot to score student work on the Grazing Area task. For each scoring criterion, **please rate the relevance of each criterion for success in entry level college work**. Do NOT rate the student work, just the scoring instrument itself.

1=NOT RELEVANT, 2=MARGINALLY RELEVANT, 3=SOMEWHAT RELEVANT, 4=RELEVANT, 5=VERY IMPORTANT

Initial results from the higher education faculty validity review (which will be reported fully in the next version of this conference paper) indicate that the ELA and mathematics task demands and rubric criteria are aligned with expectations for entry level courses in English language arts and mathematics. In most cases the content and/or skills addressed by the tasks were considered "prerequisites", "reviewed in the course", or "covered or taught in the course") and most of the dimensions of the scoring rubrics were considered "somewhat relevant", "relevant", or "important" to entry level college courses. Students represented by "Level 3"

benchmark work samples were rated in most cases as "Likely to succeed in the course", which is what we expected. In the case of the science inquiry tasks, faculty rated the content covered by the tasks as "prerequisites" or "reviewed in the course", but the skills addressed by the science inquiry tasks in some cases were "more advanced than the course". This is because it is rare that college students have the opportunity to conduct self-directed inquiries in their science courses and it is only until graduate school that science majors have this opportunity. The college faculty in some cases bemoaned this fact and expressed that although the OPAPP science learning tasks were not aligned with expectations for college freshman science courses, they wished their own students had the opportunity to engage in inquiry-oriented science.

Conclusion

The first two years of the Ohio Performance Assessment Pilot Project provided important opportunities to evaluate the viability of using performance assessments as one measure of students' college readiness in public high schools. This final section summarizes results of the pilot with regard to the question asked in the title of this paper: *Can performance assessments be reliable and vaild?* The answer to this question is examined in light of the theoretical frameworks discussed earlier in the paper regarding: (A) *scoring models as reflective of task and domain structure; generalizability and the boundaries of score meaning; (B) convergent and discriminant correlations with external variables; and fairness*; and C) *content relevance and representativeness.*

First, ***can performance assessments be reliable? And are they reliable enough for valid use as both summative and formative assessments?*** G-study and D-study analyses of score data from the first two pilot years of OPAPP suggest that teachers can score performance assessments with sufficient levels of reliability when they have received sufficient training and participated in

professional conversations that support their calibration to score a particular task.[16] (Most tasks require the use of two raters to achieve reliability coefficients of .80 or higher on the total score.)

It is likely that a distributed scoring system with blind scoring, like the system used to score the learning tasks for the G-studies and the assessment tasks, better supports reliable scoring. Given limited resources, we chose not to "audit" the local scores submitted by teachers to determine whether scores assigned by teachers to their own students' work were reliable. Instead, we conducted controlled scoring sessions in which teachers were assigned to score pre-selected samples without identifiers and used scores generated during those sessions for reliability analyses. However, an audit of previously submitted "local scores" and samples could easily be done, with additional resources, by re-scoring a small sub-sample of the work and scores submitted by each teacher. If teachers' scores of their own students' work (as in the scoring of the learning tasks) are to be used only for formative purposes, there is little rationale for investing in blind, distributed scoring of the learning tasks. However, to maintain some confidence in the comparability of teachers' scores on the learning tasks, it might make sense to periodically audit teachers' scores by having them select a small random sample of scores and student work samples (3-4 samples across the score range), and submitting them for review. A review panel could be established within each school or district to review teachers' scores on that small sample and provide feedback to teachers about the accuracy of their scores. Teachers could rotate into the panel as part of their annual professional responsibility. [17]

Other Findings from the Generalizability Studies. With regard to implications for valid use of scores, the results of our G-studies for the learning tasks showed that individual

[16]During the pilot, teachers received at least one day of training to score a particular learning task, and an additional day of practice scoring and calibration. Participants who scored the assessment tasks received several hours of training and calibration, and scored during the remainder of the day.

[17]This system has been used successfully in Queensland, Australia, and panelists find their participation to be an important professional development experience.

dimension scores, particularly for the science and ELA tasks, did not achieve sufficient levels of reliability to use dimensional or item scores for any high-stakes purposes. Therefore, only the total or average of scores using the scoring rubrics as designed should be used for making high-stakes summative decisions. In addition, given the desire for teachers to use the dimensional scores for formative purposes, the project would need to consider revising the existing rubrics to strengthen the reliability of the dimensional scores. This could be done by breaking down large dimensions into more discrete items (in the ELA and science rubrics) and then using empirical data about the relationship of items with one another (such as that obtained from factor analyses) to map the items into dimensions.

In the G-studies of assessment tasks, the chemistry and physics assessment tasks were demonstrated to be sufficiently reliable (with minor tweaks in the number of items) to support summative decisions, assuming the use of two raters to score all student samples. On the other hand, the mathematics assessment tasks were not found to be sufficiently reliableto be used for high-stakes summative purposes because of insufficient variation in scores within and across items even with an increase in the number of items and if only a total score is reported. D-studies showed that even with an increase in the number of raters and items, the level of reliability would be insufficient. These results suggest a need to revisit the design of the mathematics assessment tasks and rubrics, but also a need to develop alternative frameworks for evaluating the construct validity of performance tasks, in which the entire task comprises an "item" and the rubric dimensions used to score a single task represent different aspects of performance on a single item (rather than treating each of those rubric dimensions as separate items. For example, to maintain the diagnostic and formative purposes of scoring, performance tasks could be scored using analytic rubrics (with multiple dimensions of performance), but the

scores that are analyzed for psychometric validation could be converted to a holistic score based on total score ranges.

Implications for Revisions. As noted earlier, some revisions to the scoring instruments could improve the reliability of scoring for particular tasks and rubrics. For example, the generic ELA and science inquiry rubrics for the learning tasks could be broken down into a greater number of smaller, more discrete scoring dimensions, allowing for more analytic scoring, reducing the cognitive complexity of each scoring dimension and improving the precision of the score. Increasing the number of scoring dimensions ("items") would also improve overall score reliability. Another strategy for improving the reliability of scoring is to expand the score scale in the ranges where most students are scored (e.g., at the "2" score level on the four-point rubrics). Currently, scores that could be considered a "weak 2", "solid 2" and "strong 2" are all scored at the 2 level. This leaves too much room for interpretation and could lead more lenient rater to rate a "strong 2" performance at the "3" level or a more stringent rater to rate a "weak 2" performance at the "1" level, given that "+", "-", or "0.5" scores are not permitted. Expanding the score scale would lend greater precision to the meaning of the score scale, and allow raters to more precisely assign scores to a given work sample. This would also lead to greater variation in scores and differentiation across students.

Concurrent Validity - Relationship with External Indicators of Academic Performance. In our studies, we examined the relationship of the learning task scores to students' Ohio Graduation Test (OGT) scores, cumulative GPAs, and for a smaller sub-sample ACT scores. We found that in general, the learning task scores had relatively low but significant correlations with cumulative GPAs and almost zero order correlations with OGT scores. This makes sense because the learning tasks were curriculum embedded and were designed to be

completed much like a course assignment, with some support from teachers and students and an opportunity to receive feedback and revise work. These findings suggest that the learning tasks are measuring constructs completely different from that measured by the OGTs and that they also go beyond what course grades capture in measuring students' academic proficiencies.

In relation to ACT scores, the English language arts learning tasks seem to have the strongest relationship to the English, English/Writing, Writing, and Composite scores, but this varied across tasks. In fact, the *Hero's Journey* ELA task, which had been dismissed as being less appropriate for high school level curriculum, had the strongest correlations with ACT scores across the board.[18] The mathematics learning task scores were weakly or moderately correlated with OGT and ACT scores, and the science learning task scores were more strongly correlated with the OGT and ACT *writing*subscores than to the science subscores. This makes sense in that the science learning tasks require writing to explain students' conceptual understandings and communicate how the science inquiry was executed, whereas the OGT and ACT science tests primarily measure students' breadth of science knowledge using a selected response format and almost no writing. Thus, it is not surprising that the correlation between the science inquiry learning task scores and these other indicators of science content knowledge were low.

One important finding that was evident from these correlational analyses (as well as the regression models that followed) was the wide variation across tasks in the magnitude and nature of these relationships. Such variation is a clear indicator of the lack of comparability of tasks. The regression models also indicate that even controlling for prior student achievement and demographics as well as school-level demographics, there were differences in average scores across learning tasks within content fields. While in most cases, these differences were not

[18]*The Hero's Journey* task was also rated the most highly by college faculty as being relevant to entry level expectations for success in college English courses (more discussion on this to be added in the results of the higher education validity review).

significant (likely due to sample sizes and high standard errors), the differences were not close to zero in most cases. This evidence further highlight the limitation of the learning task scores as a way to track student progress over time from task to task and course to course, even when their work is being scored along the same dimensions. If there is a desire for dimensional scores to be reliable enough to track student progress over time on the same dimensions of performance, the comparability of task scores would need to be addressed.

Implications for Revisions. These findings again clearly point to the need for a reconsideration in the ways that both the learning and assessment tasks and scoring rubrics are designed. To return to the question posed in the title of this paper, "Can performance assessments be reliable AND valid?", we must focus on the intended uses of the learning and assessment task scores. If there is an intent to use the students' scores to track progress over time and across tasks, tasks would need to be demonstrated as being comparable in difficulty and cognitive demand.[19] However, this would be difficult to achieve even from a conceptual standpoint, as the content knowledge addressed by a task is an inherent aspect of cognitive challenge. For example, a task in which students demontrate the ability to apply science inquiry methods in a unit on force (physics) might be inherently more complex than a task in which students demonstrate the same skill in a biology class on the topic of osmosis. Task shells and other task design specifications could support the development of more comparable tasks in terms of design, but differences in the ways students interact with tasks ("items") are likely to persist. Prior research (some of which is cited in the theoretical framework section) has shown that students vary in the ways in which they interact with tasks and that the scores they receive on different tasks are not stable. This suggests that scores from one task alone (e.g., one

[19] Task equivalence is a higher standard and would require a research design that is practically unfeasible in most cases.

assessment task) would not be stable enough to be used for high-stakes summative purposes, and that scores from assessment tasks should be combined with other measures (e.g., a series of assessment tasks or even score data from learning tasks) to improve the reliability and validity of consequential decisions.

Content validity. Overall, evidence from the external evaluator's expert review of the ELA, science, and mathematics learning tasks shows in relation to a set of validated standards demonstates a moderate to high level of alignment with the Common Core State Standards, the Ohio science content standards, and the 21st Century Skills defined by the Partnership for 21st Century Skills, though again, this seemed to vary across tasks. In some cases, the OPAPP tasks went beyond expectations of the Common Core (e.g., ELA tasks and rubrics require reflective writing, persuasive writing, and other types of genre-specific writing not included in the Common Core). And in some cases, the OPAPP tasks fell short of some of the standards, particularly in terms of their coverage of content. The variation in alignment with standards across tasks within content fields again highlights the need for a set of tools and protocols that would support a more systematic way of designing tasks and rubrics to ensure greater conceptual comparability.

Lack of content breadth is another drawback of complex performance tasks that is often cited as a validity problem. However, this assumes that performance tasks like the learning and assessment tasks would be used *alone* rather than within a system of multiple measures in which the entire set of performance tasks and other assessment formats (selected response, short constructed response, extended constructed response) covers the breadth of standards. In addition, "matrix sampling" approaches may be used to sample across content and skill targets as well as across students within a school. However, this would mean that individual assessment

components in themselves would not be used in a high stakes summative way and only in combination with other assessment components.

Last, the higher education faculty validity review provided evidence that the OPAPP learning tasks do appear to be aligned with college readiness expectations for entering first-year students in ELA and mathematics, though less so for entry level science courses which focus on face-paced coverage of content. In addition, the kinds of learning and performance expectations embedded in the scoring criteria were affirmed as relevant and important for the kinds of learning and performance expected in entry level college courses. Most of the Level 3 scored student samples that were reviewed were rated as "likely to succeed" in entry level college courses, supporting the notion that these tasks provide evidence of a student's college readiness. Last, most college faculty reviewers seemed to support the idea of having students complete these kinds of performance tasks as important learning experiences and preparation for the kinds of work they would be doing at the college (or graduate school) level.

Overall, the results of these content validity studies (as well as the concurrent validity studies) suggest that the OPAPP learning tasks are aligned to college readiness expectations (as defined by the Common Core State Standards and higher education faculty), and that they go beyond what is currently measured in most student assessment programs (such as the Ohio Graduation Tests, the ACT), as well as by course grades and GPAs. These findings support a validity argument in favor of including performance based assessments in the mix of assessment formats used to evaluate student achievement and progress by expanding the scope of learning that can be assessed, allowing for a greater focus on discipline-specific skills and work habits that have previously been unmeasurable using standardized tests alone, and providing a set of tools for formative use that can have immediate impacts on curriculum and instruction.

References

Brennan, R.L. (1993). *Elements of generalizability theory*. Iowa City, IA: ACT Publications.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp.443-507). Washington, DC: American Council on Education.

Cronbach, L.J., Gleser, G.C., Nanda, H., &Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

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.

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

Herman, J. L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum.

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.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1990, October). What alternative assessments look like in science. Paper presented at Office of Educational Research and Improvement Conference, The Promise and Peril of Alternative Assessment, Washington, DC.

Shavelson, R.J., Ruiz-Primo, M.A., & Wiley, E.W. (1999). Notes on sources of sampling variability in science performance assessments.*Journal of Educational Measurement, 36*(1), 61-71.

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

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 OhioPerformance 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.

Constructing the Self

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

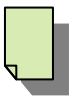
The goal of this last assignment is to help you synthesize the work you did in the two previous tasks: to make comparisons and judgments about the visions about identity set forth in those texts. Working from the idea that "texts teach," imagine that each one of the texts you selected in Part 2 contain 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

Open for Business — Rubric

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

		points	section points
1.	***Mathematics of the Task***		
	Determines the linear relationship of demand and price for Laptop, Xbox and Ipod.		
	Laptop: **q = (-3/50)p + 69** Xbox**: q = (-12/25)p + 185** Ipod: **q = (-8/25)p + 99**	5	
	Stereo: **q =(-14/19)p +**132 Calculator: **q = (-14/10)p + 150**		
	Partial Credit: 1 to 4 errors	(3)	
	Partial Credit: 3 slopes correct	(1)	
	Finds the optimum number, selling price & maximum profit for Laptop, Xbox & Ipod.		
	Laptop: 13 / $933 / $2,592	3x1	
	Xbox: 30 / $323 / $1,810	3x1	
	Ipod: 24 / $234 / $1,797	3x1	
	Shows that the Stereo and Calculator will produce smaller profits.	1	
	Determines she can sell **13** laptops, **30** XBoxes and **8** iPods to raise $5,000	2	17
2.	***Mathematical Reasoning***		
	Derives a linear relationship between *demand* and *price*. (q = mp+b)	2	
	Partial Credit: correctly derives the slope.	(1)	
	Uses the relationship that *revenue* equals *demand* times *price* (r=qp).	1	
	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).	2	
	The amount Malena earns by selling products at different prices, accounting for the percentage taken by eBay (accept earning equation or eBay overhead expression).	2	
	Construct graphs of the earnings function	2	
	Interprets optimum number sold as a whole number	1	
	Reasons how to optimize earnings to raise $5,000	1ft	11
3.	***Approach***		
	Maximizing Equation method: Finds an equation that maximizes earnings. Uses a method (perhaps derivative or vertex of the parabola) to determine the maximum.	5	
	Partial Credit		
	Graph method: Determine relationship of price to earnings taking into account the demand, revenue and cost. Creates a graph to estimate the maximum earnings. (trace function)	(4)	
	Table, spreadsheet or a repeated calculation method	(3)	
	Errors: Uses either approach but misses a significant step or steps in the process.	(2)	5
4.	***Communication***		
	Clear explanation of the process	2	
	Shows table(s)	1	
	Shows correct profit or earning equation(s)	1	
	Shows graph(s) to illustrate earnings	1	
	Summary explanation	2	
	Partial credit: maximum earnings that sum close to $5,000 or maximum of top 3	(1)ft	7
	Total Points		**40**

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: Research and prepare an introduction to your lab report sharing what you have learned about acids and bases and the process of neutralization. (**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

Version 3.0 June 1, 2009

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

Version 3.0 June 1, 2009