Through-Course Common Core Assessments in the United States: Can Summative Assessment Be Formative?

A paper presented at the annual meeting of the American Educational Research Association
New Orleans, LA

Walter D. Way
Katie Larsen McClarty
Dan Murphy
Leslie Keng
Charles Fuhrken

April 2011
Abstract

In this paper, we present a design for enhancing the formative uses of summative through-course assessments. Starting with the Common Core assessment plans articulated by the Partnership for Assessment of Readiness for College and Careers (PARCC), we present the argument that formative information used to improve teaching and learning can be best obtained using a particular through-course assessment design that leverages online testing, automated scoring approaches, complete and immediate disclosure of both tasks/prompts and student responses, and that explicitly avoids commonly used test equating practices. We argue that this design can optimize the use of assessment results in formative decision-making and still retain the desired degree of psychometric rigor that characterizes high-stakes educational assessments. We provide a specific illustration of our proposed through-course assessment design in the context of the PARCC English language arts (ELA) assessments.

Keywords: Common Core State Standards, through-course assessment, formative assessment
Through-Course Common Core Assessments in the United States: Can Summative Assessment Be Formative?

Introduction

In the United States, new approaches to large-scale summative assessment have been proposed that include “through-course” components administered during the instructional year to support teaching, learning, and program improvement. These intended uses appear to be formative in nature. Black and Wiliam (2009) and Wiliam (2010) recently argued that assessment is formative to the extent that results are used by teachers, students, or peers to better make or confirm decisions about subsequent instructional actions. That raises the question: can summative assessment be formative?

In this paper, we argue that summative through-course assessment can in fact be formative, and we further argue that this can be best achieved using a particular through-course assessment design that leverages online testing, automated scoring approaches, complete and immediate disclosure of both tasks/prompts and student responses, and that explicitly avoids commonly used test equating practices. We believe that this design can optimize the use of assessment results in formative decision-making and still retain the desired degree of psychometric rigor that characterizes high-stakes educational assessments.

We begin the paper by providing background on recent reforms in educational assessment in the U.S. and the proposed assessment designs that are currently emerging from this reform. Next, we introduce the idea of through-course assessments, and discuss some of the issues with this approach that have been raised by assessment experts. Before turning to the specifics of our proposed through-course design, we review key ideas from the content standards that are serving as the basis for new assessments in English language arts. These ideas strongly influence our proposed model and our ideas about how these through-course assessments can most fully support formative practice. Measurement and psychometric issues are discussed next, and our solution to these reveals the full intent of our model. Finally, we discuss some practical implementation issues associated with our proposed model and consider how these issues might be addressed should our ideas be pursued in practice.
Background

Educational assessment in the United States is undeniably being shaped by recent political events. Under the American Recovery and Reinvestment Act of 2009, the President and Congress invested unprecedented resources into the improvement of K-16 education in the United States. As part of that investment, the $4.35 billion Race to the Top Fund focused on a state-by-state competition for educational reform. Race to the Top also included $350 million in competitive grants that were awarded in 2010 to two state consortia for the design of comprehensive new assessment systems to accelerate the transformation of public schools: the Partnership for Assessment of Readiness for College and Careers (PARCC)\(^1\) and the SMARTER Balanced Assessment Consortium (SBAC)\(^2\). The assessments to be developed by these consortia will measure the Common Core State Standards (CCSS), which were the result of a voluntary effort to develop a set of evidence-based standards in English Language Arts and Mathematics essential for college and career readiness in a twenty-first century, globally competitive society (CCSSO & NGA, 2010a; 2010b).

The Common Core assessments are ambitious in that they have multiple goals and purposes. On the one hand, the consortia are required to implement summative assessment components in both mathematics and English language arts by the 2014-2015 school year. Yet, the assessments must “produce data (including student achievement data and student growth data) that can be used to inform (a) determinations of school effectiveness; (b) determinations of individual principal and teacher effectiveness for purposes of evaluation; (c) determinations of principal and teacher professional development and support needs; and (d) teaching, learning, and program improvement” (Department of Education, 2010, p. 18171).

The challenge facing the consortia developing the Common Core assessments can be illustrated by considering Figure 1, which is taken from a presentation by Nellhaus (2009) to the U.S. Department of Education in a public meeting discussing the Race to the Top assessment program. In this graphic, there are two overarching goals that Nellhaus argued should characterize next generation assessment systems: ensuring accountability and improving

---

\(^1\) http://www.achieve.org/parcc
\(^2\) http://www.k12.wa.us/smarter/
teaching and learning. The inference drawn in the graphic is that a *unified system* with summative, benchmark and formative assessments is needed to achieve these goals.

![Diagram of Overarching Goals](image)

**Figure 1. Overarching Goals for Next Generation Assessment Systems (from Nellhaus, 2009)**

The first priority for each consortium is to develop new Common Core summative assessments that can be implemented by 2015, yet the assessments still have to produce results and data the will inform teaching, learning, and program improvement. Given that summative, standards-based assessments in the U.S. have been almost universally criticized for not providing timely and useful information to inform teaching, learning, and program improvement, how can the consortia achieve this goal with the Common Core assessments?

**Through-Course Assessment Design**

For at least the PARCC consortium, the answer to this question lies in the use of through-course assessment, defined by the U.S. Department of Education as follows:

Through-course summative assessment means an assessment system component or set of assessment system components that is administered periodically during the

---

3 Because the SBAC consortium is not currently planning through-course assessment as part of their summative assessment design, this paper is focused on the model proposed by the PARCC consortium.
academic year. A student’s results from through-course summative assessments must be combined to produce the student’s total summative assessment score for that academic year (U. S. Department of Education, 2010, p. 18178).

The PARCC assessment design consists of two main summative components denoted as through-course assessment and end-of-year assessment (PARCC, 2010). Through-course assessments are intended to link assessment to instruction throughout the school year and to provide feedback to the teachers when they still have time in the year to intervene. Assessment can take place closer to the time when material is taught in the classroom. This is similar to interim or benchmarking practices that exist at many schools; but rather than being separate, through-course assessment results will be integrated into the final student scores at the end of the year. As conceptualized by PARCC, through-course assessments will be administered at roughly 25%, 50%, and 75% of the instructional year. There are several anticipated benefits of a through-course assessment approach including:

- Multiple assessments. Students have multiple opportunities to demonstrate their knowledge and skills. There is less pressure placed on a single test at the end of the year.

- Performance tasks in assessment. Through-course assessments will consist of performance tasks that require students to demonstrate thorough mastery of the CCSS. These are also the types of tasks that students need to be able to perform to be ready for college and careers.

- Signaling good instruction. The through-course performance tasks will model the types of activities that teachers should engage in with their students throughout the year. These tasks should be valuable instructionally as well as provide assessment information.

- Providing diagnostic feedback. Feedback will be provided to teachers to help improve instruction and inform interventions at quarterly intervals.

- Providing early indicators. When a student or class is off track, information will be available earlier, and steps can be taken to intervene and get them back on track.
Clearly, many of these benefits arise from supporting formative uses of the evidence elicited from through-course assessments. On the other hand, there are also several challenges facing the PARCC consortium in using through-course assessments. Many of those challenges involve the psychometrics and measurement needed to support the design and interpretations of the assessment results. For example, through-course assessments that consist of one or two performance tasks resulting in one or two scores present difficult obstacles to achieving reliable results that can be equated for different students. In addition, combining scores from the different components of the assessment system is not straightforward. Psychometric models that assume student proficiency is constant throughout the year or that student proficiency is the same over different types of assessment tasks may not hold. It will be important to be able to show growth in student proficiency throughout the year and to make inferences based on the test scores about student preparedness for college and careers.

Role of the Standards

The measurement challenges noted above exist for both assessments of ELA and mathematics; however, each content area also faces some domain-specific challenges. The through-course assessment design we favor has a more direct application to ELA and a primary reason for this is the nature of the CCSS in ELA. Specifically, these standards “translate the broad (and, for the earliest grades, seemingly distant) aims of the [College and Career Readiness] standards into age- and attainment-appropriate terms. The Standards set requirements not only for English language arts (ELA) but also for literacy in history/social studies, science, and technical subjects” (CCSSO & NGA, 2010a, p.3). The integrative nature of the standards can be seen, for example in Writing Standard #9 at grade 6:

Draw evidence from literary or informational texts to support analysis, reflection, and research.

a. Apply grade 6 reading standards to literature (e.g., “Compare and contrast texts in different forms of genres [e.g. stories and poems; historical novels and fantasy stories] in terms of their approaches to similar themes and topics”).
b. Apply grade 6 reading standards to literary nonfiction (e.g., “Trace and evaluate the argument and specific claims in a text, distinguishing claims that are supported by reasons and evidence from claims that are not”) (p. 44).

It is evident that this standard integrates grade 6 reading standards for both literature and informational text. Thus, a task measuring Writing Standard #9 could also measure Reading Standard for Literature #9 or Reading Standard for Informational Text #8. Furthermore, depending upon how the task was crafted, other reading and writing standards could also be measured.

Measuring these integrated standards is exactly the intent of the PARCC consortium’s use of through-course assessment for ELA: “These through-course components are designed to measure the most fundamental capacity essential to achieving college and career readiness according to the CCSS: the ability to read increasingly complex texts, draw evidence from them, draw logical conclusions and present analysis in writing” (PARCC, 2010, p. 46). The PARCC consortium is very specific in its application about the structure of the through-course assessment components.

Another feature of PARCC’s ELA through-course assessment is an emphasis on text complexity, which reflects a similar emphasis in the CCSS. (Appendix A of the CCSS includes a lengthy and detailed discussion of the role of text complexity in college and career readiness.) The PARCC (2010) application states, “A fundamental aspect of the Partnership’s proposed ELA/literacy components is the inclusion of texts that are appropriately complex and of high quality. To ensure that the texts used for the system are at appropriate level of complexity for each grade, the Partnership will need to determine a rigorous methodology for selecting texts for inclusion” (p. 198).

Maximizing Through-Course Assessments for Formative Purposes

We believe that the PARCC ELA through-course assessment can best support formative information by using a design that involves developing, field-testing, and administering a large number of constructed response tasks for each of the through-course assessments within a given
grade (e.g., 100 to 150 tasks per grade). Within a grade and assessment, these tasks are all based on texts determined to be at equivalent levels of complexity. The tasks themselves are also assumed to be randomly equivalent to each other so that psychometric equating and scaling is not necessary. Features of this design include:

- All tasks are field-tested by randomly administering them within the targeted population. Field-test responses are scored by humans and used as a basis for applying automated scoring.

- All texts, associated tasks, and all anchor papers used to train human scorers are made available to be viewed online by teachers, students, and parents. Documentation providing rationales for the anchor paper scores is developed and also made available. The human scores are used to train the appropriate automated essay scoring engine or engines.

- Field-tested tasks found to have extreme difficulty levels relative to the set of tasks to be assumed equivalent are eliminated so that the difficulty levels of the remaining tasks fall within an acceptable range.

- For the operational through-course assessments, the task administered to each student is randomly selected from the full set of available tasks. Scores on different tasks are not equated but rather assumed to be directly comparable because of the equivalence of the prompts.

- Automated scoring is applied to 100 percent of the operational online student responses, and the text, the task administered, and the student response to the task are all made available to teachers, students, and parents virtually immediately after the assessment is completed. A small percentage of unusual or “unscorable” responses as detected by the automated scoring engine are routed to human scoring. In addition, a small percentage of randomly selected responses are also scored by humans to serve as an ongoing audit for the automated scoring procedures.
• Teachers are given the opportunity to challenge the automated score if they disagree with it, provided they have been through appropriate scorer training. Some limits on challenges may be necessary, such as setting a maximum number of challenges per teacher and/or imposing a cost if the original score is upheld. A successful challenge results in the teacher’s score replacing the score originally assigned.

• Text complexity has an explicit role in scoring the through-course assessments and in measuring student growth.

• Coverage of the CCSS is achieved by varying the combinations of standards measured by each of the large number of randomly selected tasks. This provides an elegant solution to the problem that not all of the CCSS standards can be assessed in each student’s test.

Disclosing all prompts greatly enhances the formative aspects of the through-course assessments by increasing the transparency of the assessments. Furthermore, the immediacy of feedback that includes prompt, response, and score provides teachers and students with important evidence about achievement that can be used to inform or corroborate decisions about next steps in instruction. A full appreciation of our proposed through-course assessment design requires considering several associated issues, including the through-course assessment tasks, the associated scoring rubrics that might be used, and the details of the psychometric approach supporting the through-course components within the full summative assessment system. We turn to these considerations next.

Task Considerations for ELA Through-Course Assessments

Given the importance of text complexity and students reading and writing to source texts, our proposed through-course assessment design focuses on these elements. For each through-course assessment, a student will be given a piece of high quality text of appropriate complexity. The text selections for each through-course component will come from a variety of literature and informational texts, including history/social studies, science, and technical subjects in the higher grade levels. The specific description of each component of the summative assessment system is detailed below.
ELA-1 and ELA-2 Through-Course Assessments. Through-course assessments 1 and 2 will each be administered in a single class period. Students will read a single piece of high quality text and respond to two extended constructed response questions. The text for the ELA-1 assessment will come from the literature genre whereas the text for ELA-2 will come from the informational genre. In this way, a teacher will receive information about the performance of each student on both literary and informational texts from components of the summative assessment system by the time half of the course has been completed. These summative through-course components can be used in a formative way to better understand student interactions with the different types of texts.

The constructed response questions will require students to draw evidence and logical conclusions from the text (i.e., writing standard 9) in addition to covering additional CCSS standards in the reading, writing, and language domains. Specifically, the first question will address the key ideas and details found in the text. Questions will be drawn from the reading standards 1-3. For example, standard 2 under Literature/Key Ideas and Details requires grade 7 students to “determine a theme or central idea of a text and analyze its development over the course of the text.” Therefore, a grade 7 student could read an excerpt from Mildred D. Taylor’s *Roll of Thunder, Hear My Cry* and write about how the author develops the theme of the importance of family.

The second question will address craft and structure by requiring interpretation or evaluation of the text. Questions will be drawn from the reading standards 4-6. For example, standard 6 under Informational Text/Craft and Structure asks grades 11-12 students to “determine an author’s point of view or purpose in a text in which the rhetoric is particularly effective, analyzing how style and content contribute to the power, persuasiveness, or beauty of the text.” Therefore, a grade 11 or 12 student could read Ronald Reagan’s “Address to Students at Moscow University” and discuss how the author’s ideas and presentation style contribute to his persuasive purpose.

ELA-3 Through-Course Assessment. Through-course assessment 3 will be a two-day performance activity, with the assessment tasks each day occurring in a single class period. On the first day, students will read a single, long piece of quality text. Students will then provide one
extended response to a question asking students to analyze and interpret the craft and structure of the text they have read (reading standards 4-6). On the second day, students will read a second piece of text that is linked by subject or theme to the first, or, they may watch a media clip related to the text they read the previous day. They will then respond to two questions. The first question will deal with interpreting the craft and structure of the second text or media clip (reading standards 4-6). The second question will ask students to make comparisons between the two selections (reading standard 9). The paired selections could be literature/literature, literature/informational, or informational/informational in genre. In lieu of printed text, media clips could be presented, such as viewing a scene from a play for the literature genre or a speech for informational. The pairing presented to an individual student will be randomly selected from a pool of options, but there will be an increased number of informational pieces in the pool at higher grade levels as outlined in the CCSS.

For example, reading standard 9 under Integration of Knowledge and Ideas calls for grade 4 students to “integrate information from two texts on the same topic in order to write or speak about the subject knowledgeably.” Therefore, on the ELA-3 through-course assessment, a grade 4 student could read an informational text about hurricanes on day one and watch a video of a meteorologist’s report about hurricanes on day 2. Then the student could write a response that compares the main points and key details about the warning signs and dangers of hurricanes featured in the two selections.

**ELA-4: End of year assessment.** The end-of-year assessment will cover a greater breadth of content standards than could be covered in the through-course assessments. Concepts such as vocabulary and editing can be assessed in this assessment as well as other aspects of reading comprehension. Specifically, a media piece will be included to assess reading standard 7 and at least one informational piece to assess reading standard 8. Students will be expected to read texts of the appropriate complexity level, and they will respond to items that can be scored by a computer. Computer-scored items can include technology-enhanced items that assess higher order thinking skills. Students can view video clips, listen to audio clips, highlight text in passages, drag and drop paragraphs to reorder them, or edit text. This assessment will be administered on computer to students at all grade levels so that results can be reported quickly.
**ELA-5: Speaking and Listening.** In the PARCC assessment system, there is another through-course assessment component that pertains to speaking and listening. The ELA-5 through-course assessment will be completed following the ELA-3 assessment. In this assessment, students will present their analysis of the comparison between the two pieces from ELA-3 to their classmates. Using the ELA-3 example above, the grade 4 student could prepare a presentation on the warning signs and dangers of hurricanes to share with his or her classmates. Teachers will score performance on this component, but the score will not contribute to the composite ELA summative assessment score. This component is similar to traditional formative assessment.

**Scoring Rubrics for Through-Course Assessments**

We propose that the rubrics for scoring each of the through-course extended constructed response items will cover three separate domains, consistent with domains found in the CCSS: reading, writing, and language. Each domain will be scored on a 4-point scale. As shown in Table 1, the first rubric domain will assess student performance on the reading standard (1-6 or 9) assessed by the questions. The reading score will focus on the content and ideas contained in the answer. The second rubric domain score will assess the quality of student writing, focusing on writing standards 4 and 9 including development, organization, style, and drawing evidence from the text. The third score will reflect student performance on language standards 1-3 including writing conventions and knowledge of language.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Focus</th>
<th>CCSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>Content and ideas</td>
<td>1-6 or 9</td>
</tr>
<tr>
<td>Writing</td>
<td>Development, organization, style, and drawing evidence from text</td>
<td>4, 9</td>
</tr>
<tr>
<td>Language</td>
<td>Writing conventions and knowledge and language</td>
<td>1-3</td>
</tr>
</tbody>
</table>

Table 1. Rubric for Evaluating TCA Responses
This rubric will be applied to each through-course assessment question regardless of text type or question asked. A student will receive a score on each question ranging from 3-12, based on an equal combination of the subscores in the reading, writing, and language domains. The detailed rubric descriptions of student performance in each domain at each score point will facilitate interpretation of the rubric scores. The assessments are designed as integrated tasks, but the reporting of the separate subscores can assist teachers’ formative use of assessment scores for understanding and diagnosing student strengths and weaknesses.

**Developing Equivalent Through-Course Assessment Tasks**

One of the hallmarks of this assessment design is the assumption of randomly equivalent tasks. Implementing evidence-centered design (ECD) in developing the assessment tasks is one approach that can support development of approximately equivalent tasks and comparable inferences about student scores (Huff, Steinberg, & Matts, 2010). When ECD is implemented, the claims and evidence needed to support the claims are well articulated, and task models can be developed to support item development and to provide a link between the items and the claims (Hendrickson, Huff, & Leucht, 2010). The possible variations on features included in the task model provide the opportunity to create multiple items that provide students the opportunity to demonstrate evidence in support of the claim. One of the challenges of ECD is the unfamiliar terminology used to describe the process (Ewing, Packman, Hamen, & Clark Thurber, 2010). Therefore, an example is provided here for illustrative purposes. A complete ECD process would need to be undertaken to develop these assessments in practice.

Using an ECD approach, one desirable claim to make of all students taking the ELA assessment might be that “The student is able to describe key ideas and details from the text, using evidence from the text to support the answer.” Example pieces of evidence that would support this claim could include: articulation of specific relevant instances from the text, determination of story theme, in depth description of a specific character in the story, identification of the main idea of the text, description of the connection between two pieces of information in the text, and more. Then task models and items would be developed so that students could demonstrate evidence that they are able to describe key ideas and details from the text. Although students may respond to different questions based on different texts, the
inferences that are being made about students—the degree to which they can describe key ideas and details from the text using evidence—are similar.

Other parts of the assessment design also support equivalent tasks. First, students at the same grade level will be reading texts of similar levels of complexity. The complexity level of the text is one indicator of difficulty; thus, prompts addressing text of similar complexity should have similar difficulty levels. Second, students are held to the same grade level expectations in the CCSS about the skills needed to engage each text and answer questions. The skills that are needed are equal regardless of the text. Each assessment task represents a sampling from the domain of expectations. Finally, student responses will be judged against a common rubric which reflects the grade level expectations in reading, writing, and language. The specific content evaluated under the reading portion of the rubric may change slightly depending on which specific reading standard was assessed, but the resulting score will reflect the same interpretations about key ideas and details, craft and structure, or integration of knowledge and ideas.

**Advantages of the Proposed Model**

A diagram of the proposed model for the ELA assessment system is shown in Figure 2. The through-course design realizes several of the desired benefits of a through-course approach. First, it provides multiple opportunities to gather evidence of student performance. Across the school year in the through-course summative assessment system, students have multiple opportunities and methods to show their learning including writing to different types of texts, viewing or listening to media, and responding to traditional multiple choice items. Second, teachers can view the assessments students were given, including the text and student responses, which increases the ability to use the summative through-course assessments as formative assessment. Teachers could have conferences with students following the assessment to gain a deeper understanding of student misconceptions and to inform future instruction for that student. Third, these tasks model good instructional practices. Students should be reading and responding to quality texts in the classroom. The performance tasks on the assessments mirror that same process. Finally, the performance tasks resemble authentic contexts that students will encounter in their everyday lives. Outside of the classroom, students will be expected to read, understand,
and evaluate various types of texts from sources such as newspapers, journals/magazines, Internet articles, blog postings, advertisements, and manuals.

![Diagram](image)

**Figure 2. English Language Arts Comprehensive Summative Assessment Design**

**Psychometric Approach for Supporting Through-Course Assessment Components**

A necessary component of any valid assessment is that it must be reliable. The evaluation of reliability can be challenging when scores are based on a single task or a small number of constructed response tasks scored by humans and/or computer. To evaluate the reliability of constructed response tasks, it is important to consider the various components that influence a test taker’s score. One of the components that will influence a test taker’s score in our through-course assessment design is the complexity of the text.

Text complexity is one of the key criteria of the CCSS for ELA. According to a study conducted by ACT (2006), the ability to read complex text is one of the single best predictors of college performance. The focus on text complexity requires methods of categorizing text by complexity levels. Several measures of text complexity can be considered, including MetaMetrics’ Lexile scale, McNamara and Graesser’s (in press) Coh-Metrix Text Complexity Component Scores, and Landauer, Kireyev & Panaccione’s (2011) latent semantic analysis
(LSA) based measure. Subject matter experts can also be recruited to help categorize text by complexity levels. To reduce variability in test scores due to text complexity, we propose categorizing the tasks in the bank by their complexity level and type of text.

For example, suppose three categories that are deemed grade-appropriate for Grade 8 ELA include (text complexities/item type): 101-150/literary, 151-200/informational and 201-250/literary or informational. Then for ELA-1, items eligible to be administered would be from the 101-150/literary group, for ELA-2, they would be from the 150-200/informational group, and for ELA-3, they would be from the 201-250/literary or informational group. For each TCA within a grade level, tasks are randomly assigned by computer to test takers. Our assumption of randomly equivalent tasks is based on the random assignment to test takers of tasks that are the same text type at similar levels of complexity.

**Psychometric Considerations**

To evaluate the assumption of randomly equivalent tasks and assess the psychometric properties of the proposed ELA through-course assessment design, it is important to identify the sources of measurement error affecting the precision of the test scores. Measurement error can be quantified through an application of generalizability theory (G-theory; Cronbach, Gleser, Nanda & Rajaratnam, 1972; Feldt & Brennan, 1989) and/or by using a structured equation modeling (SEM; Raykov 1997, 1998) approach.

One source of measurement error results from test score variability within texts that are assumed to be of the same complexity level. This variability may be exacerbated with genres such as poems or when media clips are part of the materials encountered by students. This type of measurement error would result in test score variability due to tasks rather than to differing test-taker ability. It is also possible that test score variability would depend to some degree on the interaction of the test taker (p) with the task (t). With the G-theory approach, person × task (p × t) pilot studies can be conducted to estimate measurement error from these two sources and evaluate the assumption of randomly equivalent tasks.

A separate source of measurement error is the rater (r) scoring the test. Automated
scoring reduces variability across raters, because all tests are scored by the same rater (i.e., the computer); however, measurement error within the automated scoring engine exists. The measurement error due to automated scoring can be estimated using a G-theory pilot study by assigning the participants multiple tasks and analyzing the variability of automated scoring across tasks. By using a person x task x rater \((p \times t \times r)\) G-theory pilot study design, error variance due to tasks can be compared to that due to other sources of error such as raters. D-studies can then be conducted to estimate the reliability of the proposed through-course model and suggest assessment design changes if necessary.

In addition, a SEM approach using a model similar to the one shown in Figure 3 can be used to estimate measurement error. In the example model shown, TCA-1 to TCA-4 represent the (observed) scores obtained on each through-course assessment, the factor loadings for R1 represent variance that is accounted for by the measured factor, and the variances represented by E1 to E4 represent measurement error variance. The errors of measurement in this model are assumed to have a mean of 0 and to be uncorrelated across tasks. Therefore, in theory, errors of measurement for any particular student should approach 0 in the long run across assessments. It is important to note that this model also assumes that the same factor (i.e., R1) is measured across the TCA-1 – TCA-3 tasks. The single factor and uncorrelated error assumptions can be empirically tested by fitting and comparing alternative models.

![Figure 3: Example SEM model for proposed ELA through-course assessment design](image-url)
Scoring

As stated earlier, to the extent possible, the use of automated scoring is recommended to score not only items in TCA-4, but also the extended-response items in TCA-1 to TCA-3. This would allow for fast turnaround time of test results to inform classroom instruction. Recent research (e.g., Foltz, Gilliam & Kendall, 2000; Landauer, Laham & Foltz, 2003; Nichols, 2005) has supported the reliability and validity of automated scoring engines in the rating of performance-based tasks, such as written composition and constructed response items. The constructed response tasks for TCA-1 to TCA-3 are scored on 3 dimensions in accordance with the standards: reading, writing, and language. Each dimension is scored on a 4-point scale, and the three scores are summed to generate the final score. TCA-4 consists of machine-scored items, most of which are expected to be dichotomous.

Composite Score Reliability

Following completion of all assessment components, the ELA Summative Weighted Composite Scores would be calculated and reported as scale scores. Subscale scores, performance category classifications, and growth indices will also be provided. In considering the weights for the summative score, it seems prudent to balance the practical need to give substantial weight to the constructed-response task components and the psychometric need to have adequately reliable weighted composite scores. We would therefore suggest a scheme such as weighting each of the through-course components (ELA-1 to ELA-3) 10% and the end-of-year assessment (ELA-4) 70%. Obviously, there are other weighting schemes possible and those might be preferred based on policy considerations.

Students are tested at different times throughout the school year; therefore student proficiency cannot be assumed to be constant across the assessments. Because we assume that measurement errors are uncorrelated across the TCA tasks, an estimate of error variance in the composite score can be found by taking a weighted sum of the error variances for each task. Stratified alpha estimates reliability as 1 minus the ratio of error variance to total composite score variance:

---

4 Comprehensive reviews of the automated scoring literature can be found in Dilki (2006) and Phillips (2007).
In the above equation, $w_i$ is the weight associated with assessment $X_i$, $\sigma_{X_i}^2$ is the variance of assessment $X_i$, $\rho_{X_i,X_i'}$ is the reliability estimate of assessment $X_i$, and $\sigma_Z^2$ is the variance of the total composite score.

**Measuring Student Growth**

We propose the use of a growth-to-proficiency model to develop an “on track” indicator to measure student progress within a grade level and provide individual student feedback on their likelihood of meeting the summative passing standard at course completion. The growth-to-proficiency model would use each test-taker’s initial performance (i.e. on TCA-1) to derive growth targets on the remaining through-course assessments so that they will meet the summative passing standard upon course completion. Scores on each of the assessments are weighted by complexity to create some verticality in the scale. For example, an automated score of 8 on both TCA-1 and TCA-2 assessments are weighted such that the scores reported to the test taker reflect growth from time 1 to time 2. Each test-taker’s performance at TCA-2 and TCA-3 is compared to his or her respective growth targets to determine whether they are progressing adequately throughout the course. Interventions can then be provided if a student is progressing at a rate below his or her target.

A similar “on track” indicator can be used to measure student progress across grades towards college readiness, which would be conceptualized as meeting the summative passing standard for the final (Grade 11) ELA course.

**Practical Considerations**

Several practical considerations related to our proposed ELA through-course design are worth mentioning. These include precedents from other high stakes assessments that do not equate tasks but rather assume equivalence across tasks given across test-takers or administrations, cost and feasibility considerations associated with implementation, additional
supports that might be utilized to enhance formative information, and the question of whether our proposed through-course design can also be applied to Common Core assessments in mathematics.

**Assessments without Equating**

In the U.S., test equating carries more weight in assessment design than in perhaps any other country. Most psychometricians in the U.S. assume the need for equating as a given in most summative assessment situations. However, even in the U.S. there are examples where equating is not applied to some or even all components of high stakes assessments. One such example is the Analytical Writing Assessment (AWA) of the Graduate Management Aptitude Test (GMAT®). The GMAT includes measures of Verbal and Quantitative aptitude administered using computerized adaptive testing. In addition to these measures, the AWA consists of two 30-minute writing tasks—Analysis of an Issue and Analysis of an Argument. These are also administered by computer and scored by one human scorer and also using automated scoring technology. Students taking the AWA respond to prompts that are randomly selected from a large pool of publically-available prompts that are assumed to be randomly equivalent.\(^5\) Interestingly, students can request an independent re-scoring of their AWA results for a fee of US$45.

There are other examples of performance assessments across a number of disciplines where equating is not utilized, for example, computerized performance-based licensure assessments in medicine (Clauser, Harek & Clyman, 2000; Melnick & Clauser, 2006), architecture (Bejar, 1991; Bejar & Braun, 1994; Williamson, Bejar, & Hone, 1999), and the National Board Certification of master teachers\(^6\). Although these programs do vary assessment tasks across test-takers and/or across administrations, unlike the GMAT, they do not disclose the tasks that are used operationally within their assessments.

---

\(^5\) The complete set of AWA prompts used in GMAT administrations can be found at [http://www.mba.com/mba/TheGMAT/TestStructureAndOverview/AnalyticalWritingAssessmentSection/](http://www.mba.com/mba/TheGMAT/TestStructureAndOverview/AnalyticalWritingAssessmentSection/).

\(^6\) See [www.nbpts.org](http://www.nbpts.org) for information about the assessment approach used in this certification.
In short, the notion of sampling from a pool of assessment tasks that are not equated has precedent, particularly for performance assessments that are developed according to strict task requirements. For the Common Core through-course assessments, it is important to recognize that the associated high-stakes decisions will not be at the individual student level but rather at higher levels of aggregation, such as teachers, schools, districts and states. Thus, while results for an individual student could be affected because a particularly difficult or easy task was drawn from the pool, across a sample of students it is reasonable to assume that the difficulty effects will cancel out. The important requirement for this assumption is random assignment, which is effectively achieved through computer administration.

**The Challenge of Developing a Pool of Through-Course Assessment Tasks**

Our proposed model for through-course in ELA requires creating large pools of assessment tasks that meet the requirements for the ELA-1, ELA-2, and ELA-3 through-course assessments as outlined above. This development effort will clearly be challenging. For example, suppose 80 prompts are desired for the operational pool for each PARCC through-course assessment at each grade from 3 to 11. Assume further that some percentage of developed prompts, perhaps 20 percent, will not work. Under these assumptions, the total number of prompts that would have to be field-tested would be $3 \times 9 \times 100 = 2,700$. In addition to the large total number, development will have to be guided by a prescribed matrix of design features such as literary versus informational text, a balanced representation of the different relevant CCSS, and specific ranges of text complexity.

Some strategies may be used to reduce the number of prompts that have to be developed for the through-course assessments under our proposed design. For example, it might be possible to pair different sets of tasks that might cover different CCSS in reading and language with the same text. It might also be possible to overlap prompts across grade levels, as implied by the range and levels of text complexities for student reading presented in the CCSS. These strategies could create synergies in the use of the prompts and reduce the overall developmental burden.

On top of the sheer volume of the development effort required by our through-course design is the challenge associated with obtaining high-quality authentic texts. That is,
permissioned passages from published works can be expensive and time-consuming to obtain. Although passages from classic works (e.g., works from authors who died before 1941) may be available in the public domain, more contemporary pieces generally require a permissions arrangement. Permissioned passages sometimes have severe restrictions, such as being usable only in print on a particular test form given within a particular interval of time. Such restrictions may be acceptable in a traditional paper-and-pencil testing setting, but would not be workable within the through-course design we envision, as a critical feature for promoting formative uses of the assessments is the immediate, online availability of the passage read by each student, the questions posed, the response given, and the scores based on the application of the rubric.

One way to address this challenge is to consider using commissioned passages for some of the through-course assessment prompts. Some will argue that commissioned passages lack the authenticity and depth that are desired for the Common Core assessments. On the other hand, many state testing programs develop high quality ELA assessments that include commissioned passages. Commissioned passages are more economical to develop, do not require complex contractual negotiations, and can be targeted to meet specific needs related to content and complexity.

The good news in considering the challenge associated with developing our large proposed pool of through-course assessment prompts is that the PARCC consortium has the funding needed to support the substantial up-front development that this model requires. The payoff from a significant up-front effort is that the system can run for several years before any further development is needed. This would seem to be desirable given that it is unclear how the costs of the Common Core assessments will be sustained once the initial development funded by the federal government has run its course.

Further Supporting Formative Assessment

Although we believe it is possible for through-course assessments to be used formatively, it seems short-sighted to assume that through-course assessments in and of themselves will be sufficient to improve teaching and learning in the manner that reformists envision. Thus, the availability of assessment tools that can augment the information obtained from through-course
assessments will be an important component of Common Core assessment systems. For ELA, such tools may include diagnostic assessments that can drill down deeper into weaknesses that students display on the through-course assessments. In addition, assessment tools that provide prompts very similar to those on the through-course assessments but that include research-based scaffolds to assist students in achieving the skills and capacity needed to succeed on the through-course assessments will be critical resources for teachers to make use of. These tools may be developed by the consortia themselves, pulled from open-source content repositories, or offered as new products by vendors.

Perhaps more important in building out PARCC’s through-course assessment model, however, is providing appropriate opportunities and training to increase the assessment literacy of teachers. In particular, focused programs will be needed to train teachers and school administrators about the purpose of the through-course assessments, the connection between these assessments and the CCSS, the details of the scoring rubrics, and the connections of student performance to inferences about student growth and the goals of college and career readiness. Such training will serve to enhance the formative uses of the through-course assessments and better equip teachers to use complementary assessment strategies to impact student learning.

Through-Course Mathematics Assessments

In this paper we have concentrated on a through-course assessment solution for ELA that we believe fits within the parameters of the PARCC Common Core assessment plans. We have not explicitly addressed mathematics in this paper for several reasons. First, scope and sequence considerations are more prominent with mathematics instruction than ELA and present challenges for through-course Common Core assessments in mathematics that are not present in ELA. Second, whereas the emphasis on reading and writing to text sources of increasing complexity effectively frames the ELA through-course assessments, in the mathematics standards there is not a similar unifying spine. Rather, there are different domains within the mathematics standards, different clusters of standards within domains, and because mathematics is a connected subject, these different domains and different clusters within domains may in some cases be closely related and in others not at all related. Finally, with the ELA through-
course assessments, writing is the primary vehicle through which students construct their responses, and the technology and research base for automated scoring of writing seems—at least in our view—to be well enough established to support widespread operational applications. In mathematics, technology-enhanced problems may allow students to enter responses in ways that can be scored automatically, but the diversity of problems and responses that are conceivable for through-course assessments in mathematics suggests that the comprehensive use of automated scoring will not be feasible.

Despite these differences in how we envision through-course assessments playing out in the PARCC Common Core mathematics assessments, there are aspects of our proposed model for ELA that we believe can and should be retained for mathematics. These include: 1) the development of large numbers of problems that are available for each through-course assessment; 2) the random selection of problems from the operational pool to be administered to each student; 3) the disclosure of through-course assessment problems, student solutions, and scores; and 4) the assumption that scores on the through-course assessments are not equated but rather considered comparable under the assumption that the through-course assessment problems are randomly equivalent. We believe that these features are essential to promoting formative uses of through-course assessment results. We plan to consider in more depth the mathematics content issues related to scope and sequence as well as content sampling in a future paper on mathematics through-course assessment design.

Summary and Conclusions

In this paper we have presented a case for how through-course assessments as proposed by the PARCC consortium can most effectively support formative uses of assessment information. The educational reform movement in the U.S. has resulted in funding for Common Core assessments that are expected to serve—among other purposes—to inform teaching and learning within the schools. The intent of the through-course assessment components is to provide the assessment data that will realize this goal. Our proposed through-course assessment design serves to facilitate this purpose by making not only the results but the full set of assessment artifacts readily available to teachers, students, and other stakeholders in the assessment process. While this focus is decidedly aimed at supporting teaching and learning, we
believe our design will also provide data that are appropriate for summative purposes. Although there is clearly a significant up front cost to our proposed through-course design, there is also a simplicity that mitigates a plethora of psychometric issues that more traditional approaches would introduce. Given the through-course assessment model that the PARCC consortium has adopted, we are confident that our design can provide the needed balance to achieve both formative and summative purposes.
References


