Assessing What Matters: Literacy Design Collaborative (LDC) Writing Tasks as Measures of Student Learning

Ruth Chung Wei, Ph.D.
Stanford University
rchung@stanford.edu

Kenneth Cor, Ph.D.
University of Alberta, Canada
mcor@ualberta.ca

This report is based on a paper presented at the
2015 Annual Meeting of the American Education Research Association.

July 2015

SCALE
Stanford Center for Assessment, Learning, & Equity

Stanford
GRADUATE SCHOOL OF
EDUCATION

Ruth Chung Wei is Director of Assessment Research & Design at the Stanford Center for Assessment, Learning and Equity (SCALE), Stanford University Graduate School of Education

Kenneth Cor is Director of Assessment, Faculty of Pharmacy and Pharmaceutical Sciences
Edmonton Clinic Health Academy, University of Alberta

The **Stanford Center for Assessment, Learning, and Equity (SCALE)** is a research and development center that provides technical consulting and support to schools and districts that have committed to adopting performance-based assessment as part of a multiple-measures system for evaluating student learning and measuring school performance. SCALE's mission is to improve instruction and learning through the design and development of innovative, educative, state-of-the-art performance assessments and by building the capacity of schools to use these assessments in thoughtful ways, to promote student, teacher, and organizational learning.

http://scale.stanford.edu

This research was supported by the Bill and Melinda Gates Foundation.

Special thanks to Claudia Long for her editorial feedback.

Abstract

The disconnect between large-scale external assessments and the enacted curriculum is often minimized as schools and teachers are ranked, punished, and rewarded based on accountability policies that depend heavily on these external assessments. This paper examines results from a 2012-2013 study of instructionally-embedded performance assessments, in this case, Literacy Design Collaborative (LDC) writing tasks. Instructionally-embedded performance assessments are embedded within units of study, frequently assigned as summative assessments at the end of a unit. They are more directly aligned with the taught curricula then external assessments. We ask: Would an assessment system that includes both external measures and instructionally-embedded measures provide a set of measures that is more valid and balanced by taking into account multiple and varied kinds of information about student learning? We explore the technical quality and validity of instructionally-embedded performance assessments, specifically Common-Core aligned, text-dependent writing tasks that were developed by teachers using the LDC templates for designing writing tasks. We examine the relationships of scores on these performance tasks with some on-demand measures of student learning in similar domains, including standardized external assessments, and explore the technical merits of the performance tasks for measuring student learning. Last, we make a validity argument about the potential uses of instructionally-embedded writing tasks and the interpretation of scores generated by those tasks.

Assessing What Matters:
Literacy Design Collaborative (LDC) Writing Tasks as Measures of Student Learning

Introduction

In the last decade, large-scale standardized assessments have been the primary measure used by state accountability policies for schools and teachers. In some cases, such external assessments have driven changes in instruction that focus on "test prep" activities, putting an over-emphasis on tested matter (i.e., language arts and mathematics) over a more well-rounded curriculum that includes science, social studies, art, and music. In other cases, these external assessments are only marginally related to the kinds of curriculum and learning opportunities offered to students, and are largely disconnected from the kinds of assessments regularly administered by teachers in their classrooms. This disconnect between the external, standardized assessments and the enacted curriculum is often minimized, even though schools and teachers are ranked, punished, and rewarded based on accountability policies that depend heavily on these external assessments. As a result, these accountability policies have created an environment of fear among educators, a loss of professional autonomy, and a distrust of external assessments in general.

This paper examines results from a 2012-2013 study of instructionally-embedded performance assessments, specifically those offered by the Literacy Design Collaborative (LDC), a Common Core implementation initiative. Instructionally-embedded assessments are classroom assessments that are typically administered as end-of-unit or culminating assessments. They follow a series of learning activities and lessons – referred to as "instructional ladders" – designed to scaffold and support student learning and success on the assessments. The LDC gives teachers the opportunity to use Common Core standards-aligned templates to design their own text-dependent performance tasks in the form of argumentative or explanatory writing tasks along with task-specific "instructional ladders" that scaffold students' ability to access texts and engage them in the writing process.[1]

Specifically, the research study examines the following two questions as they pertain to the LDC tasks:

1. What is the relationship between instructionally-embedded performance assessments and external on-demand measures of student learning in terms of student performance and the information provided about student competencies?
2. How well do instructionally-embedded performance assessments meet standards of technical quality (validity, comparability, reliability)?

The ultimate purpose of our study was to investigate the possibility of using instructionally-embedded performance assessments to measure student learning and growth. If fair and valid assessments are those that are both aligned to a state's adopted standards and reflective of students' opportunities to learn, it is critical that the assessments used for accountability purposes be aligned to both the curriculum and the instruction to which students have had access. This study raised the question of whether an assessment system that includes both external on-demand assessments and instructionally-embedded assessments would be a fairer and more valid assessment system for students and teachers.

Background

Standardized selected-response tests designed and implemented by large-scale testing companies have long been criticized as being too disconnected from teachers' curriculum and instruction, and consequently, student learning (see e.g., Shepard, 2000; Shepard, 2003; Wiggins, 1990). While large-scale standardized tests are aligned to state standards and provide reliable, comparable information about student achievement relative to a specific standard of proficiency, results of these tests are less useful for classroom teachers because they usually come long after students have left a teacher's class. In addition, although the machine-scored

[1] See www.ldc.org for more information on writing task tools and templates provided by LDC. An entire LDC "module" is composed of both a writing task ("teaching task") and an "instructional ladder," a sequence of formative assessments ("mini-tasks") and instructional strategies designed to scaffold students' ability to access the text(s) and complete the writing task.

selected-response (multiple-choice) items that make up the bulk of these tests can assess a breadth of students' knowledge, they are limited in terms of their ability to assess student's depth of knowledge in any one content area.

On the other hand, teachers have been designing their own classroom assessments for generations, with test content closely tied to the taught curriculum. These instructionally-embedded assessments are not only more directly related to the taught curriculum, they also generate more timely information that can help teachers make immediate instructional adjustments and provide feedback that can be used by students to monitor their own learning. The instructionally-embedded assessments are also more formative in nature. They provide opportunities for "assessment for learning" (Black & Wiliam, 1998; Stiggins, 2002) in that they are often scaffolded through a series of learning tasks or activities that help students develop their responses to an assignment, with opportunities for feedback and revision. **Table 1** below displays the key differences between instructionally-embedded and external, standardized assessments.

Table 1
Comparison of Instructionally-Embedded Assessments versus External, Standardized Assessments

Instructionally-Embedded Assessments	External, Standardized Assessments
Closely tied to the taught curriculum and units of study	Loose or assumed connection to taught curriculum (through state standards)
Teacher-developed or teacher-selected	Externally developed and validated
Customized to school/classroom context	Standardized across schools
Scaffolded, with opportunities for feedback and revision	On-demand, no opportunities for feedback or revision
Provides immediate, targeted information and feedback	Provides information annually, on broad level goals

In addition, instructionally-embedded assessments often are, or include, *performance* assessments. This type of assessment is able to assess more complex and higher-order thinking skills critical to college and career readiness that are generally not assessable using selected-response items (Darling-Hammond & Adamson, 2010; Lane, 2010). In this paper, we define performance assessments as tasks that ask students to construct a product or execute a performance that demonstrates application of knowledge, understandings, and skills through work authentic to the discipline and/or real world. The LDC writing tasks that were examined in this study are performance assessments that require students to authentically demonstrate key college and career readiness competencies – e.g., the ability to analyze and synthesize textual sources, develop and support their own ideas using textual evidence, organize their ideas into a coherent writing product, and revise and refine their writing over multiple drafts. Such competencies cannot be adequately assessed using multiple-choice or short-answer item formats.

Although instructionally-embedded assessments designed by teachers have been part of classrooms for decades, they have also been criticized as being too idiosyncratic, unreliable, and lacking comparability across teachers to be useful as credible measures of student competencies. Even the statewide performance assessment initiatives of the 1990s in which teachers designed their own tasks (e.g., the Vermont portfolio, the Kentucky portfolio system) were criticized for the lack of comparability among the teacher-designed tasks. This issue, among others, weakened the validity argument for these systems, and contributed to their discontinuation (Koretz, 1998; McDonnell, 2004; Stecher, 1998). These concerns about assessment validity, comparability, and technical quality continue to be raised in other more recent efforts to validate locally designed assessments (e.g., Rhode Island's Graduate Diploma System, Wyoming's Body of Evidence system, Student Learning Objectives-SLO assessments), and the scrutiny is intensified when the assessments are used to evaluate individual students' achievement or to evaluate teachers (CEP, 2011; NRC, 2003; Goe and Holdheide, 2011).

The Literacy Design Collaborative has sought to overcome some of these challenges with instructionally-embedded assessments by supporting teachers' design of assessments in two ways: 1) providing teachers with templates and technological tools to build reading and writing performance tasks that are text-dependent and aligned to writing expectations in the Common Core State Standards; and 2) providing teachers with a bank of "juried" writing tasks that have undergone a peer review process by a national cadre of "jurors" trained to evaluate the writing tasks and instructional modules using a common set of criteria. These juried writing tasks serve as models for teachers as they design their own writing tasks. Since 2013, LDC has made the templates, technological tools, juried tasks, and jurying criteria freely available to all educators. The LDC also regularly provides "jurying training" at national and regional events, with the goal of building a national consensus around "high quality" writing tasks and instructional modules among practitioners who are using the LDC templates and coaches who are teaching practitioners to use the tools.

What does it mean to be a valid and reliable assessment?

As noted earlier, both validity and reliability are two important technical criteria for determining the fairness and usefulness of an assessment. Cronbach (1971) argues that validity has to do with the meaning or interpretation of scores, as well as consequences of score interpretation (for different persons or population groups, and across settings or contexts). Therefore, in this study, score validity is examined in light of two primary contexts for assessment use: 1) for summative assessment (i.e., as a potential measure of student progress over time) and 2) for formative assessment (i.e., as a driver of instruction and student learning).

Further, according to the most recent conceptions of validity, validation involves both an interpretive argument that specifies the proposed and intended uses of test scores and 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).

The criteria used to evaluate the validity of a performance assessment vary. For example, Messick's (1989) evaluation criteria include content, substantive, structural, generalizability, external, and consequential aspects of validity, while Linn, Baker, and Dunbar's (1991) evaluation criteria include consequences, fairness, transfer and generalizability, cognitive complexity, content quality, and content coverage.

In this study, we focus on the following criteria for supporting a validity argument, based on the data available to us: 1) structural criteria, that is, the scoring models (e.g., the LDC writing rubrics) as reflective of task and domain structure; 2) generalizability and the boundaries of score meaning; 3) external criteria – in this case convergent and discriminant correlations with external variables; and 4) cognitive complexity and content quality (through a review of writing tasks submitted by teachers in our study sample). We also have available to us, through evaluation studies conducted by Research For Action and the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA, information about how teachers use the LDC tools to guide their instructional planning, an aspect of consequential validity.

Reliability is the other factor that must be considered when evaluating the technical quality of performance assessments. It has been said that you cannot have validity without reliability. The traditional approaches to evaluating reliability for selected-response items, which are machine-scored as correct or incorrect, have not worked well for performance assessments, which are typically scored by human raters using rubrics that assess the quality of a student's responses. Although some alternative frameworks for evaluating reliability in performance assessments have been proposed (see for example Moss, 1994), there is a policy imperative to quantify the reliability of scoring when performance assessment scores are used for consequential decisions, high-stakes or otherwise. Therefore, in this study, we drew 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. (See Brennan, 2001, and Cronbach, Gleser, Nanda, & Rajaratnam, 1972, for more detail on Generalizability theory.) The results of these analyses are critical to understanding the design of the scoring rubrics, the construct validity of the scoring dimensions, and for inferring how the performance task scores may be used validly as measures of student achievement and progress.

Methodology

Although our sample size turned out to be small, we used multiple sources of data in our study, including student scores from on-demand performance assessments, instructionally-embedded performance assessments (LDC writing tasks), and external (state) standardized assessments. We also used multiple analysis strategies to examine that data. These data sources and analysis strategies are described in more detail below.

Data Sources. To address the first research question with reference to LDC tasks – i.e., What is the relationship between instructionally-embedded performance assessments and external on-demand measures of student learning in terms of student performance and the information provided about student competencies? – SCALE requested from one LDC research site (through a previously established agreement with that site) local administrative data on students taught by LDC-implementing teachers in the seven small districts within that site. The site was selected based on a convenience sampling strategy. The site had been implementing LDC for at least one year prior to the study and had a significant number of teachers involved in LDC. These seven districts (and the state) had also made commitments to continue to implement LDC through foundation grant funding. (Because this site was composed of seven small districts, seven separate data requests were made.) We worked with an administrator at the site to recruit participants for the 2012-13 academic year. Given that LDC is meant to be used for teaching writing across the curriculum, our target goal was to recruit at the site a total of at least 30 teachers at the high school level, including 10 ELA teachers, 10 history teachers, and 10 science teachers. Per IRB regulations, we required consent from the teachers and their students (and guardians). Participating teachers received a small stipend and were asked to select one class and collect and submit the following data from that class:

- Student responses on on-demand course-specific Pre- and Post-Assessments that the teacher administers to the class
- Student responses (essays) that the students complete as part of an instructionally-embedded LDC writing task designed by their teacher, and that the teacher administers in spring 2013
- The student roster for the class completing the assessments (Pre and Post, and LDC writing task) with State Student IDs (which would allow us to link the student scores with administrative data on students)

We worked with a testing company, Measured Progress, to select and administer during the school year the above noted on-demand pre- and post-assessments in four different areas – ELA, Science, World History and United States History – to match the courses taught by teachers in our sample. The pre- and post-assessments were identical (the same assessment was administered during an administration window near the beginning of a course and again at the conclusion of the course). These assessments, which had previously been developed by CRESST for a different project administered by Measured Progress, included a combination of selected-response items that were machine-scored; constructed-response items that were hand-scored; and an essay component that was hand-scored using a writing rubric. (See **Table A1** in the Appendix for the data collection timeline for the study).

In addition to the above data, we requested from the seven districts the following information about students participating in the study: demographic information (e.g., race/ethnicity, gender, economic disadvantage status, language status, special education status, attendance) and concurrent and prior achievement scores going back as far as possible (to third grade if available). In particular, we asked for scores on the State Literature Exam (these scores came from students in the classes of the participating English Language Arts teachers only) and scores on the state grade-level (gr. 3-8) exams in reading and writing). We also asked for SAT/ACT test scores, which were only occasionally available.

The resulting sample of teachers who participated in the study and completed all data collection components is displayed in **Table 2** below.

Table 2
Number of Teachers Who Completed All Data Collection Components

	Completion of All Data Collection
ELA	12
Biology/Science	6
U.S. History	5
World History	7
Total	**30**

As indicated by the table, in this LDC site the total number of teachers who completed all portions of data collection was 30 teachers, which met our target goal. However, the target goal of having 10 teachers each in ELA, science, and history was not met due to insufficient numbers of science and history teachers implementing LDC within the research site. Also, because not all courses have corresponding end-of-course exams, and because of gaps in current year and prior year state test scores, even with a low attrition rate of teachers from the study, the sample sizes of the student data that could be used were reduced significantly. In the end, we relied on the data generated by the students of only the English Language Arts teachers in our study. A subset of those students took an end-of-course "State Literature Exam" in that same year (typically taken in 10th grade, with multiple opportunities to take and pass the exam through 12th grade), completed the on-demand Pre- and Post-Assessments (with a writing component) administered by their teachers, and submitted LDC essays written in response to LDC writing tasks developed by their own teachers. The data for this subset of students was used in this study. See **Table 3** below for descriptive statistics on these measures for students of the ELA teachers in the study.

Table 3
Student Descriptive Statistics – LDC Study, Students of English Language Arts Teachers*

	Students with ELA Performance Assessment Data				
	Mean	SD	Min	Max	N
Items from Administrative Data					
2013 State Literature Exam Scaled Score	1536.648	61.781	1410	1770	105
Pupil Demographics					
White	83.61%				299
Black	5.02%				299
Hispanic	7.36%				299
Asian	2.68%				299
Other Race	1.34%				299
Female	52.67%				300
Economically Disadvantaged	29.43%				299
ELL	2.01%				299
Special Education	11.16%				233
Gifted	7.69%				299
Items from Performance Assessment Data					
Pre-Assessment Essay Score	7.130	2.776	5	17	376
Pre-Assessment Total Score	14.440	5.285	0	28	393
Post-Assessment Essay Score	7.359	2.917	5	17	343
Post-Assessment Total Score	15.473	5.308	2	29	351
LDC Essay Total Score	14.891	5.979	7	28	290

*Only data generated by students of English Language Arts teachers were used in the study.

Measured Progress collected and scored all of the on-demand Pre- and Post-Assessments, as well as the LDC Essays. They provided SCALE with all of the student score data with student identifiers, which allowed us to match the performance assessment scores with district administrative data about the students' demographics, prior test scores, and concurrent year test scores.

Other factors that further limited our sample size were the lack of availability of Grade 3-8 state test scores for students with performance assessment data, mismatches in IDs provided by the teachers and the districts, and missing scores for many students on the Pre- and Post-Assessments and LDC writing tasks. In the end, the regression models rely on a small sample, ranging from 66 students to 98 students. These missing data and small sample sizes may bias the data and results of this study.

To address the second research question relevant to LDC tasks – How well do instructionally-embedded performance assessments meet standards of technical quality (comparability and reliability)? – we gathered data from two separate sources:

1) To evaluate LDC task comparability, we collected all of the LDC writing tasks that had been designed and implemented by the participating teachers in the study, trained external raters to rate the LDC writing tasks using the LDC Task Jurying Rubric, and had them provide ratings on those tasks. We had all trained raters (at least three for each subject area) score all of the submitted writing tasks in order to get a reliable estimate of each Teaching Task score, conducting back-reads by lead trainers to adjudicate conflicting scores, and examined the ratings of the writing tasks to evaluate their comparability.

2) To evaluate the reliability of the LDC scoring rubrics, we hired Measured Progress raters (those who had scored all of the student essays that were submitted by participating teachers) to re-score a set of 20 randomly selected student samples from two different ELA modules – one that uses an Argumentation Writing Rubric and the other an Informational/Explanatory Writing Rubric. We conducted G-studies (Generalizability analyses) on both sets of data to examine the reliability of the scores from the two rubrics and sources of error variance due to raters, the tasks, and the students.

Data Analysis Strategies. To address the question of the relationship between student performance (scores) on instructionally-embedded assessments and student performance on other assessments such as large-scale external assessments and on-demand assessments, we used multiple regression analyses to examine how students' prior achievement data, their on-demand Pre- and Post-Assessment scores, and their LDC Essay scores students are related to concurrent achievement data (in this case, 2013 State Literature Exam scores) and how these measures were related to each other. See Appendix, page 42 for a description of the regression models.

To address the question of comparability and reliability, we conducted two analyses: 1) we evaluated the comparability of the LDC writing tasks by examining the descriptive statistics (averages, frequency, range) for the ratings of the LDC tasks using the LDC Task Jurying Rubric; and 2) we examined the reliability of the LDC task scores by using a G-study methodology (as noted above) to examine the reliability coefficients for the LDC writing rubrics as a whole and for each scoring dimension, as well as the sources of error variance from raters, the tasks, and the students. (Two G-studies were conducted, one on the Argumentation Writing Rubric and one on the Informational/Explanatory Writing Rubric, which are nearly identical with minor variations.) The results of the G-studies were used to evaluate the reliability of raters, the reliability and usefulness of the rubric dimensions, and the usefulness of each task for capturing student variation in scores. (For more specific information about the technical aspects of both of these analyses see Appendix, page 51.)

Results and Discussion

This section describes several findings yielded from our analyses that pertain to the two research questions we initially asked. This section also discusses implications of these results for understanding the potential use of LDC writing tasks as measures.

***Finding 1**: The relationship between student scores on the LDC essay (an instructionally-embedded assessment) and student scores on an external state assessment (the State Literature Exam), taken during the same school term, is relatively weak in comparison to other on-demand assessments and prior measures of reading achievement. In addition,* *LDC essay scores do not have any relationship with prior state measures of reading and writing achievement. This suggests that instructionally-embedded performance assessments like the LDC writing tasks can provide different kinds of information about student learning and performance that is not measured on external state assessments.*

As described above, our analyses compared students' LDC essay scores and their on-demand Pre- and Post-assessment scores with their State Literature Exam scores and their prior state grade-level exam scores (in reading and writing). Based on these analyses, the LDC writing task scores have a weak relationship with the State Literature Exam scores. The state's *reading* scores for Grade 8 turned out to be the strongest predictor of students' performance on the secondary State Literature Exam, but the on-demand Pre- and Post-Assessment scores and the LDC essay scores were stronger predictors of student performance on the State Literature Exam than the state's *writing* scores for Grade 8 (which had no statistically significant relationship with the State Literature Exam scores).[2] (A detailed, technical explanation of these findings is found in the Appendix, beginning on page 42.)

This finding is reinforced when we use the performance measures as outcomes, with prior state test scores as the predictors. The state *reading* scores for Grade 8 were a moderately strong predictor of student performance on the on-demand Pre- and Post-Assessments that teachers administered at the high school level. The state *writing* scores for Grade 8 were a weak but significant and positive predictor of the Post-Assessment essay component score for students in ELA classes. However, neither the Grade 8 Reading scores nor the Grade 8 Writing scores were significant predictors of the students' LDC Essay scores. A technical explanation of these findings is found in the Appendix, page 47.

One hypothesis for why LDC essay scores have a low, non-significant relationship with the Grade 8 Reading/Writing scores and a weak though statistically significant relationship with the State Literature Exam is that there is a more limited range of scores on the LDC essays and the

[2] However, low regression coefficients for the performance measures and only incremental increases in the explanatory value of the models (r-squared values) in which the performance measures are added as predictive variables suggest that the performance-based measures seem to be measuring something different from the State Literature Exam.

average essay score skewed toward the lower end of the score scale. Typically, when teachers first begin using complex performance tasks in which students are required to write a sustained, coherent response (i.e., an essay), scores are initially depressed toward the lower end of the scale. The LDC rubric has seven score levels: 1.0 ("Not Yet"), 1.5, 2.0 ("Approaching Expectations"), 2.5, 3.0 ("Meets Expectations), 3.5, and 4.0 ("Advanced"). When we examined the LDC Essay score distribution for the students in our sample (including all scores, not just those matched to standardized test scores), we found that about 64% of students had an average total score (based on seven dimension scores) below 2.5; about 15% of students had an average score between 2.5 and 3.0 ("Meets Expectations"), and only about 22% of students had an average score of "Meets Expectations" or above (average scores at 3.0 – 4.0). See **Figure 2** below for the LDC essay score distributions (ELA only).

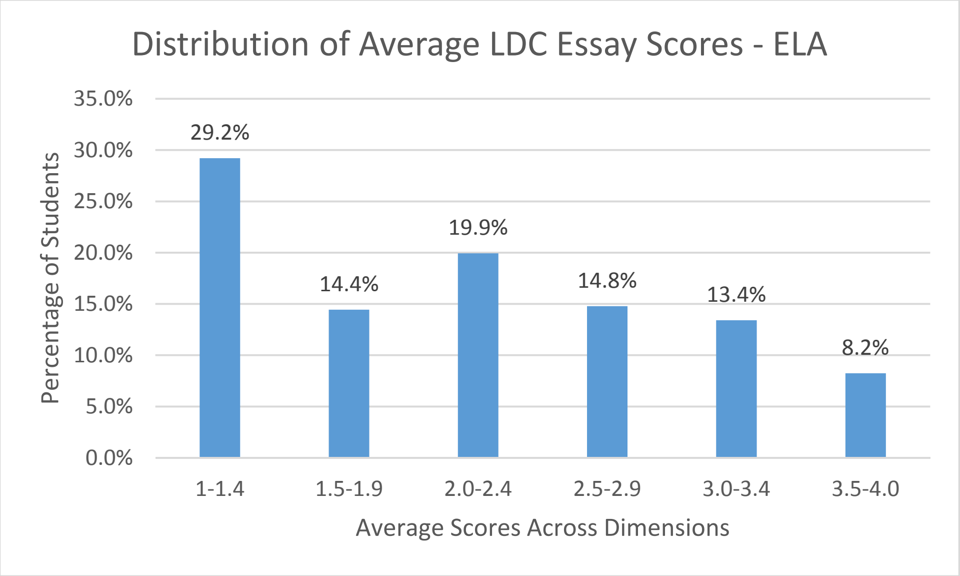


***Figure 2.* Distribution of Average LDC Essay Scores (Average of 7 Dimension Scores)**
Note: Each dimension is scored on a 4-point scale where 1=Not Yet, 2=Approaches Expectations, 3=Meets Expectations, 4=Advanced

In contrast, we see that the distribution of scaled scores on the State Literature Exam for the Spring 2013 administration (using all student data provided by the 7 districts, not just those matched with the LDC essay scores) fits a normal distribution pattern (see **Figure 3** below), and when the scaled scores are converted into the performance levels as determined by the state's standard setting panel, we find that the majority of students reach the "Proficient" level or higher (see **Figure 4** below). In fact, the distribution of scaled scores is skewed toward the upper end of the performance levels.

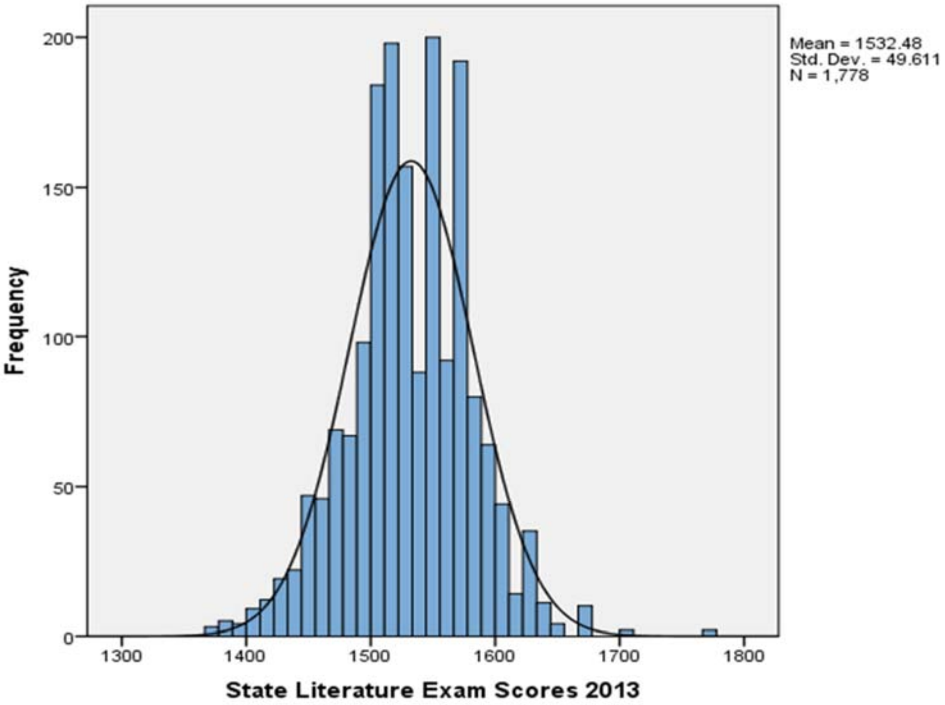


Figure 3. **Distribution of Scaled Scores on the State Literature Exam 2013**

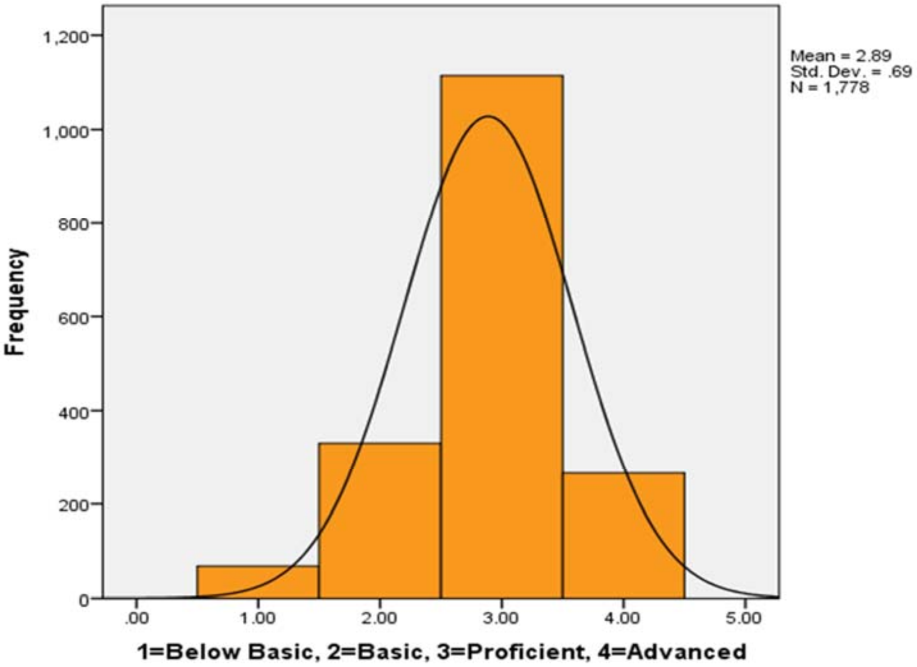
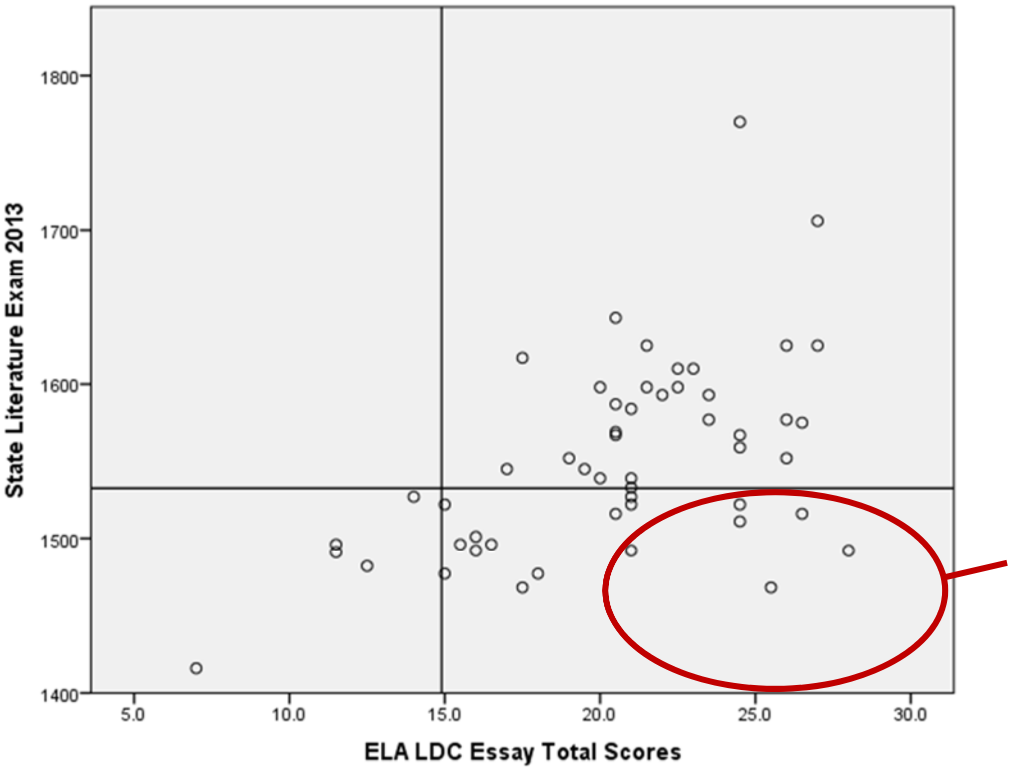


Figure 4. **Distribution of Performance Levels on State Literature Exam 2013**

These varying score distributions suggest that in comparison to the State Literature Exam, which consists primarily of multiple-choice items, with a few constructed-response questions, the LDC writing tasks are relatively more challenging and result in lower average scores. This may change as teachers and students gain more experience with performance tasks that require sustained writing using textual sources. As many states are now beginning to use state assessments that include a limited number of performance tasks (e.g., Smarter Balanced and PARCC), the alignment of the external state tests with the LDC Essay scores may improve, resulting in stronger predictive relationships between LDC Essay scores and the external state tests. Further research should be conducted in the immediate next year to investigate the relationships between the new state test scores and LDC Essay scores to determine whether LDC Essay scores provide stronger predictive information when the external assessments are more consistent in terms of the performance expectations measured.

Finding 2: *A small percentage of students in the sample who scored below average on the State Literature Exam performed well on the LDC writing tasks (scoring Meets Expectations or higher on the LDC writing rubric). This suggests that instructionally-embedded performance assessments, such as the LDC writing tasks, can potentially provide students with opportunities to demonstrate understandings and skills that are not captured adequately by external standardized state tests.*

The weak relationship described above (in Finding 1) between student performance on the LDC tasks and student performance on external assessments suggests that instructionally-embedded performance assessments like the LDC writing tasks may provide different kinds of information about student learning and performance that is not measured on external standardized state assessments. This is further supported when we closely examine the relationship between the State Literature Exam scores and the LDC essay scores (the sum of seven dimensional scores, with each dimension being scored on a 4-point scale) using a scatter plot (see **Figure 5** below). We find that while there does appear to be a generally linear relationship between the two variables, there is also wide variation in State Literature Exam scores at some of the LDC rubric score levels (e.g., see the range of scores on the State Literature Exam, represented on the vertical axis, at the "Meets Expectations" level – represented by an average total score of 21 on the horizontal axis – and higher). Students who are low-scoring on the LDC rubric score scale are also low-scoring on the State Literature Exam score scale. This suggests that low LDC rubric scores are likely to be fairly accurate in estimating the literacy (reading and writing) skills of students because they are confirmed by external standardized test scores. This is less true at the upper end of the LDC score scale. In the limited sample of students for whom we have both LDC essay scores and State Literature Exam scores, we see that there are some students who performed below average (Mean: 1532) on the State Literature Exam, but had relatively higher scores on the LDC rubric. (In **Figure 5** below, those score cases are circled.) This suggests that students who may not otherwise demonstrate strong performance on a state standardized test (i.e., they have below average scores) may demonstrate higher proficiency on instructionally-embedded assessments such as the LDC writing tasks. These score relationships should be further examined through qualitative study of students who perform differentially across these two measures. Also, the number of score pairs represented in **Figure 5** is only 52 cases, limiting the robustness of these findings.



Students who score below average on the State Literature Exam (1532) but score relatively higher on the LDC Essay ("Meets Expectations" or "Advanced" scores)

Figure 5. **Scatter Plot of Literature Exam Scores and LDC Essay Total Scores**
(Note: Total cases: 52. Average scores on State Literature Exam in this sample, represented by the horizontal line, is 1532. Average LDC Essay Total Score in this sample, represented by the vertical line, is 14.9)

These analyses were replicated with additional student score pairs, including the scores for students who completed LDC writing tasks in world history.[3] (See **Figure A1** and an explanation in the Appendix.)

Finding 3: *There is a moderate correlation between students' performance on the LDC essay and their performance on the on-demand Pre- and Post-assessments (which include an essay component). Almost 47 percent of students completing the on-demand*

[3] The sample of score pairs for students who completed science and U.S. history LDC writing tasks and also had scores on the State Literature Exam from that year was insufficient.

Post-Assessment essay scored the lowest possible score (5 out of a possible score of 20), while among those same students, there was a wider range of LDC essay scores. This suggests that instructionally-embedded performance assessments such as the LDC writing tasks may provide students with opportunities to demonstrate understandings and skills that standardized on-demand performance assessments do not.

Our study found that the correlation between students' LDC essay score and their on-demand Post-Assessment essay score is relatively low (0.49). (See **Table A5** in the Appendix.) This suggests that scores from standardized on-demand performance assessments (in writing) provide information about students' literacy and writing skills that is different from the information provided by scores on instructionally-embedded writing tasks. This makes sense, given that the standardized tests are given under on-demand conditions, while the LDC essays are completed in and out of the classroom, with teacher scaffolding and support. These results highlight a key difference between on-demand performance assessments, which are administered under standardized conditions, and instructionally-embedded assessments (LDC writing tasks), which provide opportunities for students to discuss their ideas with their peers, generate drafts, receive feedback, and polish their work. While some testing experts may argue that this use of performance assessment *for learning* reduces the validity of a piece of writing as a representation of students' understandings and skills, the skill of developing a piece of writing and using feedback to improve one's writing is a standards-aligned skill and one that is valued as a "college readiness" attribute.

A closer look at the distribution of essay scores for the on-demand ELA Post-Assessment essay (see **Figure 7** below) indicates that 46.9% of ELA students with Post-Assessment essay scores got the lowest possible score on the essay (scoring an average of 1.0 on a 4-point rubric across five dimensions of performance). It is not clear why such a large percentage of students completing the ELA Post-assessment essay received the lowest possible score. It is possible that many students felt little to no motivation for completing a test for a research study that had no stakes for them, and so they put little effort into their essays. It is also possible that many students did not even enter a response. Therefore, the trustworthiness of the scores on the ELA Post-Assessment writing task is limited.

The distribution of scores on the ELA Post-Assessment essays is skewed lower on the score scale, with an average score of 1.47 on a 4-point scale (average of four dimension scores). The distribution of scores on the ELA LDC essays (see **Figure 8** below) is also skewed toward to the lower end of the score scale, but the average score is higher – 2.13 on a 4-point scale.

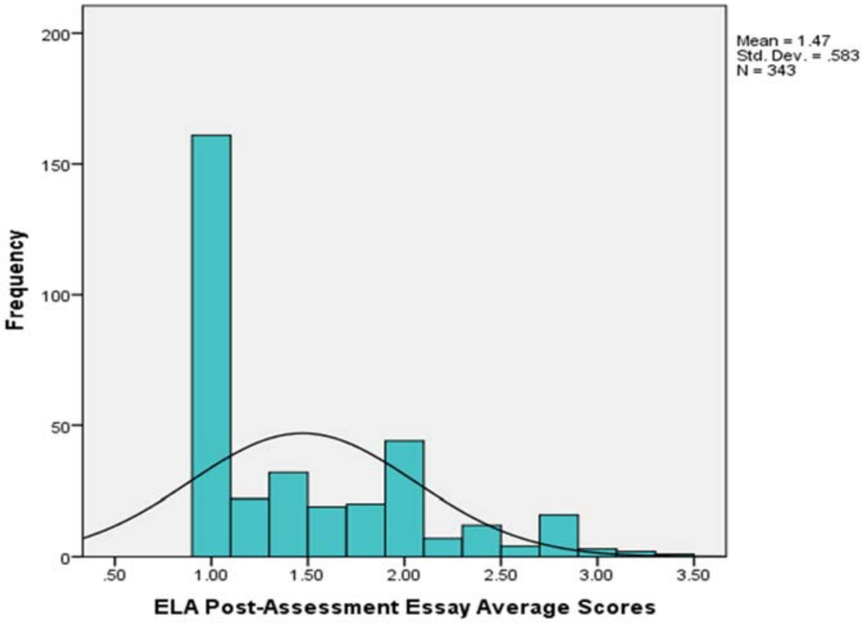


Figure 7. **Distribution of Average Scores on ELA Post-Assessment Essay (On Demand)**

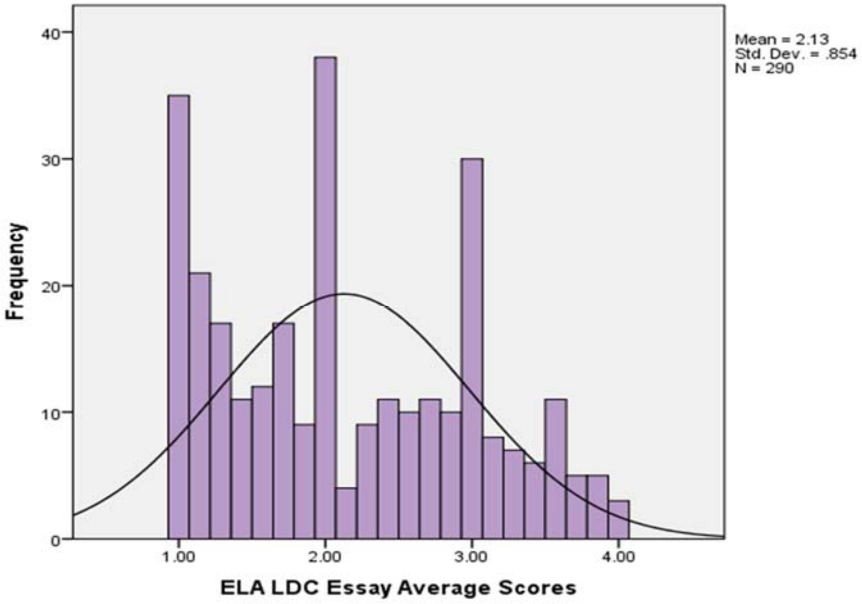


Figure 8. **Distribution of Average Scores on ELA LDC Essays**

With the limitations cited above, due to the low-stakes nature of the Post-Assessment essay for students and the possibility of high rates of non-responses, it is difficult to draw robust conclusions. However, students often find writing essays on an on-demand test to be challenging, and may even feel so intimidated by them that they cannot start writing. In other studies in which we have examined the performance of students on on-demand performance assessments we have found the same problem – students either do not start the assessment writing task, or, at best, they write one or two sentences as a response, resulting in scores of N/S (Not Scorable) or the lowest possible score. In these cases, the scores provide poor evidence for making inferences about what students know or can do, or for diagnosing students' learning needs.

In this sense, for those students who might otherwise perform poorly on an on-demand performance assessment (like those found in state assessments), instructionally-embedded performance assessments like the LDC tasks may provide students an opportunity to demonstrate greater levels of proficiency on the assessment targets because of the tasks' scaffolding and other supports that help students both gain entry into the performance assessment and persist to complete it. **Figure 9** below shows that when we select *only* the students who received the lowest possible score on the ELA Post-Assessment essay (average of 1.0 on a 4-point scale), the distribution of the same students' scores on the LDC essays varied widely on the score scale (also a 4-point scale), though a significant number of them (about 20%) also received an average score of 1.0 on the LDC writing rubric.

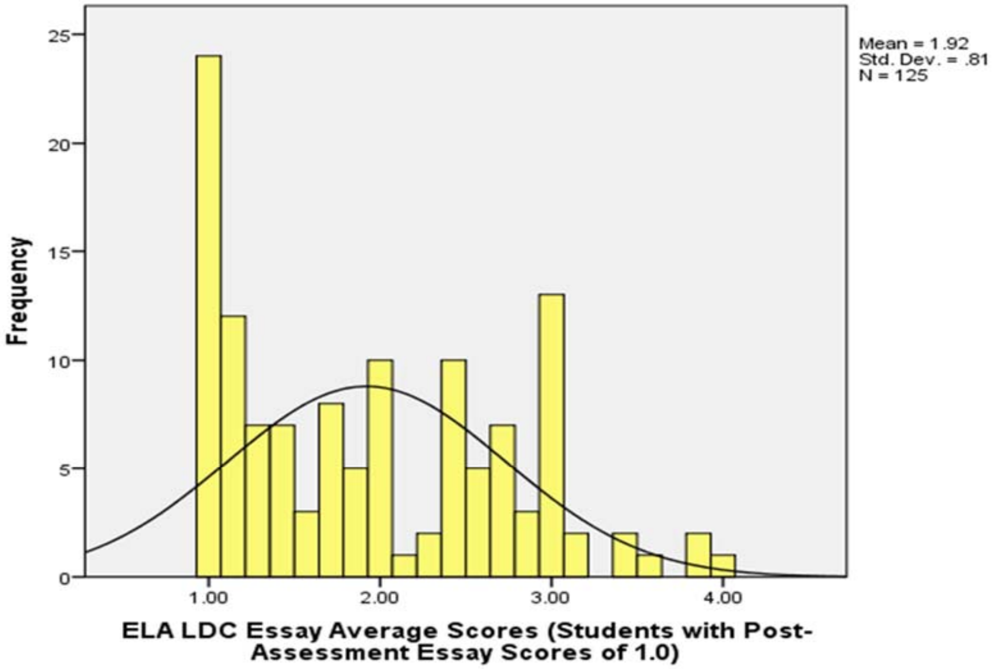


***Figure 9.* Distribution of Average Scores on ELA LDC Essays for Students with Average Score of 1.0 on the ELA Post-Assessment Essay**

Thus, these discrepancies in students' scores on the LDC writing tasks and their scores on the standardized on-demand writing task (Post-Assessment essay) suggest that the instructionally-embedded assessment provides students with opportunities to demonstrate understandings and skills that on-demand performance assessments do not.

***Finding 4:** The quality of the LDC writing tasks that were submitted by participating teachers varied widely in quality, based on ratings on the LDC Task Jurying Rubric. In addition, it appears that task quality may have a relationship with student performance, with students scoring higher on tasks that met the LDC Jurying Rubric's "Good to Go"* *threshold for task quality. This suggests that teacher-designed performance assessments lack sufficient comparability to be used for high-stakes consequences, or that sufficient controls would need to be put into place to vet and review teacher-designed tasks for quality, if the goal is to generate scores with some confidence in their comparability.*

One of the questions that we wanted to investigate in this study was the question of whether the LDC templates provide sufficient structure so that the tasks that are designed by teachers might be comparable. As part of this study, we asked teachers to submit the LDC writing tasks associated with the LDC essays completed by students and submitted by teachers. (These were tasks that teachers had designed using the LDC task templates for Argumentation Writing and Informational/Explanatory Writing tasks, and represented tasks to be used in ELA, history, and science courses.) We trained local secondary teachers representing ELA, history, and science content areas on the LDC Task Jurying Rubric using the LDC anchors for task jurying. We also checked for calibration and used only the data generated by raters who met calibration standards. We had each writing task scored by at least three raters. We used the "consensus scores" generated by those ratings.[4] Lead trainers from SCALE (who also serve as trainers for the national LDC module jurying training sessions) generated the master scores for the anchor modules, and also back-read the scores of all raters. In all, we rated 35 tasks from three different content fields – English language arts, history, and science.[5]

The LDC Task Jurying Rubric is used to evaluate writing tasks along four major dimensions of quality: 1) Task Clarity/Coherence, 2) Content, 3) Texts, and 4) Writing Product. It is also used to assign a holistic rating – "Exemplary," "Good to Go," or "Work In Progress" – based on those dimension scores. (The criteria for the holistic ratings are explained in **Figure A2** in the Appendix.)

[4] "Consensus scores" – the most frequent score assigned by raters on a given rubric dimension. For example, if there were three raters, and two raters scored the dimension "Work In Progress" (Level 1) and one rater scored the dimension "Good to Go" (Level 2), then the consensus score was "Work in Progress." If all three raters scored the dimension differently, the lead trainer adjudicated the score by conducting a "back-read" of the task.

[5] While 35 teachers from the research site submitted LDC teaching tasks as part of the study, not all completed data collection (did not submit student essays associated with those teaching tasks).

The juried task scores indicated wide variation in the quality of writing tasks designed by teachers in the study, based on the criteria in the LDC Task Jurying Rubric. This is illustrated in **Figure 10** and **Figure 11** below. A vast majority of writing tasks were rated as "Work in Progress" holistically, and only two were rated as "Exemplary." This suggests that for large-scale assessment purposes in which the results of the assessment have high stakes (e.g., teacher evaluation, student promotion, graduation), teacher-designed writing tasks are not ideal measures because of their lack of comparability, even when they are using common templates. This is not such a surprising finding given that the writing tasks were not designed specifically with summative assessment purposes in mind, and that the writing tasks were locally designed with little training or vetting. Instead, the writing tasks and instructional ladders were designed for formative purposes, with a focus on using the writing tasks as a means of helping students *learn* how to write a response to text-dependent prompts. It should also be noted that the teachers in the study had no introduction to or training on the LDC jurying rubric criteria prior to designing their LDC writing tasks *because the criteria did not yet exist*. So none of them had received training on a common set of criteria for evaluating their LDC writing tasks.

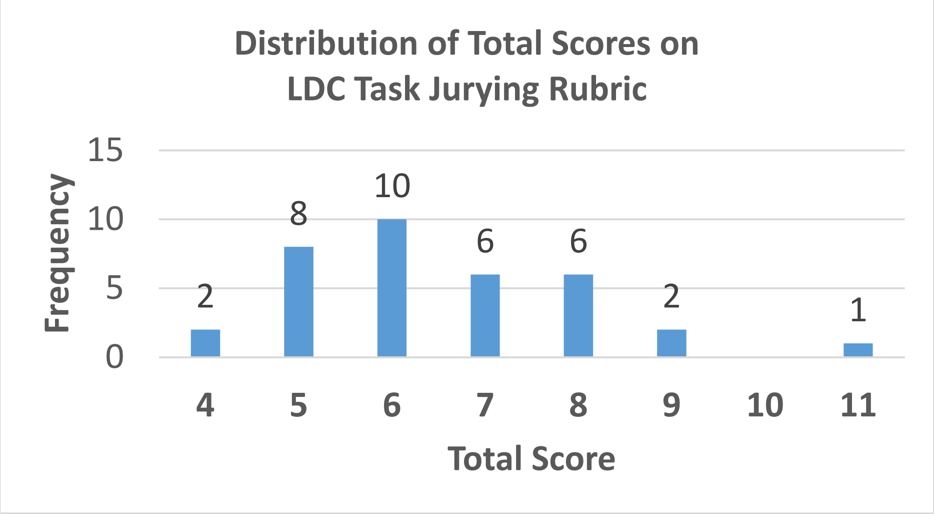


Figure 10. **Distribution of Total Scores on the LDC Task Jurying Rubric (Maximum Score = 12)**
Note: Total score is the sum of four dimensional scores, each scored on a three-point scale: Task Clarity/Coherence, Content, Texts, and Writing Product.

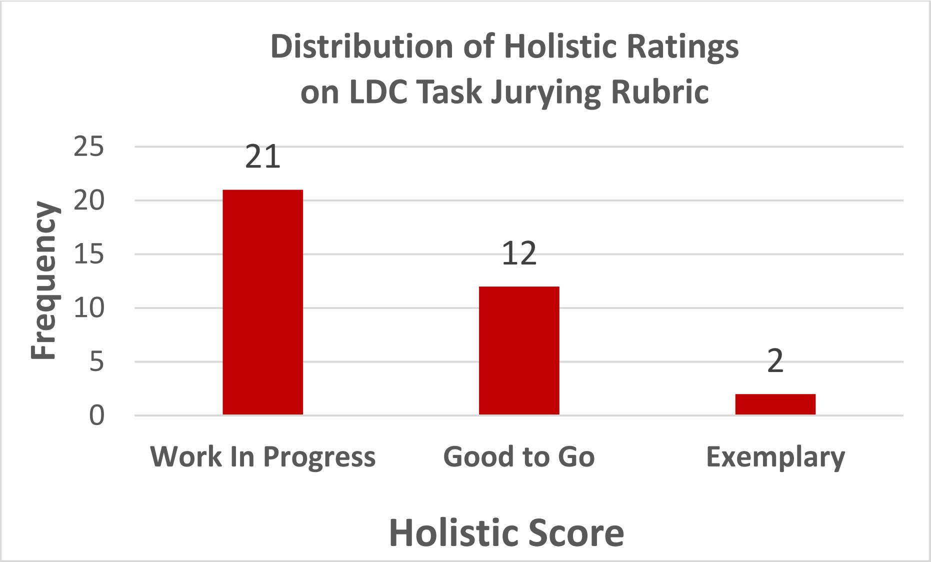


Figure 11. **Distribution of Holistic Ratings on LDC Task Jurying Rubric**

Another question that we asked was whether the quality of the tasks (as rated on the LDC Task Jurying Rubric) had any bearing on the scores students received on the essays (with scoring completed by external raters – in this case, raters trained by Measured Progress). We approached this question by analyzing the relationship between the LDC essay scores and the juried task scores (holistic rating as well as the sum of four dimension ratings). We found that the correlation between LDC essay scores and the juried task scores was statistically significant but low (0.133). This is likely due to a lack of variation in the task quality ratings, which were skewed, as noted earlier, toward the lower end of the scale, as well as to limited variation in the LDC essay scores (see **Figure 2** above).

When we looked at students' average total scores (the sum of 7 dimension scores) on the LDC tasks in relation to the different holistic ratings for the task (i.e., 1=Work In Progress, 2=Good to Go, 3=Exemplary), we found that students completing tasks that were rated "Work In Progress" had average total scores that were significantly lower than the average total scores of students completing tasks that were rated "Good to Go" or "Exemplary" (in many cases by a complete half point on the scoring rubric). This was also true of the seven separate dimensional scores. There were no significant differences in essay scores for students completing writing tasks rated as "Good to Go" and "Exemplary," but the sample size for the students completing "Exemplary" tasks was low (2 tasks, 35 students), limiting the power of this analysis. (See

Table A6 in the Appendix for a display of the average LDC essay scores across the different task quality ratings.)

This result suggests that students assigned to complete lower quality tasks (as measured by the LDC Task Jurying Rubric) might have had less of an opportunity to demonstrate high quality writing skills because of the poorer quality of the writing task. We cannot definitively claim that the cause of lower essay scores was lower quality writing tasks, because our design does not allow for making inferences about causality. An alternative explanation is that, in general, lower-achieving classes of students might have been assigned "Work In Progress" writing tasks based on the teachers' lower expectations for those students. Because we do not know the extent to which variation in student skills across teachers' classrooms is due to the quality of the writing tasks as opposed to variations in student characteristics and prior achievement across classrooms, or due to variations in teaching quality or other variations in teaching and learning contexts, we should be careful not to over-interpret this finding. This is, however, a worthwhile area for further investigation.

Based on these analyses, it is fairly clear that teacher-designed performance assessments lack sufficient comparability to be used for high-stakes consequences, and that if the goal is to have some confidence in the comparability of instructionally-embedded performance assessments, sufficient controls would need to be put into place to vet and review teacher-designed tasks for quality.

Finding 5: *LDC Essays can be scored reliably using the LDC Argumentation or Informational/Explanatory Rubrics. Error variation due to raters is very low across most dimensions (with the exception of the Conventions dimension), and score reliability for the "Total Score "– scores on all dimensions of the rubric, with one rater, exceeds the minimum standard of reliability for hand-scoring (0.80).*

These findings are based on two Generalizability studies that were conducted using Measured Progress raters who were trained to score the LDC essays submitted for this study. Two teacher-designed LDC writing tasks – one Argumentative writing task and one Explanatory writing task – were selected from among the tasks that were submitted by English language arts teachers in the study. The data sources and methodology of G-study are described in more detail in the Appendix beginning on page 51.

Results for the first ELA task (Argumentative Writing Task) are presented first below. **Table 4** below and **Figure A3** in the Appendix both show estimated reliability of each dimension of the rubric as a function of the number of raters used to calculate scores.

Table 4
Generalizability Study - Estimated Reliability of an LDC Argumentative Writing Task as a Function of Number of Raters Used to Calculate Scores

	Number of Raters			
Rubric Dimension	1	2	3	4
Focus	0.82	0.90	0.93	0.95
Controlling Idea	0.84	0.91	0.94	0.95
Reading Research	0.87	0.93	0.95	0.96
Development	0.71	0.83	0.88	0.91
Organization	0.71	0.83	0.88	0.91
Conventions	0.77	0.87	0.91	0.93
Content Understanding	0.83	0.91	0.94	0.95
Total Score	**0.89**	0.94	0.96	0.97

Scores for each dimension are highly reliable with as few as one rater being required to achieve a total score reliability greater than .80. The sub-scores for Development, Organization, and Conventions demonstrated only slightly lower levels of reliability.

Despite having high levels of reliability, the dimension scores are determined to be highly multi-colinear based on the estimated true score correlations (see **Table A8** in the Appendix). In other words, rather than functioning as independent dimensions, the tool appears to be measuring one underlying construct so that students that perform well on one dimension are highly likely to perform well on the other dimensions.

The results for the Explanatory Writing task are almost identical. The reliability of the individual sub-scores and total score are highly reliable (**Table 5 below** and **Figure A4** in the Appendix). These nearly identical results for the Explanatory writing task and the Argumentative writing task suggest that common rubrics support comparability when they are used to score writing tasks that were created using the LDC writing templates, even when the writing tasks themselves are different in purpose – Argumentative vs. Informational/Explanatory – and different in content.

Table 5
Estimated Reliability of the LDC Explanatory Writing Task as a Function of Number of Raters Used to Calculate Scores

	Number of Raters			
Rubric Dimension	1	2	3	4
Focus	0.92	0.96	0.97	0.98
Controlling Idea	0.86	0.92	0.95	0.96
Reading Research	0.83	0.91	0.94	0.95
Development	0.89	0.94	0.96	0.97
Organization	0.83	0.90	0.93	0.95
Conventions	0.69	0.82	0.87	0.90
Content Understanding	0.87	0.93	0.95	0.97
Total Score	**0.91**	0.96	0.97	0.98

The estimated correlations between dimensional scores for the Explanatory Writing task are also very high, once again suggesting multi-colinearity (see **Table A10** in the Appendix).

Taken together, the results suggest the LDC writing rubrics produce highly reliable Total Scores and somewhat reliable dimensional scores. However, the estimated true score correlations suggest the dimensions do not differentiate unique underlying facets of writing ability and that the tools measure a single underlying construct – writing ability.

Because the LDC writing rubrics are used primarily as a means for producing feedback for students, rather than to generate scores, we recommend not returning to holistic scoring as in the past. But reducing any potential redundancies in the scoring dimensions and reducing the number of dimensions would increase efficiency of scoring without sacrificing reliability. Feedback from LDC users suggests that there is a demand for revision of the LDC writing rubrics to streamline the dimensions (e.g., Focus and Controlling Idea are probably measuring the same thing), and clarify the distinction between other dimensions (e.g., the difference between Reading/Research and Development). Another problem is that because the original LDC rubrics were based on the Anchor Standards of the Common Core, there are no distinctions made between grade-level expectations. We suggest a careful round of review to produce the next iteration of the LDC writing rubrics.

Discussion

The Validity of Using Instructionally-Embedded Assessments for Summative Assessment Purposes

Although we found that essays written in response to LDC writing tasks can be scored reliably when raters are trained and calibrated, and that the LDC scores generated by the writing rubrics may provide reliable information about students' writing ability, there remain limitations on the usefulness of teacher-designed LDC writing tasks as measures for high-stakes or large-scale purposes because of their lack of comparability. And it appears that there may be a relationship between task quality and student performance. When students are responding to tasks that have not been rated as meeting a minimum level of quality (i.e., "Good to Go"), their essay scores are lower on average.

For districts or states contemplating the use of LDC writing tasks as "assessments" that can produce results for teacher evaluation or other high-stakes purposes, a "common task bank" or "common task" approach would be more appropriate than relying on writing tasks that individual teachers develop themselves. A common task approach is the use of the same set of performance tasks across teachers within a subject field and grade level to assess student progress or achievement. However, this approach may not appeal to teachers because the common task may not be aligned to teachers' local curricula, or teachers may feel that their curriculum and instruction are being driven by an external test, which is not much different from what we currently have in the form of state standardized tests. A more desirable approach may be the ***common task bank*** approach, in which teachers submit tasks they have designed and piloted to be considered for a central repository of highly vetted tasks that have been rated as "Good to Go" or higher. Teachers would be able to select from a menu of grade-level appropriate tasks that have been previously piloted, refined, and validated as being tightly aligned to the targeted learning outcomes, and that represent a range of curricular choices within their courses.

The Literacy Design Collaborative has been collecting expert-vetted and peer-juried writing modules that have undergone the jurying process and that meet at least the "Good to Go" criteria, as well as modules that have been rated as "Exemplary." These modules are available to the public in LDC's online open bank of modules known as "CoreTools."[6] States and districts may also consider developing, within CoreTools, their own LDC module collections that are customized to their state or local standards or curricula. These modules could be submitted for jurying by the national LDC organization, or could be vetted locally through the jurying tools and protocols that have been disseminated by LDC. In either case, if LDC tasks are to be used as potential measures of student growth or for high-stakes purposes such as educator effectiveness systems, the tasks should be reviewed and validated as meeting a standard of quality as established by the national LDC organization or other standards of quality (e.g., Achieve's EQuIP rubric). These processes of review and validation may not *ensure* comparability of the writing tasks or the scores, but they provide a moderation process that supports the quality and comparability of instructionally-embedded assessments used as measures.

The "Common Assignment Study (CAS)," a project in which SCALE has been involved over the last two years, illustrates another approach to the "common task bank" model. The CAS project has developed a set of common units of study across two states, Kentucky and Colorado. In this project, design teams of teachers led by curriculum experts and teacher leaders are developing common units of study for middle school and high school ELA, history, and science courses, aligned to the Common Core State Standards, Next Generation Science Standards, and/or state-specific standards for history and science. Embedded within each of these units of study is a common LDC writing task (either Argumentation or Informational/ Explanatory) as well as 2-3 other formative and summative assessments. These assessments and units of study have been piloted over the past two years and have been revised and refined, using feedback from teachers and insights from analysis of student work.[7] There are six teams in all across the two states, working to develop two units of study per course. In future years, these teams will be developing additional common units of study for their courses, to be vetted by curriculum experts, piloted in multiple schools, and refined prior to acceptance into the CAS unit bank. There are also plans to expand the project to include more grades and subject areas,

[6] http://ldc.org/coretools

[7] The technical and psychometric properties of the common assessments are currently being studied by the National Center for the Improvement of Assessment in Education.

particularly those that are not currently tested at the state level. CAS offers a promising strategy for developing a common task bank of disciplinary assessments embedded within units of study, supporting both the quality and comparability of the tasks developed and used by teachers so that they can be used for measurement purposes.

The Validity of Using Instructionally-Embedded Assessments as a Driver of Instruction and Student Learning

As noted at the beginning of this article, part of the validity argument for using instructionally-embedded assessments as a measure of student learning is that they are more closely connected to the curricula that teachers use and the learning experiences that students have in the classroom prior to completing the assessment. An additional aspect of validity that is related to the formative use of instructional-embedded assessments is consequential validity – how the assessments impact student learning and teachers' instruction.

While this study does not examine these impacts directly, other research studies conducted by external evaluators have examined the use of the LDC tools by teachers and their impact on instruction. Research for Action, a non-profit education research group, conducted surveys, interviews, and classroom observations of teachers engaged in the use of the LDC tools during the 2011-12 school year (Reumann-Moore & Sanders, 2012). They repeated the surveys and interviews in 2013 across 21 states, with over 1,500 teachers participating in the survey (54% response rate) (Research for Action, 2015). In 2013, 72% of surveyed teachers reported that using LDC modules had helped them find effective strategies for teaching their subject content, and 80% of teachers reported that they developed new ways to teach literacy skills in their content areas (Research For Action, 2015, p. 3). In the 2011 survey, between 70-80% of surveyed teachers reported that the LDC tools helped them to teach key areas of reading and writing such as summarizing, evaluating strength/weakness of evidence, comparing arguments, formulating a thesis statement, writing an introduction, and citing textual evidence to support claims. Science and social studies teachers, in particular, seemed to benefit from using the LDC tools to teach literacy skills. Eighty-seven percent of science teachers and 77% of social studies teachers reported that the LDC tools helped them to teach literacy (Reumann-Moore & Sanders, 2012, pp. 14-15). In addition, about 72% of surveyed teachers in 2013 reported that they had begun using LDC instructional strategies even when they were not teaching an LDC module and that they were infusing strategies from LDC modules into their ongoing instruction (Research For Action, 2015, p. 5).

With regard to the uses of the LDC tool for formative assessment purposes, 72% of surveyed teachers reported that the LDC assessments helped them learn detailed information about their students' strengths and weaknesses in literacy; about 66% of teachers reported that using the LDC tools helped them to include more formative assessments ("mini-tasks") in their instruction; and 71% reported that using the LDC tools had helped them give detailed feedback to students on their writing (Research For Action, 2015, pp. 3-4).

With regard to student learning, 78% of teacher respondents agreed that the LDC tools were effective in helping to make instruction more engaging for students (Research for Action, 2015, p. 7). In addition, 92% of surveyed teachers reported that the LDC tools were effective in improving students' literacy skills; 79% agreed that the LDC tools had helped to improve their students' writing; and 81% agreed that the LDC tools helped improve their students' understanding of content (Research for Action, 2015, p. 8).

Similar positive findings came out of a CRESST study that examined student learning outcomes in Kentucky. The study found that eighth-grade students whose teachers were using LDC tools outpaced the learning of a comparison sample of eighth-graders in reading by 2.2 months of learning (Research for Action, 2015, p.9). A more detailed explanation of this study can be found in Herman & Epstein (2014). While the short-run impacts of LDC tools on student learning reported here are modest, when more teachers (across grade levels and subject areas) begin using instructionally-embedded performance assessments such as LDC writing tasks, we can anticipate that such cumulative effects will have greater potential to impact student learning gains, particularly if the state external assessments are more aligned with the assessments in focus and format.

These outcomes related to changes in instruction and to student learning, though limited by their methodologies (teacher self-report) and measures (state achievement tests composed of multiple-choice items), provide support for a validity argument for instructionally-embedded performance assessments. The evidence suggests that such assessments are more instructionally useful than external standardized tests and that they are more likely to engage students and support their learning of important understandings and skills. If we think about the formative purposes of assessment, then instructionally-embedded assessments like the LDC writing tasks have a distinct advantage over external standardized tests.

Conclusions

The results of this study have important implications for the possible uses of instructionally-embedded assessments in large-scale assessment systems. Clearly, using instructionally-embedded assessments has important benefits for teachers and students, and there is little doubt that a focus on such assessments would be an effective instructional improvement strategy. However, the results on whether such assessments can be used as reliable and comparable measures of student learning and growth are mixed.

The G-study results suggest that performance assessments can be scored with high levels of reliability with sufficient training of raters and distributed scoring systems (in which teachers do not score their own students' work). The total scores that are generated by the LDC common writing rubrics for both Argumentation writing and Informational/Explanatory writing appear to be sufficiently robust and reliable, even when the tasks are different in purpose and the content/texts included in the writing tasks are different. This comparability in the reliability of scoring may have been facilitated by the use of common rubrics and common templates for generating the writing tasks.

However, it is unlikely that teacher-designed performance assessments like the LDC writing tasks can currently meet standards of comparability for inclusion in high-stakes, large-scale assessment and accountability purposes, unless controls are put into place for reviewing and certifying such assessments as meeting a minimum threshold for quality. We saw wide variability in the quality of the writing tasks developed by teachers, and that these variations in writing task quality appear to have a bearing on students' ability to demonstrate proficiency in writing. It is more likely that for instructionally-embedded performance assessments to be accepted as policy instruments, they will need to be drawn from a common task bank, such as the juried modules in the LDC CoreTools resource bank, or the collaboratively-designed, expert-vetted, and practitioner-tested LDC writing tasks developed by the Common Assignments Study described earlier. LDC and other groups, both for-profit and non-profit organizations, such as

Amplify, Achieve, and the CCSSO's Innovation Lab Network, have been working on developing systems for reviewing unit and assignment quality, and building curated resource banks of vetted curriculum resources.

Second, the results of this study suggest that instructionally-embedded and on-demand performance assessments provide different kinds of information about students' ability to demonstrate their disciplinary knowledge and skills *under different administrative conditions*, and that instructionally-embedded tasks may provide unique information about students that external, standardized assessments cannot provide. In the literacy and writing domain, these include the ability to collect and synthesize information from a variety of authentic sources, and the ability to engage in the writing process (craft, edit, and polish an original thesis), skills that cannot be authentically assessed in on-demand, timed conditions.

Instructionally-embedded assessments such as the LDC writing tasks also appear to be more accessible than on-demand writing assessments for students because of the instructional supports and scaffolding provided to them while they are planning and working on their writing in collaboration with their peers and with feedback from their teachers. Instructionally-embedded assessments involving writing tasks may reduce students' aversion to writing and the negative affects associated with on-demand testing. Such assessments give students opportunities to access the task, persist, and complete their work, and they provide teachers with more helpful diagnostic information than on-demand, standardized writing tasks, on which a large percentage of students are often simply too intimidated to even start writing, resulting in high non-response rates, or non-scorable/overly brief responses that result in the lowest possible score.

Given the limitations of our findings due to low sample sizes and biases in the findings due to missing data, further investigation of instructionally-embedded assessments is needed to understand their possible role in a more balanced assessment and accountability system that includes both external assessments and local, instructionally-embedded assessments administered by teachers. In particular, we need more robust information about how scores from instructionally-embedded assessments are related to scores on external measures, and what is truly different about how students perform on external versus instructionally-embedded assessments, administered under different conditions. For example, do instructionally-embedded assessments advantage or dis-advantage students who perform poorly on external assessments or on-demand assessments? Are students with learning challenges (e.g., English language learners, special education students) better able to demonstrate their learning and achievement on instructionally-embedded assessments than on external assessments or on-demand assessments? These questions, among others, provide rich opportunities for further inquiry into the usefulness of developing an assessment system that includes both external standardized assessments and instructionally-embedded performance assessments.

References

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Center on Education Policy (CEP). (2011). Profile of state high school exit exam policies: Rhode Island. Washington, DC: S. McIntosh.

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.

Darling-Hammond, L. & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Goe, L. & Holdheide, L. (2011). *Measuring teachers' contributions to student learning growth for non-tested grades and subjects.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from: http://files.eric.ed.gov/fulltext/ED520722.pdf

Herman, J., & Epstein, S. (2014). *Supporting middle school content teachers' transition to the Common Core: The implementation and effects of LDC.* Los Angeles, CA: National Center for Research, Evaluation, Standards, and Student Testing. Retrieved on May 29, 2015 from: http://ldc.org/sites/default/files/research/CRESST%202014%20LDC%20Report.pdf

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

Koretz, D.M. (1998). Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education, 5*(3), 309-334.

Lane, S. (2010). *Performance assessment: The state of the art. (SCOPE Student Performance Assessment Series)*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

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

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

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5-12.

McDonnell, L.M. (2004). *Politics, persuasion, and educational testing.* Cambridge, MA: Harvard University Press.

National Research Council (NRC). (2003). *Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment.* Washington DC: Committee on Assessment in Support of Instruction and Learning.

Research for Action (2015, February). LDC's influence on teaching and learning (Brief One). Philadelphia, PA: Author. Retrieved on May 29, 2015 from: http://ldc.org/sites/default/files/research/RFA%20%2B%20CRESST%20LDC%20Brief%20%231_Feb%202015.pdf

Reumann-Moore, R. & Sanders, F. (2012, September). Robust implementation of LDC: Teachers' perceptions of tool use and outcomes (Brief Two). Philadelphia, PA: Research for Action. Retrieved on May 29, 2015 from: http://ldc.org/sites/default/files/research/RFA%20Brief%202_Robust%20Implementation%20of%20LDC.pdf

Shepard, L.A. (2000). The Role of Assessment in a Learning Culture. *Educational Researcher, 29* (7), 4-14.

Shepard, L.A. (2003). Reconsidering large-scale assessment to heighten its relevance to learning. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom,* pp. 121-146. Arlington, VA: National Science Teachers Association Press.

Stecher, B. (1998). The local benefits and burdens of large-scale portfolio assessment. *Assessment in Education, 5*(3), 335-351.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-765.

Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation, 2*(2). Retrieved March 27, 2014 from http://PAREonline.net/getvn.asp?v=2&n=2

APPENDIX: Tables and Figures

Table A1
LDC Study - Data Collection Timeline

Autumn 2012	Teacher recruitment (Determine who is implementing LDC modules at the HS level, get agreement from districts, schools, teachers)
January–February 2013 (courses at the high school level were one-semester courses that begin in August and January, with the second semester ending in May)	Teachers administer Pre-Assessment (2 days) • Collect consent forms • Collect student rosters with State Student ID
January - May 2013	Teachers implement LDC module, collect student essays
April-May 2013	Teachers administer Post-Assessment (2 days) to same students who completed Pre-Test. Submit LDC essays for same students.
April 2013	Students take end-of-course state tests
November 2013	Initial request for administrative data
April 2014	Final receipt of data from district(s)

Table A2

Relationship between 2013 State Literature Exam Scores and Prior State Test Scores in Reading and Writing

Main Predictors	Model 1		Model 2		Model 3	
First Prior Reading Score	0.649	***	0.571	***	0.373	***
(Grade 8)	(0.028)		(0.043)		(0.050)	
First Prior Writing Score	0.219	***	0.177	***	0.114	**
(Grade 8)	(0.027)		(0.042)		(0.042)	
Second Prior Reading Score					0.324	***
(Grade 7)					(0.050)	
Second Prior Writing Score					0.021	
(Grade 7)					(0.034)	
Statistical Controls						
Current Grade Level	0.143	***	0.230	***	0.236	***
	(0.065)		(0.067)		(0.064)	
Ethnicity	0.068		0.149		0.123	
0 = non-Caucasian; 1 = Caucasian	(0.039)		(0.102)		(0.064)	
Gender	0.061		0.076		0.083	
0 = Male; 1 = Female	(0.046)		(0.062)		(0.029)	
Economic Disadvantage	-0.084		-0.116		-0.099	
0 = no; 1 = yes	(0.046)		(0.073)		(0.071)	
ELL Status	-0.199		-0.131		-0.059	
0 = no; 1 = yes	(0.020)		(0.461)		(0.044)	
Special Ed Status	-0.217	***	-0.311	**	-0.253	*
0 = no; 1 = yes	(0.063)		(0.111)		(0.107)	
r2	0.609		0.548		0.589	
N	1036		456		456	

Note: These models are for all students in the administrative data and do not take into consideration whether the student participated in the CRESST or LDC performance assessments.

Regression Models for Examining Relationships among State Standardized Tests and Performance Measures

In the first regression model, prior achievement data are examined as predictors of 2013 State Literature Exam scores to create a base regression model. Then the other measures (on-demand Pre- and Post-Assessment scores, LDC Essay scores) were added to that base model as new independent variables, step by step, to assess their relationship with the 2013 State Literature Exam scores (the outcome variable). In addition, the change in the percentage of the score variance accounted for by the additional independent variables was recorded. Student demographic variables were also included in the regression models as controls.

The step-by-step data analyses are summarized below:

- Run outcomes (the State Literature Exam scores, the on-demand Post-Assessment scores, the LDC essay scores) as a function of all the prior state tests available. Save the r^2.
- Run (1) but now include the on-demand Pre-Assessment score. See the additional r^2 and test the significance.
- For all but the LDC essay scores, rerun (2) but now include the on-demand post-assessment scores.
- Correlate the outcome variables in (1)

When the Grade 3-8 state tests in Reading and Writing are entered into a series of regression models as predictors of the 2013 State Literature Exam Scores, there is a strong and positive coefficient for the Grade 8 Reading score and a weaker positive coefficient for the Grade 8 Writing score. **Table A2** above shows the magnitude and significance of the relationships. In Model 1, the percentage of variance in the 2013 State Literature Exam Scores accounted for by the Grade 8 Reading score is moderately high (60.9%). In Model 2, when we run the same analyses using only the student data for whom we have both Grade 8 and Grade 7 Reading/Writing scores, we find that the predictive value of the Grade 8 scores decline slightly, as does the total variance accounted for in Model 2 (from 60.9% to 54.8%). When an additional prior year's test scores (Grade 7 Reading and Writing Test) are added to Model 3, we see that the coefficient for the Grade 8 Reading score declines and the Grade 7 Reading score is introduced as a slightly weaker but still significant coefficient. In addition, the r^2 increases slightly from 54.8% to 58.9%.

This tells us that both years' prior test scores could be used in the base model; however, to maximize sample size, and given the r^2 improvement is negligible, only the Grade 8 scores are used as predictors in the base model.

Using Model 1 in **Table A3** as the base model, we add our independent variables of interest--our performance assessment scores. We start with the ELA teachers and students in our sample. See **Table A4** below for the regression models using the 2013 State Literature Exam scores as the outcome measure and the performance assessment scores as the predictors.

Despite relatively low samples sizes, we find that the Total Pre-Assessment Score, the Total Post-Assessment Score, and the LDC Essay Score are all positive and significant predictors of the 2013 State Literature Exam Scores, with coefficients of 0.340, 0.374, and 0.273 respectively. This means, for example, that when the LDC Essay Score increases by one standard deviation, the State Literature Exam Score increases by 0.273 of a standard deviation. The Total Pre-Assessment Score, Total Post-Assessment score, and LDC Essay Score are less predictive than Grade 8 State Reading Test scores but more predictive than the Grade 8 State Writing Test score (Model 4, 8 and 10). In each case, when performance task data are added, the r^2 increases by a modest 2-4% indicating that the explanatory value of the model is slightly improved with the performance task data.

Table A3

Predicting 2013 State Literature Exam Scores - ELA Performance Tasks

	Essay Pre-Test		Total Pre-Test		Essay Post Test		Total Post Test		LDC Essay Score	
Main Predictors	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
First Prior Reading Score	0.945 ***	0.768 ***	0.950 ***	0.695 ***	0.944 ***	0.817 ***	0.949 ***	0.725 ***	1.005 ***	0.855 ***
(Grade 8)	(0.119)	(0.139)	(0.118)	(0.14)	(0.118)	(0.124)	(0.118)	(0.127)	(0.14)	(0.146)
First Prior Writing Score	0.106	0.128	0.107	0.165	0.051	0.023	0.052	0.052	-0.023	-0.028
(Grade 8)	(0.114)	(0.112)	(0.114)	(0.111)	(0.113)	(0.109)	(0.113)	(0.105)	(0.127)	(0.121)
Essay Pre-Test		0.216 *								
		(0.094)								
Total Pre-Test				0.340 **						
				(0.111)						
Essay Post-Test						0.233 *				
						(0.089)				
Total Post-Test								0.374 ***		
								(0.106)		
Total LDC										0.273 *
										(0.109)
Statistical Controls										
Current Grade Level	0.189	0.157	0.199	0.187	0.226	0.145	0.234 *	0.156	0.484 *	0.695 **
	(0.113)	(0.111)	(0.112)	(0.107)	(0.116)	(0.116)	(0.115)	(0.109)	(0.213)	(0.219)
Ethnicity	0.134	0.158	0.130	0.233	0.214	0.203	0.211	0.221	0.424	0.424
0 = non-Caucasian; 1 = Caucasian	(0.240)	(0.235)	(0.240)	(0.232)	(0.238)	(0.229)	(0.237)	(0.221)	(0.379)	(0.361)
Gender	0.25	0.232	0.262	0.244	0.362 *	0.338 *	0.375 *	0.318 *	0.113	0.116
0 = Male; 1 = Female	(0.155)	(0.152)	(0.154)	(0.147)	(0.160)	(0.154)	(0.158)	(0.149)	(0.202)	(0.193)
Economic Disadvantage	-0.080	-0.040	-0.069	-0.038	-0.012	0.055	0.000	0.028	0.238	0.220
0 = no; 1 = yes	(0.214)	(0.210)	(0.213)	(0.204)	(0.214)	(0.208)	(0.213)	(0.199)	(0.306)	(0.291)
ELL Status	-0.388	-0.266	-0.399	-0.203	-0.540	-0.565	-0.550	-0.470	-2.485 **	-2.468 **
0 = no; 1 = yes	(0.439)	(0.432)	(0.438)	(0.423)	(0.428)	(0.412)	(0.426)	(0.398)	(0.904)	(0.861)
r2	**0.657**	**0.673**	**0.658**	**0.688**	**0.676**	**0.699**	**0.677**	**0.720**	**0.665**	**0.696**
N	94	94	95	95	83	83	84	84	59	59

Note: Special Ed Status was excluded from the analysis due to a lack of data that would have reduced the sample size substantially. All main predictors and the criterion are standardized to a mean of 0 and a standard deviation of 1.

*Significant at the 0.05 level **Significant at the 0.01 level ***Significant at the 0.001 level

Table A4
Predicting ELA Performance Assessment Scores with Prior State Reading and Writing Test Scores

	Essay Pre-Assessment	Total Pre-Assessment	Essay Post Assessment	Total Post Assessment	LDC Essay Score
ELA Performance Assessment					
First Prior Reading Score	0.413 ***	0.507 ***	0.084	0.417 ***	-0.012
(Grade 8)	(0.100)	(0.091)	(0.101)	(0.094)	(0.143)
First Prior Writing Score	0.005	-0.098	0.200 *	0.025	0.109
(Grade 8)	(0.086)	(0.083)	(0.090)	(0.084)	(0.110)
r2	0.154	0.212	0.048	0.186	0.260
N	129	134	118	120	95

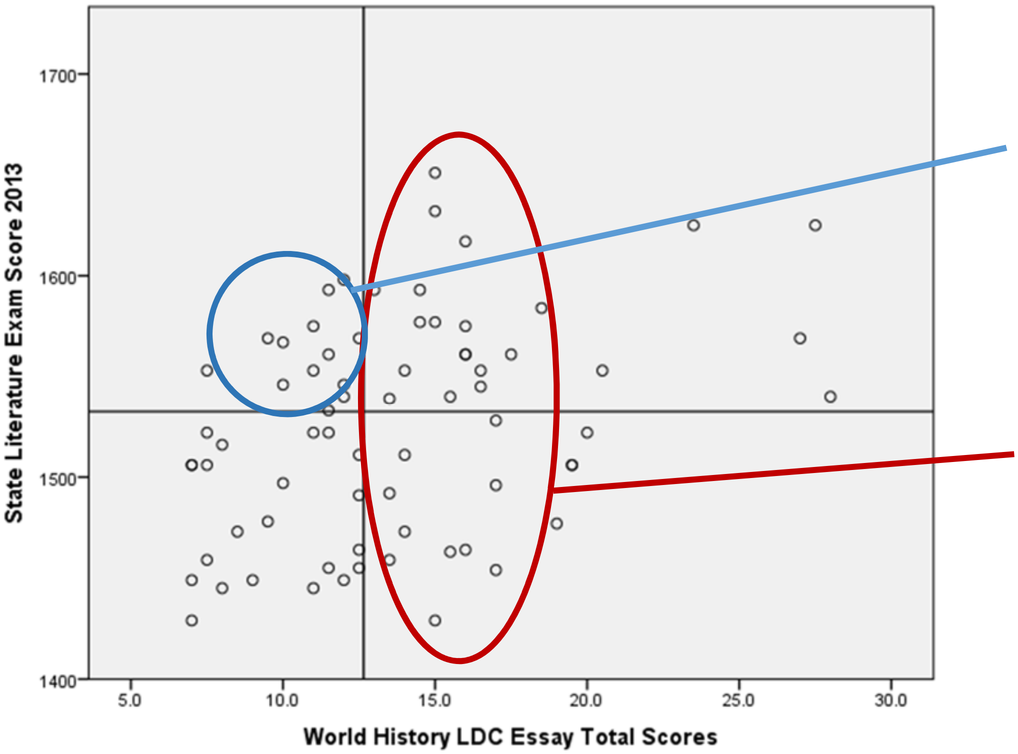
Note: All models also include controls for student grade level, ethnicity, gender, economic disadvantage, ELL status, and special education status fixed effects. All main predictors and the criterion are standardized to a mean of 0 and a standard deviation of 1.

Table A5
Pairwise Correlations among State Tests and Performance Assessment Scores

		1	2	3	4	5	6	7	8
ELA Performance Assessment									
1	State Literature Exam	1.00							
2	Essay Pre-assessment	0.59	1.00						
3	Total Pre-assessment	0.64	0.91	1.00					
4	Essay Post-assessment	0.65	0.67	0.68	1.00				
5	Total Post-assessment	0.72	0.72	0.80	0.90	1.00			
6	LDC Essay Score	**0.38**	0.42	0.48	**0.49**	0.52	1.00		
7	First Prior Reading Test Score	0.79	0.65	0.69	0.68	0.71	0.40	1.00	
8	First Prior Writing Test Score	0.47	0.31	0.31	0.46	0.40	0.31	0.59	1.00

When we examine the Pre- and Post-Assessments as the outcomes of interest, we find that state reading test scores from Grade 8 have a positive and significant coefficient, meaning that they are moderately strong predictors of performance on the on-demand, Pre- and Post-Assessments. The Grade 8 Writing test is also a weak but significant and positive predictor of the Post-Assessment Essay score for the ELA on-demand writing task. This makes sense, given the similarities in the format of the tests and the on-demand, standardized conditions in which they were administered. See **Table A4** above.

On the other hand, neither the Grade 8 Reading nor Writing test scores are significant predictors of students' LDC Essay scores. **Table A5**, which shows the pair-wise correlations between the performance measures and the State Literature Exam scores also shows a relatively lower correlation (0.38) between the LDC essay score and the State Literature Exam score. (Bi-variate Pearson correlation coefficients tell us the extent to which rank-ordering of students on one measure is consistent with the rank-ordering of the same students on another measure.)



Students who score above average on the State Literature Exam (1532) but score relatively poorly on the LDC Essay (between “Not Yet” and “Approaches” scores)

Students who score between “Approaches” and “Meets Expectations” on the LDC Essays have a wide range of scores on the State Literature Exam.

***Figure A1*. Boxplot of Literature Exam Scores and World History LDC Essay Total Scores**
(Note: Total cases: 71. Average score on State Literature Exam in this sample is 1532. Average LDC Essay Total Score in this sample is 12.7)

We see more students scoring higher on the State Literature Exam but scoring below average on their world history LDC essay, and a wide range of scores on the State Literature Exam in the middle of the LDC essay score scale. The first observation suggests that the world history LDC Essays measure somewhat different competencies or knowledge domains from what is assessed in the State Literature Exams. The bivariate Pearson correlation between the two measures is only 0.359 (statistically significant), likely due to the wide spread of scores in the middle of the score ranges.

✓ ***Exemplary:*** The task can be used with students with high confidence in the intended results, can be used or easily adapted by other educators, and is a model for emulation.

Tasks scored at the Exemplary level are truly models to share, emulate, and adopt AS IS. They are not perfect, but the weaknesses within the task are negligible or require only minor tweaks. This represents a very high standard, requiring technical, conceptual, and pedagogical precision.

✓ ***Good to Go:*** The task can be used with students with some confidence in the intended results.

Tasks scored at the Good to Go level are those that have many strong features, and are likely to yield good results with students. There is nothing really "wrong" with these tasks, but they are generally less customized and detailed, less focused on discipline-specific approaches to literacy and thinking, less intellectually rigorous, or less authentic or engaging as assignments for students.

✓ ***Work in Progress:*** There are one or more aspects of the task that suggest a need for revision to be a useful writing assignment for students. Some of those revisions may be significant, and other revisions may be minor "tweaks."

Tasks scored at the Work in Progress level often have some strengths that show strong potential and one or more clear areas for improvement. WIP tasks frequently have one or more significant problems that suggest a clear need for improvement before it is used with students. This level is applied to communicate with the task author that revision is advised.

***Figure A2*. Criteria for Holistic Task Rating on LDC Task Jurying Rubric**

Table A6

Average LDC Essay Scores By Holistic Ratings of Task Quality

		N	Mean	Std. Dev.	Std. Error	95% Confidence Interval for Mean		Min	Max
						Lower Bound	Upper Bound		
Focus	1=Work In Progress	454	1.695	.6825	.0320	1.632	1.758	1.0	4.0
	2=Good to Go	256	2.180	.8569	.0536	2.074	2.285	1.0	4.0
	3=Exemplary	35	2.100	.7356	.1243	1.847	2.353	1.0	3.5
	ALL	745	1.881	.7836	.0287	1.824	1.937	1.0	4.0
Controlling Idea	1=Work In Progress	454	1.605	.7013	.0329	1.540	1.669	1.0	4.0
	2=Good to Go	256	2.154	.8737	.0546	2.047	2.262	1.0	4.0
	3=Exemplary	35	2.043	.7894	.1334	1.772	2.314	1.0	3.5
	ALL	745	1.814	.8115	.0297	1.756	1.872	1.0	4.0
Reading/ Research	1=Work In Progress	454	1.641	.6992	.0328	1.576	1.705	1.0	4.0
	2=Good to Go	256	2.238	.8811	.0551	2.130	2.347	1.0	4.0
	3=Exemplary	35	2.157	.8114	.1372	1.878	2.436	1.0	4.0
	ALL	745	1.870	.8224	.0301	1.811	1.930	1.0	4.0
Development	1=Work In Progress	454	1.611	.7026	.0330	1.546	1.676	1.0	4.0
	2=Good to Go	256	2.176	.8979	.0561	2.065	2.286	1.0	4.0
	3=Exemplary	35	2.086	.7523	.1272	1.827	2.344	1.0	3.5
	ALL	745	1.828	.8223	.0301	1.768	1.887	1.0	4.0
Organization	1=Work In Progress	454	1.753	.6774	.0318	1.691	1.816	1.0	4.0
	2=Good to Go	256	2.240	.8413	.0526	2.137	2.344	1.0	4.0
	3=Exemplary	35	2.129	.7984	.1349	1.854	2.403	1.0	3.5
	ALL	745	1.938	.7779	.0285	1.882	1.994	1.0	4.0
Conventions	1=Work In Progress	454	1.801	.6996	.0328	1.736	1.865	1.0	4.0
	2=Good to Go	256	2.234	.7969	.0498	2.136	2.332	1.0	4.0
	3=Exemplary	35	2.100	.7154	.1209	1.854	2.346	1.0	3.5
	ALL	745	1.964	.7625	.0279	1.909	2.019	1.0	4.0
Content Understanding	1=Work In Progress	454	1.726	.7563	.0355	1.656	1.796	1.0	4.0
	2=Good to Go	256	2.232	.8446	.0528	2.128	2.336	1.0	4.0
	3=Exemplary	35	2.200	.7784	.1316	1.933	2.467	1.0	4.0
	ALL	745	1.922	.8251	.0302	1.863	1.981	1.0	4.0
Total Score	1=Work In Progress	454	11.831	4.6350	.2175	11.404	12.259	7.0	28.0
	2=Good to Go	256	15.455	5.7812	.3613	14.744	16.167	7.0	28.0
	3=Exemplary	35	14.814	5.2104	.8807	13.024	16.604	7.0	25.0
	ALL	745	13.217	5.3658	.1966	12.831	13.603	7.0	28.0

* For all dimensions and the total score, the average score differences between students completing tasks rated as "Work in Progress" vs. "Good to Go" are statistically significant (p<.001), while the average score differences between students completing tasks rated as "Good to Go" vs. "Exemplary" were not significant.

Data Sources and Methodology for Generalizability Studies of LDC Tasks and Rubrics

Two teacher-designed ELA writing tasks – one Argumentation writing task and one Informational/Explanatory writing task - were selected from among the tasks that were submitted by ELA teachers in the study, and 20 random samples were selected from student essays for each task. Three raters were assigned to score the same 20 samples for the Argumentation writing task, and a different three raters were assigned to score the same 20 samples for the Informational/Explanatory writing task. (Since a significant amount of time had passed between their initial training to score the LDC essays and the G-study scoring, the raters were re-calibrated by the lead trainer through a discussion of two calibration papers prior to independent scoring began.) The LDC writing rubrics each have seven different dimensions including: Focus, Controlling Idea, Reading/Research, Development, Organization, Conventions, and Content Understanding. Each dimension is scored on a seven point scale: 1, 1.5, 2, 2.5, 3, 3.4, and 4.

Reliability of the two ELA tasks is estimated using Generalizability Theory (G-Theory). G-Theory facilitates the modeling of reliability based on tractable sources of error. The ELA task scores vary depending on differences in student ability, differences in rater severity, and differences in the interaction between student ability and rater severity. In other words, scores from the ELA tasks can be represented using a student by rater (s x r) multivariate model.

G-Theory was used to estimate the variance attributable to each component of the observed scores (see **Table A7** and **Table A9** below). These estimates were used to calculate the expected reliability depending on the number of raters for each dimension of the rubric as well as the total score as calculated as the sum of the scores across the individual dimensions. Results for the first ELA task (Argumentation writing task) are presented first. **Table 4** on page 29 and **Figure A3** below both show estimated reliability of each dimension of the rubric as a function of the number of raters used to calculate scores.

Table A7

Source Table for the LDC Argumentation Writing Task

Source	Focus	Controlling Idea	Reading Research	Development	Organization	Conventions	Content Understanding
Student	37.85 (91%)	37.52 (86%)	32.28 (83%)	37.7 (88%)	32.92 (81%)	23.66 (60%)	30.48 (86%)
Rater	0.39 (1%)	0 (0%)	0 (0%)	0.2 (0%)	0.99 (2%)	5.15 (13%)	0.59 (2%)
Error(s x r)	3.36 (8%)	6.29 (14%)	6.58 (17%)	4.8 (11%)	6.93 (17%)	10.68 (27%)	4.41 (12%)

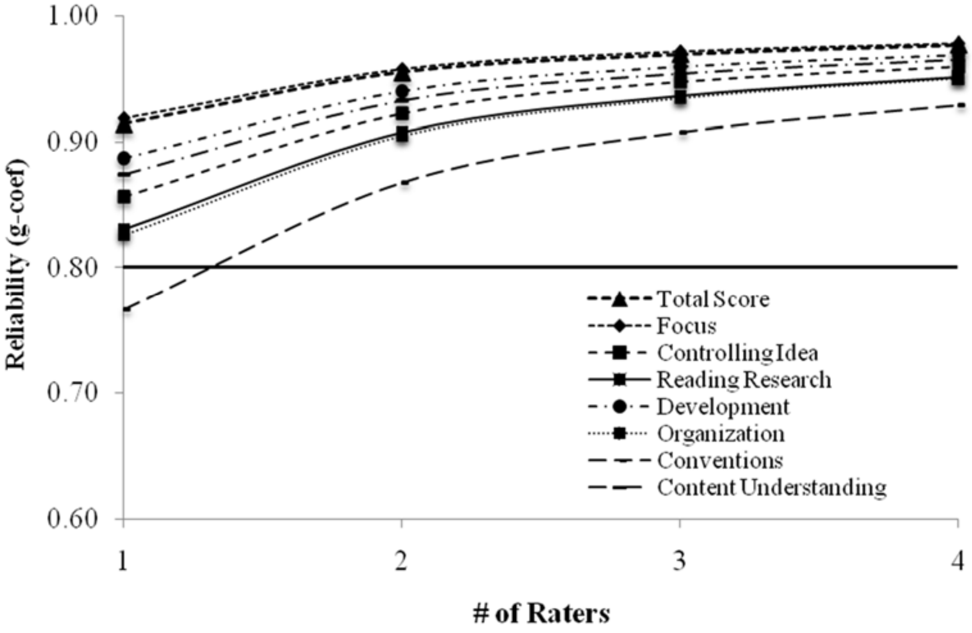


Figure A3. **Estimated Reliability of the Argumentative Writing Task as a Function of Number of Raters Used to Calculate Scores**

Table A8
Argumentation Writing Task - Estimated True Score Correlations

	2	3	4	5	6	7
1. Focus	0.91	0.85	0.83	0.86	0.75	0.85
2. Controlling Idea		0.93	0.98	0.89	0.85	0.90
3. Reading/Research			~1	0.93	0.82	0.92
4. Development				0.96	0.90	0.99
5. Organization					0.92	0.87
6. Conventions						0.83
7. Content Understanding						na

Table A9
Source Table for the LDC Informational/Explanatory Writing Task

Source	Focus	Controlling Idea	Reading Research	Development	Organization	Conventions	Content Understanding
Student	37.85 (91%)	37.52 (86%)	32.28 (83%)	37.7 (88%)	32.92 (81%)	23.66 (60%)	30.48 (86%)
Rater	0.39 (1%)	0 (0%)	0 (0%)	0.2 (0%)	0.99 (2%)	5.15 (13%)	0.59 (2%)
Error(s x r)	3.36 (8%)	6.29 (14%)	6.58 (17%)	4.8 (11%)	6.93 (17%)	10.68 (27%)	4.41 (12%)

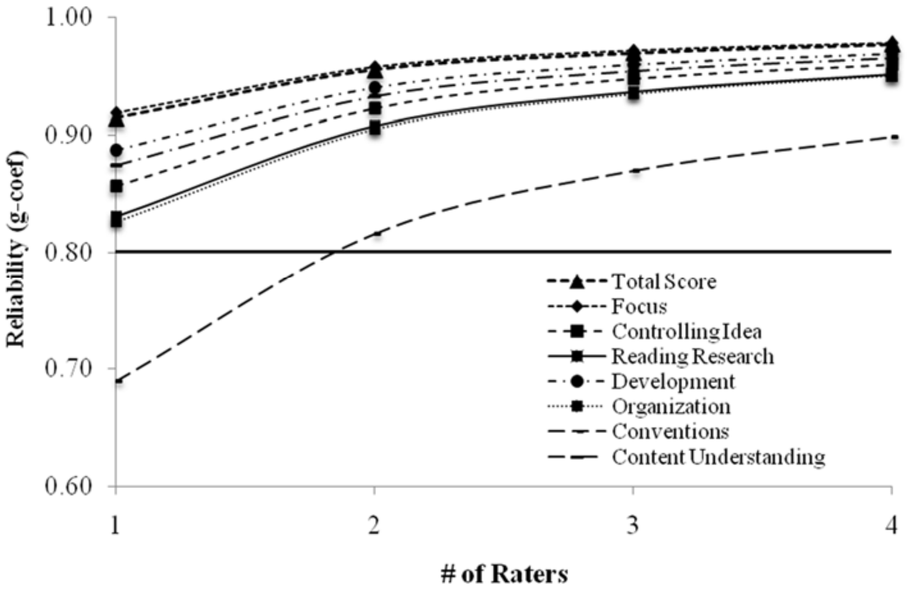


Figure A4. **Estimated Reliability of the Informational/Explanatory Writing Task as a Function of Number of Raters Used to Calculate Scores**

Table A10

Informational/Explanatory Writing Task Estimated True Score Correlations

	2	3	4	5	6	7
1. Focus	0.99	0.95	0.94	0.97	0.98	0.99
2. Controlling Idea		0.97	0.98	0.97	0.99	0.99
3. Reading/Research			~1.00	0.99	~1.00	1.00
4. Development				0.98	~1.00	0.99
5. Organization					0.98	0.97
6. Conventions						0.98
7. Content Understanding						na