Smarter Balanced Assessment Consortium:

Cognitive Laboratories Technical Report

Developed by: The American Institutes for Research

September 23, 2013

Executive Summary

The Smarter Balanced Assessment Consortium conducted cognitive laboratories to better understand how students solve various types of items. A cognitive laboratory uses a think-aloud methodology in which students speak their thoughts while solving a test item. The interviewer follows a standardized protocol to elicit responses and record what a student says. While this one-on-one process is time consuming, the type of information elicited is often difficult to obtain by other means. This report presents the results of a series of cognitive laboratory observational studies. The studies were conducted with small numbers of students in order to gather in-depth qualitative data about how students react to different types of items, formats, etc. Due to the small number of subjects studied and the ad hoc nature of the achieved sample of participants, the findings should be used to point the way to more systematic studies, rather than be cited as an authoritative source of scientific findings.

This executive summary presents the major findings from various protocols. Most protocols were developed at multiple grade bands (e.g., 3, 6, and 11). A grade band is the level of content for which the protocol is targeted. Protocols were usually targeted to answer a specific question in one or more content areas (e.g., ELA, mathematics). Results are organized under topics or questions of interest.

Summary and Findings of Cognitive Lab Results by Research Question

Research Question 1: Do mathematics multi-part selected-response (MPSR) items provide similar information about the depth of understanding by the test taker as do traditional constructed-response (CR) items?

An MPSR item has students select several examples of a correct response rather than just one, as in the typical selected-response (SR) item. The intention of this research question was to see whether the MPSR items provided depth of understanding similar to that provided by CR items. If effective, an MPSR item would be a more efficient way to measure the content measured by CR items. Within a form, parallel items were constructed in both formats and presented to the same students. In the protocols the MPSR and CR items were presented in random order.

This research question sought to address two hypotheses. The first hypothesis examined whether students who get full credit on MPSR items reveal, through their think-aloud sessions, greater understanding than those students who do not achieve full credit. The second hypothesis examined whether students who get full credit on MPSR items reveal depth of understanding similar to that of students who get full credit on similarly challenging CR items measuring the same target. In most cases the depth of knowledge (DOK) demonstrated by the student either equaled or exceeded the DOK demonstrated for the CR items.

Students who got full credit on the MPSR items also revealed greater understanding of the material than those who did not obtain full credit. The percentage of students understanding the material is also quite similar for the MPSR and CR items. A typical interviewer comment was, “based on the

accuracy of the student's responses to both types of items, it appears that item type is not a factor in determining how well the students respond[s]."

Research Question 2: Do TE item types and multi-part SR items approach the depth of knowledge of CRs?

The question is designed to assess whether different types of technology-enhanced (TE) items approach the DOK of CR items for specific content claim/targets and DOK levels. SR items were also included, where available, as a comparison item format. Comparisons were examined for specific TE item types at specific DOK levels for specific content claims/targets. CR and SR items were matched to specific content claims/targets and DOK 4 items in one of the three formats (SR, TE and CR) appeared in each form. Multiple forms were administered, each form to a different sample of students. It was hypothesized that students responding to items of a specific type would reveal that they are using thought processes consistent with a specific DOK level for items measuring a specific target. Different item types were administered to different students.

For ELA, students demonstrated a higher DOK level for most of the TE item types than for the matched CR items. Two exceptions were two targets in the "select text" item type: "justifying interpretations" (grade band 6) and "analyzing the figurative" (grade band 11). A similar pattern was observed for the matched SR items versus the CR items. The same TE item types had higher percentages than did the CR items, with the exception of the "select text" items for the "writing or revising strategies" target (grade band 7) and the "citing to support inferences" target (grade band 11).

For the SR items in ELA, the percentage receiving the maximum score was higher than both the CR and TE formats for the following "select text" items:

- "select text" for justifying interpretations, claim 1, DOK 2 in grade band 6
- "select text" for citing to support inferences, claim 1, DOK 2 in grade band 11
- "select text" for analyzing the figurative, claim 1, DOK 2 in grade band 11

For mathematics, the pattern is less clear. The TE item types that showed a higher percentage of students demonstrating thought processes consistent with the DOK level included:

- "placing points" for fractions, claim 1, DOK 2 in grade band 3
- "single lines" for equations and inequalities, claim 1, DOK 2 in grade band 11
- "tiling" for fractions, claim 1, DOK 2 in grade band 3
- "tiling" for equations and inequalities, claim 1, DOK 2 in grade band 11 ("Student indicated use of multiple steps and solved correctly.")
- "vertex-base quadrilaterals" for lines, angles, and shapes, claim 4, DOK 3 in grade band 4

The item types in which the CR items had a higher percentage of DOK-consistent thought processes included:

- "select and order" for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6

- "tiling" for everyday mathematic problems, claim 4, DOK 3 in grade band 4
- "tiling" for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6
- "tiling" for everyday mathematic problems, claim 2, DOK 3 in grade band 11
- "vertex-base quadrilaterals" for lines, angles, and shapes, claim 1, DOK 2 in grade band 4

The TE item types for which a higher percentage of students received full credit included only:

- "tiling" for equations and inequalities, claim 1, DOK 2 in grade band 11, and
- "vertex-base quadrilaterals" for lines, angles, and shapes, claim 1, DOK 2 in grade band 4.

In other cases the percentage of students receiving full credit was lower than for the comparable CR items. It should be noted that the percentage receiving full credit was generally low in mathematics. Even the matched SR items generally did not perform any better than either the CR or TE items.

Research Question 3: The Impact of Labeling on Mathematics Multi-Part Selected-Response (MPSR) Items: For multi-part selected response (MPSR) items where students may select more than one answer choice, which wording best indicates to the student that he or she is allowed to select more than one option? For multipart (e.g., YES/NO) dichotomous choice items, do students know that they need to answer each part?

Smarter Balanced sought to investigate whether students might become confused with MPSR items in mathematics and perhaps not complete the entire item. In order to investigate this, items were constructed with different amounts of labeling. *Labeling* is the identification of the parts of the problem with indicators such as "a," "b," "c" or "1," "2," "3." A labeled and a non-labeled condition were investigated. An example of an item in the labeled and unlabeled format can be found in Exhibit 2.

This question is designed to assess whether labeling or not labeling an MPSR mathematics item produces a difference in performance. Results are reported in five grade bands. Each form contains one MPSR item followed by one CR item. The labeled and non-labeled items appeared in different forms of the test and thus were taken by different students.

Even though the labeling of MPSR items was intended to clarify the mathematic tasks for the students, in many cases it actually seemed to confuse the students. Little difference was observed between the labeled and non-labeled items in the lower grade bands (grade bands 3–6). However, students in grade band 7 tended to score higher with non-labeled items. Also, grade band 7 and 11 students tended to be confused by the labeling. In addition, the labeled items tended to receive more comments related to not understanding the instructions. The interviewer confirmed this, suggesting that the grade band 7 and 11 students better understood the instructions in the non-labeled condition than in the labeled condition.

Research Question 4: Does the ability to move one or more sentences to different positions provide evidence of students' ability to revise text appropriately in the consideration of chronology, coherence, transitions, or the author's craft?

Smarter Balanced is considering using items that have students reorder sentences to measure an editing/revising standard. Claim 2 of the standards states that students should be able to revise one

or more paragraphs demonstrating specific narrative strategies (use of dialogue, sensory or concrete details, description), chronology, appropriate transitional strategies for coherence, or authors' craft appropriate to the purpose of the item (closure, detailing characters, plot, setting, or an event).

This question was designed to assess whether students' movement of one or more sentences to different positions provided evidence of students' ability to demonstrate consideration of chronology, coherence, transitions, or author's craft. Six ELA items were included in a test form.

Students who performed well on the items were more likely to consider the targeted writing skills (consider chronology, coherence, transitions, and author's craft) when answering the questions. Also, students who made appropriate sentence moves were more likely to consider the targeted writing skills than those who made inappropriate sentence moves. A high percentage of students considered chronology, coherence, and transitions; however, they were less likely to consider author's craft.

Research Question 5: Do Students Who Construct Text Reveal More Understanding of Targeted Writing Skills Than Students Who Manipulate Writing Through the Manipulation of Text (MT) Tasks?

Many believe that the best way to measure writing is to have students write. However, in a testing environment, it is often difficult to adequately sample the writing content domain with an assessment composed exclusively of CR items. An effort is ongoing to find items that are efficient but that can adequately measure the components of the writing domain, thus allowing a broader selection and greater number of items to be delivered. The question examines whether comparable understanding of the targeted writing skills can be achieved using a set of MT tasks in comparison to comparable CR tasks. Examples of the item types can be found in Exhibit 2.

Four pairs of ELA items were developed. Each pair contained one MT item and one CR version of the same item. Two forms were created, and each form contained a single version of an item. Each form contained two MT items and two CR items. The MT items were almost exclusively "select and order" items, though two items—one in grade band 3 and one in grade band 11—were "reorder text" items. All items assessed claim 1, target 1.

The results showed that the targeted writing skills are considered by students who manipulate text at a level comparable to (or greater than) that encountered when they are constructing text. The grade band 3 and 6 students showed comparable (or greater) levels of understanding when the items were in an MT format. For the grade band 11 students the results were mixed, but students tended to be more effective in applying the targeted writing skills in the CR format, particularly for transitions and author's craft. Score distributions were comparable for MT and CR item formats.

Research Question 6: Do different types of directions (minimal, concise or extensive) have an effect the performance of technology enhanced (TE) items in ELA and Mathematics?

The optimal amount of direction that should be given to a student working with TE items is unclear. With minimal directions students may not know how to approach an item; with extensive directions students may be distracted or slowed to a point where the item becomes inefficient. This may be particularly true with elementary school students, who may take longer to process text. This question

examined this issue for ELA and mathematics items. Three types of directions were used (minimal, concise, and extensive).

In most cases in ELA the level of instruction did not make a difference. For most grade bands and item types, neither the level of instruction nor the item type showed a differential effect in ELA. Cases in which differences were observed included "select text" items when the directions were "concise." With the "reorder text" items the grade band 3 students did less well with minimal directions. The grade band 11 students also had some difficulty with the "reorder text" items when the directions were "extensive."

In mathematics, the level of instruction also did not make a difference for many item types and grade bands. "Select and order" items were difficult (grade bands 6 and 11) regardless of the direction type, however, no direction type proved better than another. High percentages of students received full credit on "select defined partition" and "straight lines" items; however, the direction type did not make a difference. Finally, "tiling" items were generally difficult, but no benefit was shown for different types of directions. Differences were observed in items including "placing points" items under the minimal and concise directions in grade band 11; however, under extensive directions all students received the maximum score. With "placing points and tiling" items a higher percentage of students received full credit with fewer instructions (grade band 6). Finally, "vertex-based quadrilateral" items seemed to benefit from minimal directions in grade band 11.

When asked if they had difficulty using the computer, ELA students, in grade band 3, under minimal directions, said they had trouble with both select text and reorder text items. The ELA grade band 11 students also seemed to have some difficulty with the "reorder text" items. Since these are related to specific item types it suggests that there was uncertainty about how to perform the task, rather than using the computer itself. Mathematics students did not seem to have any problems using the computer.

Research Question 7: Smarter currently intends to administer the passage first, and then administer the items one item at a time. Does this affect student performance?

Smarter Balanced is interested in the possibility of administering items adaptively within a passage. This would require administering items sequentially so that the ability estimate could be updated after each item. Presenting items one at a time may take longer, and students may object to not knowing what is coming next. This question is designed to assess whether administering an item set takes longer when the items are presented sequentially and whether there is a difference in confusion or frustration level when students are presented a passage and all the items together or are presented a passage with the items then being presented one at a time. The item sets were not administered adaptively.

Two sets of items were created for a given test form. Both sets contained passages of equivalent length and difficulty as well as items of equivalent difficulty.[1] The first set in a form presented the passage with all the items together. The second set presented the passage with the items presented one at a time.

The forms were administered, within grade band, to different samples of students. Each sample contained both a general education group (Gen Ed) and a group that received English language accommodations (ELL) students. One sample was timed without thinking aloud during the administration. Each item set in these forms was separately timed. This sample provided timing information only. The second sample involved thinking aloud while responding to the questions and was not timed.

The primary questions of interest were:

1. Does presenting the items individually after the passage appear to take longer (timed condition)?
2. Does presenting the items individually after the passage increase the student's negative emotional states (e.g., frustration, confusion; think-aloud condition)?
3. Do students prefer one approach or another (think-aloud condition)?

The time it took to complete the sets when all items were presented together or one at a time varied by grade band and sample. For the grade band 3 and grade band 11 samples, timing differed little whether the items were presented at once or one at a time. However, for grade band 6, presenting the items one at a time took substantially longer for both the Gen Ed and ELL samples. While there is some variability between the ELL and the Gen Ed samples, the differences are not large and show the same pattern within grade band.

There appears to be slightly more *confusion* for both the Gen Ed and the ELL samples in grade band 3 when all the items are presented together. However, similar *frustration* levels were observed under the two formats for the grade band 3 students. Students working on the grade band 6 ELL sample showed similar patterns of frustration and confusion in both presentation formats. However, the Gen Ed grade band 6 students showed slightly more confusion when the items were presented one at a time.

[1] Comparable passage difficulty was achieved through the use of readability and lexile measures. Comparable item difficulty was achieved through DOK measures.

The grade band 6 students tended to score higher when the items were presented all at once (for both the Gen Ed students and the ELL students). The grade band 3 students showed similar results, regardless of sample or administration format. The grade band 11 Gen Ed students scored higher when the items were presented one at a time, while the grade band 11 ELL sample students scored higher when the items were presented altogether.

Both the ELL and Gen Ed grade band 3 students preferred to have the items presented one at a time. Grade band 11 students had a slight bias toward having the items presented one at a time. Conversely, grade band 6 students preferred to have the items presented together.

Research Question 8: Smarter intends to present relatively long passages. Do longer passages reduce student engagement?

Smarter Balanced is interested in using passages that are longer than those presently used. The Smarter Balanced recommended passage lengths are: for grades 3–5: 450–562 words for short passages and 563–750 words for long passages; for grades 6–8: 650–712 words for short passages and 713–950 words for long passages; and for high school, 800–825 words for short passages and 826–1100 words for long passages. There is concern that the longer passages may tax the processing abilities of ELL students and students with disabilities (SWD).

This question is designed to assess whether longer passages reduce student engagement, hamper the completion of the longer passages, or affect the depth of processing of the passage. Two sets of items were created. Both sets contained passages of equivalent difficulty with four items of equivalent difficulty attached to each passage. Both sets present the passage and all the items together. Each form contained a standard-length passage and an extended-length passage. The first set contained a passage of standard length. The second set contained a passage that is longer than standard length (extended-length, the length equivalent to that intended for use by Smarter Balanced).

The design was intended to compare the performance of two groups of students—ELL/SWD and Gen Ed students—across three grade bands: 3, 6, and 11. Twelve students took the forms. Of these, nine were grade band 3 Gen Ed students and one grade band 3 student was classified ELL/SWD. The single grade band 6 student was an ELL/SWD student. The two grade band 11 students were Gen Ed students.

All the ELL/SWD students were unaffected by the use of the longer passage. They were able to read the entire passage regardless of passage length and demonstrated that the longer passage was processed at a deep level. The ELL/SWD students also were not bored or distracted while reading either passage.

On the contrary, Gen Ed students did appear to be affected by the longer passage in grade bands 3 and 11. About 75 percent of the grade band 3 students and all of the grade band 11 students were affected by the use of the longer passage. Only 43 percent of the grade band 3 Gen Ed students and 50 percent of the grade band 11 Gen Ed students demonstrated a level of deep processing. Also, some percentage of the Gen Ed students were bored, regardless of the length of the passage

Research Question 9: How long does it take for students to read through complex texts, performance tasks, etc.? Is timing affected by the way students are presented the passage and items?

One way of making items more difficult is to increase their complexity. Complex items often take longer to solve or answer. In computer adaptive tests, added complexity may decrease the time a high ability student has to complete the test if the items are made more difficult through increased complexity. This potentially creates some fairness issues in an adaptive test if there is a time limit on the test. This question was designed to assess the time it takes for students to answer complex and simpler items. Complexity was defined as a function of the DOK demanded by the test question. It was hypothesized that more complex tasks would take more time.

Each ELA form had six items. These items varied in item complexity (simple or complex) and item format (SR, TE, or CR). The TE items were all "hot text" (HT) items. These items require the student to either highlight the text or drag the text to answer the item.

Forms were constructed in ELA at two grade bands: grade band 3–5 (referred to as grade band 3) and grade band 6 and 7 (referred to as grade band 6). Two forms were administered in grade band 3. One form was administered in grade band 6.

It was hypothesized that more complex items would take longer to complete than simpler items, but no evidence was found to support this hypothesis. SR items were answered in the shortest time. HT items took about one minute longer than SR items. CR items took the most time to answer, about 75 seconds longer then the hot text items.

Research Question 10: Working mathematics problems on computer: Communicating mathematics on computer—feasibility of measuring student understanding of items for Claims 2–4 on computer.

With paper tests some students write in their test books while working out mathematics problems. When mathematics items are presented on computer, scratch paper is often provided if students want to transfer the problem to paper and work it out there. Because scratch paper is often destroyed after an online testing session, the degree to which scratch paper is used is not known; neither is the importance of scratch paper in working out a problem (or potentially for use in scoring). This research question examines the need for paper when solving mathematics problems.

Each student was presented with three grade-appropriate items. The interviewer recorded whether the student made a comment, and the nature of the comment, while working the mathematics problems. The students first tried to work a problem without paper. Scratch paper was then offered to the student to rework the problem, if desired. The interviewer noted whether students chose to add anything additional and noted the nature of the addition (more text, equations, graphics). Note that there were only three comments for the third item in the lowest grade band, 3.

The general conclusion is that a subset of students benefit from being able to work mathematics problems on paper. This appears to be especially important when students are beginning to learn algebra concepts.

Grade band 3 students did not need paper to work the problems. However, in the grade band 6 and grade band 7 groups, 30–42 percent indicated they wanted to write an equation. In grade bands 6, 7, and 11, the additional information recorded on paper would have improved the response according to the rubric. Responses for specific items in grade bands 6 and 11 were improved by 15 percent of the students, and responses for all items in grade band 7 were improved when information on the scratch paper was taken into account. Improvement for this group ranged between 10 and 20 percent of the responses. ("Confused me, I didn't know how to write an equation." "Tried the keypad, but it wouldn't work." "It was much easier with paper.") This was supported by interviewer observations. About 5–10 percent of students in each grade band found the online system difficult to use, but few specifics were recorded.

Research Question 11: Usability of equation editor tool—can students use the tool the way it is meant to be used?

Although students begin to use technology at a very early age, it is prudent to verify that young students are able to use the assessment interface to be used during testing. This question sought to evaluate the ability of grades 3–5 students to use the equation editor tool to be included in the Smarter Balanced delivery system. Three mathematics items were presented to the students (*N*=33). The first item only required the student to copy his or her response. The second item was a simple mathematics item, and the third item was a more challenging mathematics item. The first item would demonstrate whether the student could use the equation editor tool. The second and third items would provide evidence of whether the ability to use the tool interacted with item difficulty.

Elementary students had some difficulty using the equation editor. Between 15 and 30 percent of the students indicated that they had difficulty using the equation editor. The examiner's assessment concurred that about 35 percent of students had difficulty using the equation editor and that about 50 percent of the students would get a given item correct.

Research Question 12: Can students compare the size of a product to the size of one factor, on the basis of the size of the other factor, without performing the indicated multiplication?

This question is designed to assess whether students with a strong understanding of fractions and the multiplication and division of fractions complete the items without performing the indicated multiplication. The task asked students to compare the size of a product to the size of one factor, on the basis of the size of the other factor, without performing the indicated multiplication. Also of interest was whether students who complete an item as intended (without using multiplication) spent less time on an item than those who did not. To investigate this question a single form was administered for grades 3–5.

There seemed to be little relationship between whether a student has a strong understanding of the multiplication and division of fractions and whether he or she used multiplication to solve the items. However, students who did not need to perform the multiplication completed the items in less time than students who had to perform the multiplication. While most students said they understood the questions, 70 percent had to use multiplication to solve them. Only about 40 percent of the students had a firm understanding of the multiplication/division of fractions, according to the interviewers.

Research Question 13: Contextual glossaries are item-specific glossaries that provide a definition of a word that is targeted to, and appropriate for, the context in which the word is used in the item. Are these a fair and appropriate way to support students who need language support?

This question addressed the efficacy of the use of contextual glossaries with non-native speakers when solving mathematics problems. Two sets of items were created that were parallel in difficulty. The first set of items contained no contextual glossaries with only single words translated. The second set of items contained contextual glossaries. The interviewer was asked to determine whether the student was having trouble understanding a word and whether the contextual glossary aided in the interpretation of the word or sentence.

Only three ELL students participated: one from grade 3 and two from grade 6.

The contextual glossaries appeared to be somewhat effective, but the impact was not always reflected in the score the student received for an item. The contextual glossaries appeared to be incomplete in that they did not include words the student needed. This limited the use of the glossaries in these situations. Interviewer's comments suggested that performance was improved when the students used the contextual glossaries.

Research Question 14: Under what conditions does the use of text-to-speech (TTS) help students with lower reading ability focus on content in ELA and mathematics?

TTS can provide access to an assessment for students with low reading ability. In order for this technology to be effective the language produced from the voice-pack must be clear enough to be understood. This is particularly true for non-native speakers of English.

Only students familiar with TTS were included in the study. Overall, 77 students used TTS at least once. Among them, 58 students were limited English proficient (LEP), 13 students had reading difficulties (IEP), and six were Gen Ed students.

In ELA four forms were administered with both high- and low-quality voice-packs. In mathematics, two forms were administered in grade bands 3 and 11. Only a single form was administered in grade band 6. The mathematics forms were only administered with high-quality voice-packs.

TTS improved access in ELA regardless of the quality of the voice-pack. Greater access was achieved when high-quality voice-packs were used. LEP students and students with reading difficulties tended to benefit more from the use of TTS. Using TTS with high-quality voice-packs improved focus on content in ELA. The use of TTS with low-quality voice-packs tended to distract students in ELA, whereas high-quality voice-packs did not. In mathematics, access was improved only for grade band 3 students. All Gen Ed, IEP, and grade band 6 LEP students found the high-quality voice-pack distracting. This was in part a function of trying to describe a table verbally.

Introduction

Smarter Balanced has conducted cognitive laboratories to better understand how students solve items in different formats. A cognitive laboratory uses a think-aloud methodology in which students speak their thoughts while solving a test item. The interviewer follows a standardized protocol to elicit responses and record what a student says. While this one-on-one process is time consuming, the type of information elicited is often difficult to obtain by other means. Due to the nature of the process the sample sizes are often small; however, they are sufficient to detect large effects. In addition, because each student's comments are recorded, smaller, non-primary effects may be brought to light. Most protocols were developed at multiple grade bands (e.g., 3, 6, and 11). A grade band is the level of content for which the protocol is targeted.

What follows are in-depth analyses for each research question outlined in the executive summary. Because of the differences in the samples, study design, and questions asked, each research question result is presented separately. A summary of the findings for each research question is provided at the end of each research question section. Research questions have been organized into sections of similar content to improve integration of the material. Finally, a conclusions section appears at the end of the document. The overall demographics for the cognitive labs sample can be found in Appendix B.

Processing Selected-Response (SR), Technology-Enhanced (TE), and Constructed-Response (CR) Items

Research Question 1: Do mathematics multi-part selected-response (MPSR) items provide information about the depth of understanding of the test taker similar to traditional constructed-response Items?

An MPSR item has students select several examples of a correct response rather than just one, as in the typical SR item. The intention of this research question was to see whether the MPSR items provided depth of understanding similar to that of CR items. If effective, an MPSR item would be a more efficient way to measure the content measured by CR items. Also of interest was whether similar results would be obtained at different educational levels. To investigate these questions, forms were constructed at four grade bands: grades 3–4 (referred to as grade band 3), grades 6–7 (referred to as grade band 6), grades 7–8 (referred to as grade band 7), and grade 9–10 (referred to as grade band 11). Within a form, parallel items were constructed in both formats and presented to the same students. In the protocols the MPSR and CR items were presented in random order. In the tables below the SR and CR data for each item are presented adjacent to each other to facilitate comparisons between the two item formats.

Interviewers were asked to assess the highest level of DOK the student demonstrated during the think-aloud session. Table 1 (ELA) and Table 2 (mathematics) show the rubrics the interviewers used during this process.

Two hypotheses related to research question 1 were examined. The first hypothesis examined whether students who get full credit on MPSR items reveal, through their think-aloud sessions, greater understanding than those students who do not achieve full credit. The second hypothesis

examined whether students who get full credit on MPSR items reveal understanding similar to that of students who get full credit on similarly challenging CR items measuring the same target.

Table 1. Depth of Knowledge Chart (ELA)

DOK Level	Definition	Types of statements
1	Recall and Reproduction	1. Recalls facts, details, and events 2. Uses word relationships (synonym/ antonym) to determine meaning 3. Recognizes or retrieves information from tables and charts
2	Basic Skills and Concepts	1. Summarizes information 2. Identifies central ideas 3. Uses context to determine word meanings 4. Analyzes text structure and organization 5. Compares literary elements, facts, terms, or events
3	Strategic Thinking and Reasoning	1. Uses supporting evidence to explain, generalize, or connect ideas 2. Analyzes or interprets author's craft (literary devices, viewpoint, potential bias) to critique a text 3. Develops a logical argument and cites evidence

Table 2. Depth of Knowledge Chart (Mathematics)

DOK Level	Definition	Types of statements
1	Recall and Reproduction	I remembered it. We learned the answer in class. I did what it said. I recognized it.
2	Basic Skills and Concepts	1. Any statement indicating putting two or more pieces of knowledge together 2. An statement indicating that they executed a sequence of steps that was not given to them 3. Any inference relating two different things 4. Expression of a hypothesis or guess about a relationship
3	Strategic Thinking and Reasoning	1. Any statement indicating that they are applying abstract concepts to concrete phenomenon, e.g., "Both patterns reflect exponential growth" 2. Statements indicating that the students evaluated several different approaches to solving the problem, accompanied by the ability to explain why they selected the solution path they chose 3. Explanations of their choices or decisions using data and information from multiple sources to construct a coherent and logical argument

Results

Twenty students were administered the grade band 3 form, 37 students were administered the grade band 6 form, 31 students were administered the grade band 7 form, and 19 students were administered the grade band 11 form.

Table 3 presents the average demonstrated DOK level by students who received full credit on an item for each grade band/target. Table 4 shows the correspondence between the target labels and the full target description. Blank cells are the result of incomplete data, either in the score or in the demonstrated DOK. In most cases the DOK the student demonstrated either equals to exceeds the DOK demonstrated for the CR items. Interviewers commonly commented that the student did equally well on both item formats.

Table 3. Average DOK Demonstrated by Students Who Received Full Credit for Paired MPSR and CR Items Measuring the Same Target

Grade Band	Target	Item Format	Avg. DOK
3	Geometric Measurement: Perimeters (J)	MPSR	2.00
		CR	1.50
3	Reason with Shapes (K)	MPSR	1.80
		CR	1.67
6	One Variable Equations (F)	MPSR	1.57
		CR	1.60
6	Analyze Proportional Relationships (A)	MPSR	
		CR	1.25
6	Generate Equivalent Expressions (C)	MPSR	2.00
		CR	2.00
6	Apply Arithmetic to Algebra (E)	MPSR	1.60
		CR	2.00
7	Analyze Proportional Relationships (A)	MPSR	
		CR	1.83
7	Generate Equivalent Expressions (C)	MPSR	1.77
		CR	2.00
7	Solve Linear Equations (D)	MPSR	2.00
		CR	1.80
11	Equivalent Problem Solving (E)	MPSR	2.33
		CR	1.75
11	Graph Equations and Inequalities (J)	MPSR	
		CR	1.70
11	Use of Functions (K)	MPSR	2.10
		CR	2.00

Table 4. Correspondence Between Target Label and the Full Target Description

Target Label	Full Target Description
Geometric measurement: Perimeters	Geometric measurement: recognize perimeter as an attribute of plane figures and distinguish between linear and area measures
Reason with Shapes	Reason with shapes and their attributes
Place Value: Whole Numbers	Generalize place value understanding for multi-digit whole numbers
Converting Units of Measure	Solve problems involving measurement and conversion of measurements from a larger unit to a smaller unit
Geometric measurement : Perimeters	Geometric measurement: recognize perimeter as an attribute of plane figures and distinguish between linear and area measures
One Variable Equations	Reason about and solve one-variable equations and inequalities
Apply Arithmetic to Algebra	Apply and extend previous understandings of arithmetic to algebraic expressions
Generate Equivalent Expressions	Use properties of operations to generate equivalent expressions
Analyze Proportional Relationships	Analyze proportional relationships and use them to solve real-world and mathematical problems
Solve Linear Equations	Analyze and solve linear equations and pairs of simultaneous linear equations
Equivalent Problem Solving	Write expressions in equivalent forms to solve problems
Graph Equations and Inequalities	Represent and solve equations and inequalities graphically
Use of Functions	Understand the concept of a function and use function notation

The second hypothesis examined whether students who get full credit on the MPSR items reveal greater understanding of the material than those who do not obtain full credit. Table 5 presents these findings. In all cases those who receive full credit for an item showed greater understanding than those who did not receive full credit. The percentage understanding is also quite similar for the MPSR and CR items.

Table 5. Percentage of Students Who Appear to Understand the Material, by Item Type, Grade Band, and Whether Full Credit Was Received

Grade Band								
	3		6		7		11	
Item	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit
MPSR1	20	50	17	89	38	78	64	-
CR1	12	100	25	100	55	-	40	100
MPSR2	0	57	29	100	42	83	45	100
CR2	7	75	23	90	36	75	67	67
MPSR3	0	-	10	67	48	100	33	75
CR3	0	-	8	100	50	75	58	67

Summary

This research question sought to address two hypotheses. The first hypothesis examined whether students who get full credit on MPSR items reveal, through their think-aloud sessions, greater understanding than those students who do not achieve full credit. The second hypothesis examined whether students who get full credit on MPSR items reveal depth of understanding similar to that of students who get full credit on similarly challenging CR items measuring the same target. In most cases the DOK the student demonstrated either equaled or exceeded the DOK demonstrated for the CR items.

Students who got full credit on the MPSR items also revealed greater understanding of the material than those who did not obtain full credit. The percentage of students understanding the material is also quite similar for the MPSR and CR items. A typical interviewer comment was, “based on the accuracy of the student’s responses to both types of items, it appears that item type is not a factor in determining how well the students respond[s].”

Research Question 2: Under what conditions do specific types of TE items (and SR items) approach the depth of knowledge (DOK) of a written constructed response in ELA and mathematics?

The question is designed to assess whether different types of TE items approach the DOK of CR items for specific content claim/targets and DOK levels. SR items were also included, where available, as a comparison item format. Comparisons were examined for specific TE item types at specific DOK levels for specific content claims/targets (see Appendix A for a full description of the claims and targets). Where possible, parallel items were created in each item format at the same DOK level and content claim/target; however, some combinations were not available. In ELA, items in the different formats were administered for most item type/content target/DOK combinations. In mathematics, however, some item formats were not administered for all claim/target/DOK conditions and some data were incomplete. This limited the comparisons that could be made. Four items in one of the three formats (SR, TE, and CR) appeared in each form. Multiple forms were administered, each to a different sample of students. It was hypothesized that students responding to items of a specific TE type would reveal that they are using thought processes consistent with a specific DOK level for items measuring a specific target.

Forms were constructed in ELA at five grade bands: grade 3 (referred to as grade band 3), grades 4–5 (referred to as grade band 4), grades 6–7 (referred to as grade band 6), grades 7–8 (referred to as grade band 7), and grade 11 (referred to as grade band 11). In mathematics, forms were constructed at four grade bands: grades 3–4 (referred to as grade band 3), grades 4–5 (referred to as grade band 4), grades 6–7 (referred to as grade band 6), and grade 11 (referred to as grade band 11). Note that the grade band relates to the level of the material in the assessment and not necessarily the grade of the students to which the assessment is administered. A single form was administered in each grade band. This was a between-subjects design in which different item types were administered to different students. For this question, the comments presented are made by the interviewer, as opposed to the student, due to the nature of the information being captured (e.g., DOK level demonstrated).

Results

Table 6 shows the sample sizes within a grade band by item format across item types and content area. The ELA forms tended to have been administered to larger samples than were the mathematics forms.

Table 6. Sample Sizes Within Grade Band, by Content Area and Item Type

		Grade Band				
Content	Item Format	3	4	6	7	11
ELA	SR	18	16	13	8	6
ELA	TE	12	14	10	8	14
ELA	CR	14	13	13	15	10
Mathematics	SR	7	6	23	-	10
Mathematics	TE	7	4	13	-	3
Mathematics	CR	4	11	8	-	3

Tables 7a (ELA) and 7b (Mathematics) list the percentage of students whose thought processes were consistent with the DOK level of the items for the respective content areas. For each TE item type, the percentage of students who demonstrated thought processes consistent with the grade band/content claim and target/DOK was recorded. SR and CR items were matched to the same grade band/content claim and target/DOK levels. The primary comparison of interest is between the TE and CR formats.

For ELA, students demonstrated a higher DOK level for most of the TE item types than for the matched CR items. (“Well thought out. Uses evidence she feels supports the main idea of the item.”) Two exceptions were two targets in the “select text” item type: “justifying interpretations” (grade band 6) and “analyzing the figurative” (grade band 11). A pattern similar to that of the TE item types was observed for the matched SR items versus the CR items.

Table 7a. Percentage of Students Demonstrating That They Are Using Thought Processes at the Specified DOK level, by Item Type, Claim, Target, and DOK Level (ELA)

					% of Students With Consistent Thought Process		
TE Item Type	Grade Band	Target	Claim	DOK	TE	SR	CR
Drag and Drop (Tiling)	6	Justifying interpretations (11)	1	3	63	78	40
Drag and Drop (Tiling)	7	Writing or revising strategies (6)	2	2	100	80	61
Reorder Text	3	Writing or revise strategies (3)	2	2	81	69	54
Reorder Text	6	Organizing ideas (3)	2	2	60		
Select Text	6	Justify interpretations (11)	1	2	33	50	60
Select Text	7	Identifying text to support inferences (1)	1	2	94	79	64
Select Text	7	Writing or revising strategies (6)	2	2	100	80	61
Select Text	11	Citing to support inferences (1)	1	2	72	82	69
Select Text	11	Analyzing the figurative (7)	1	2	33	50	55

For mathematics, the pattern is less clear. The TE item types that yielded a higher percentage of students demonstrating thought processes consistent with the DOK level included:

- "placing points" for fractions, claim 1, DOK 2 in grade band 3 ("This student had a thorough understanding of these fractions and how they related to the number line. He thoroughly and accurately placed each fraction and explained how/why using various steps.")
- "single lines" for equations and inequalities, claim 1, DOK 2 in grade band 11
- "tiling" for fractions, claim 1, DOK 2 in grade band 3 ("This student clearly understood and explained how to solve this item using multiple methods. He used multiple steps to solve each item.")
- "tiling" for equations and inequalities, claim 1, DOK 2 in grade band 11 ("Student indicated use of multiple steps and solved correctly.")
- "vertex-base quadrilaterals" for lines, angles, and shapes, claim 4, DOK 3 in grade band 4

Places where equal percentages were observed for the TE and CR formats included:

- "select and order" for fractions, claim 1, DOK 2 in grade band 3
- "select and order" for fractions, claim 1, DOK 2 in grade band 6
- "selecting points" for fractions, claim 1, DOK 2 in grade band 3
- "single lines" for everyday math problems, claim 2, DOK 2 in grade band 11

Item types for CR items yielding a higher percentage of students who demonstrate consistent thought processes included:

- "select and order" for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6
- "tiling" for everyday mathematic problems, claim 4, DOK 3 in grade band 4
- "tiling" for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6 (The student was able to explain his answer in multiple steps and with a clear understanding of the distributive property.")
- "tiling" for everyday mathematic problems, claim 2, DOK 3 in grade band 11
- "vertex-base quadrilaterals" for lines, angles, and shapes, claim 1, DOK 2 in grade band 4 ("This student understood right angles. She also understood that she had to name a similarity and a difference.")

Table 7b. Percentage of Students Demonstrating That They Are Using Thought Processes at the Specified DOK Level, by Item Type, Claim, Target, and DOK Level (Mathematics)

					% of Students With Consistent Thought Process		
TE Item Type	Grade Band	Target	Claim	DOK	TE	SR	CR
Placing Points	3	Fractions (F)	1	2	50	53	0
Select and Order	3	Fractions (F)	1	2	0	53	0
Select and Order	6	Apply arithmetic to algebraic expressions (E)	1	2	40	67	79
Select and Order	6	Everyday math problems (A)	2	3	50		
Selecting Points	3	Fractions (F)	1	2	0	53	0
Single Lines	11	Equations and inequalities (I)	1	2	50	82	42
Single Lines	11	Everyday math problems (A)	2	2	100		100
Tiling	3	Fractions as numbers (F)	1	2	71	53	0
Tiling	4	Everyday math problems (A)	4	3	0	33	52
Tiling	6	Apply arithmetic to algebraic expressions (E)	1	2	60	67	79
Tiling	11	Equations and inequalities (I)	1	2	50	82	42
Tiling	11	Everyday math problems (A)	2	3	0	39	50
Vertex-Based Quadrilaterals	4	Lines, angles, and shapes (A)	4	3	67	33	52
Vertex-Based Quadrilaterals	4	Lines, angles, and shapes (L)	1	2	33	83	72

Also of interest was how students performed on these item types. Since not all items are 1-point items, the percentage obtaining the maximum score was used. Table 8a presents this information for ELA; Table 8b presents this information for mathematics. In ELA, the pattern is very similar to the consistency of thought process table. The same TE item types had higher percentages than the CR

items, with the exception of the "select text" items for the "writing or revising strategies" target (grade band 7), and the "citing to support inferences" target (grade band 11).

For the SR items in ELA, the percentage receiving the maximum score was higher than both the CR and TE formats for the following "select text" items:

- "select text" for justifying interpretations, claim 1, DOK 2 in grade band 6
- "select text" for citing to support inferences, claim 1, DOK 2 in grade band 11
- "select text" for analyzing the figurative, claim 1, DOK 2 in grade band 11

Table 8a. Percentage of Students Receiving Full Credit for an Item (ELA)

					% of Students Who Answered Correctly		
TE Type	Grade Band	Target	Claim	DOK	TE	SR	CR
Drag and Drop (Tiling)	6	Justifying interpretations (11)	1	3	80	67	18
Drag and Drop (Tiling)	7	Writing or revising strategies (6)	2	2	67	22	47
Reorder Text	3	Writing or revise strategies (3)	2	2	64	0	44
Reorder Text	6	Organizing ideas (3)	2	2	12		
Select Text	6	Justifying interpretations (11)	1	2	10	70	25
Select Text	7	Identifying text to support inferences (1)	1	2	77	19	41
Select Text	7	Writing or revising strategies (6)	2	2	0	22	47
Select Text	11	Citing to support inferences (1)	1	2	22	67	40
Select Text	11	Analyzing the figurative (7)	1	2	8	46	31

In mathematics, the TE items in which a higher percentage of students received the maximum possible score included only:

- "single lines" for equations and inequalities, claim 1, DOK 2 in grade band 11
- "tiling" for equations and inequalities, claim 1, DOK 2 in grade band 11
- "vertex-base quadrilaterals" for lines, angles, and shapes, claim 1, DOK 2 in grade band 4

For the SR items in mathematics, the percentage receiving the maximum score was higher than both the CR and TE formats for the following items:

- "placing points" for fractions, claim 1, DOK 2 in grade band 3
- "select and order" for fractions, claim 1, DOK 2 in grade band 3
- "selecting points" for fractions, claim 1, DOK 2 in grade band 3
- "tiling" for fractions, claim 1, DOK 2 in grade band 3
- "tiling" for everyday math problems, claim 4, DOK 3 in grade band 4

- “vertex-base quadrilaterals” for fraction equivalence and ordering, claim 1, DOK 2 in grade band 3

In other cases the percentage receiving the maximum score was lower than for the comparable CR items. It should be noted that the percentage receiving the maximum scores was generally low in mathematics.

Table 8b. Percentage of Students Receiving Full Credit for an Item (Mathematics)

					% of Students Who Answered Correctly		
TE Type	Grade Band	Target	Claim	DOK	TE	SR	CR
Placing Points	3	Fractions (F)	1	2	17	35	25
Select and Order	3	Fractions (F)	1	2	14	35	25
Select and Order	6	Everyday math problems (A)	2	3	0		
Select and Order	6	Apply arithmetic to algebraic expressions (E)	1	2	0	21	44
Selecting Points	3	Fractions (F)	1	2	14	35	25
Single Lines	11	Everyday math problems (A)	2	2	0		80
Single Lines	11	Equations and inequalities (I)	1	2	33		27
Tiling	3	Fractions (F)	1	2	33	35	25
Tiling	4	Everyday math problems (A)	4	3	0	25	5
Tiling	6	Apply arithmetic to algebraic expressions (E)	1	2	31	21	44
Tiling	11	Everyday math problems (A)	2	3	33	16	40
Tiling	11	Equations and inequalities (I)	1	2	33	0	27
Vertex-Based Quadrilaterals	3	Fraction equivalence and ordering (F)	1	2	21	35	25
Vertex-Based Quadrilaterals	4	Lines, angles, and shapes (A)	4	3	0	25	5
Vertex-Based Quadrilaterals	4	Lines, angles, and shapes (L)	1	2	67	50	0

Summary

For ELA, students demonstrated a higher DOK level for most of the TE item types than for the matched CR items. Two exceptions were two targets in the “select text” item type, “justifying interpretations” (grade band 6) and “analyzing the figurative” (grade band 11). A similar pattern was observed for the matched SR items versus the CR items. In ELA, the pattern is very similar to the consistency of thought process table. The same TE item types had higher percentages than did the

CR items, with the exception of the "select text" items for the "writing or revising strategies" target (grade band 7) and the "citing to support inferences" target (grade band 11).

For the SR items in ELA, the percentage receiving the maximum score was higher than both the CR and TE formats for the following "select text" items:

- "select text" for justifying interpretations, claim 1, DOK 2 in grade band 6
- "select text" for citing to support inferences, claim 1, DOK 2 in grade band 11
- "select text" for analyzing the figurative, claim 1, DOK 2 in grade band 11

For mathematics, the pattern is less clear. The TE item types that showed a higher percentage of students demonstrating consistent thought process with the DOK level included:

- "placing points" for fractions, claim 1, DOK 2 in grade band 3
- "single lines" for equations and inequalities, claim 1, DOK 2 in grade band 11
- "tiling" for fractions, claim 1, DOK 2 in grade band 3
- "tiling" for equations and inequalities, claim 1, DOK 2 in grade band 11 ("Student indicated use of multiple steps and solved correctly.")
- "vertex-base quadrilaterals" for lines, angles, and shapes, claim 4, DOK 3 in grade band 4

Places where equal percentages were observed for the TE and CR formats included:

- "select and order" for fractions, claim 1, DOK 2 in grade band 3
- "select and order" for fractions, claim 1, DOK 2 in grade band 6
- "selecting points" for fractions, claim 1, DOK 2 in grade band 3
- "single lines" for everyday math problems, claim 2, DOK 2 in grade band 11

Item types where the CR items had a higher percentage of consistent thought processes included:

- "select and order" for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6
- "tiling" for everyday mathematic problems, claim 4, DOK 3 in grade band 4
- "tiling" for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6
- "tiling" for everyday mathematic problems, claim 2, DOK 3 in grade band 11
- "vertex-base quadrilaterals" for lines, angles, and shapes, claim 1, DOK 2 in grade band 4

The TE item types where a higher percentage of students received full credit included only:

- "tiling" for equations and inequalities, claim 1, DOK 2 in grade band 11
- "vertex-base quadrilaterals" for lines, angles, and shapes, claim 1, DOK 2 in grade band 4

For the SR items in mathematics, the percentage receiving the maximum score was higher than both the CR and TE formats for the following items:

- "placing points" for fractions, claim 1, DOK 2 in grade band 3
- "select and order" for fractions, claim 1, DOK 2 in grade band 3
- "selecting points" for fractions, claim 1, DOK 2 in grade band 3

- “tiling” for fractions, claim 1, DOK 2 in grade band 3
- “tiling” for everyday math problems, claim 4, DOK 3 in grade band 4
- “vertex-base quadrilaterals” for fraction equivalence and ordering, claim 1, DOK 2 in grade band 3

In other cases the percentage receiving full credit was lower than for the comparable CR items. It should be noted that the percentage receiving full credit was generally low in mathematics.

Research Question 3: For multi-part selected response (MPSR) items where students may select more than one answer choice, which wording best indicates to the student that he or she is allowed to select more than one option? For multipart (e.g., YES/NO) dichotomous choice items, do students know that they need to answer each part?

Smarter Balanced sought to investigate whether students might become confused by MPSR items in mathematics and perhaps not complete the entire item. In order to investigate this, items were constructed with different amounts of labeling. Labeling is the identification of the parts of the problem with indicators such as “a,” “b,” “c” or “1,” “2,” “3.” A “labeled” and a non-labeled” condition were investigated. An example of items in the labeled and unlabeled format is presented below (Exhibit 1).

This question is designed to assess whether labeling or not labeling an MPSR mathematics item produces a difference in performance. Results are reported in five grade bands. The five grade bands are designated as grade band 3 (which includes form difficulty levels 3 and 4), grade band 4 (which includes form difficulty levels 4 and 5), grade band 6 (which includes form difficulty levels 6 and 7), grade band 7 (which includes form difficulty levels 7 and 8), and grade band 11 (which includes form difficulty level 11). Each form contains one MPSR item followed by one CR item. The labeled and non-labeled items appeared in different forms of the test and thus were taken by different students.

Exhibit 1. Example of a Labeled Item

Marcus has 36 marbles. He is putting an equal number of marbles into 4 bags.

Indicate whether each equation could be used to find the number of marbles Marcus puts in each bag.

1. $36 \times 4 = \square$ ○ Yes ○ No

2. $36 \div 4 = \square$ ○ Yes ○ No

3. $4 \times \square = 36$ ○ Yes ○ No

4. $4 \div \square = 36$ ○ Yes ○ No

Example of an Unlabeled Item

Marcus has 36 marbles. He is putting an equal number of marbles into 4 bags.

Indicate whether each equation could be used to find the number of marbles Marcus puts in each bag.

$36 \times 4 = \square$ ○ Yes ○ No

$36 \div 4 = \square$ ○ Yes ○ No

$4 \times \square = 36$ ○ Yes ○ No

$4 \div \square = 36$ ○ Yes ○ No

Results

Ninety-six students were administered the grade band 3 forms, 66 students were administered the grade band 4 forms, 133 students were administered the grade band 6 forms, 33 students were administered the grade band 7 forms, and 85 students were administered the grade band 11 forms.

Table 9 shows the percentage of students receiving full credit on the items by grade band and labeling condition. For grade bands 3, 4, 6, and 11 little difference between the labeled and non-labeled conditions is observed. However, in grade band 7 a higher percentage of students received full credit in the non-labeled format.

Table 9. Percentage of Students Receiving Full Credit, by Grade Band and Labeling Condition.

	Grade Band				
Condition	**3**	**4**	**6**	**7**	**11**
Non-Labeled	32	32	20	62	16
Labeled	29	31	18	34	9

Table 10 shows whether the students understood the instructions under the different item labeling conditions. Up through grade band 6 the type of instructions received seemed to have little impact on the understanding of the instructions. However, in grade bands 7 and 11 a higher percentage of students tended not to understand the instructions when the items were labeled. The interviewers commented that “Student did not have a complete understanding of instructions” and “He said he understood, however, he only selected one bubble.”

Table 10. Percentage Understanding the Instructions, by Grade Band and Labeling Condition

	Grade Band				
Condition	**3**	**4**	**6**	**7**	**11**
Non-Labeled	63	83	93	97	84
Labeled	78	93	93	69	61

Table 11 shows the percentage of students who made comments about not understanding the instructions. Grade bands 3 and 11 had more comments about not understanding the instructions than the other grade bands, but the pattern was similar for labeled and non-labeled items. However, in grade band 7, non-labeled items generally received no comment, with labeled items receiving more comments. This is consistent with a lower percentage of grade band 7 students understanding the instructions in the “labeled” condition.

Table 11. Did the Student Make Comments About not Understanding the Instructions (Percentage Making Comments)?

	Grade Band				
Condition	**3**	**4**	**6**	**7**	**11**
Non-Labeled	34	17	15	3	33
Labeled	32	26	8	38	41

Summary

Even though the labeling of MPSR items was intended to clarify the mathematic tasks for the students, in many cases it actually seemed to confuse the students. Little difference was observed between the labeled and non-labeled items in the lower grade bands (grade bands 3–6). However, students in grade band 7 tended to score higher with non-labeled items. Also, grade band 7 and 11 students tended to be confused by the labeling. In addition, the labeled items tended to receive more comments related to not understanding the instructions. The interviewer confirmed this, suggesting that the grade band 7 and 11 students better understood the instructions in the non-labeled condition than in the labeled condition.

Research Question 4: Does the ability to move one or more sentences to different positions provide evidence of students' ability to revise text appropriately in the consideration of chronology, coherence, transitions, or the author's craft?

Smarter Balanced is considering using items that have students reorder sentences to measure an editing/revising standard. Claim 2 of the standards states that students should be able to revise one or more paragraphs demonstrating specific narrative strategies (use of dialogue, sensory or concrete details, description), chronology, appropriate transitional strategies for coherence, or authors' craft appropriate to purpose (closure, detailing characters, plot, setting, or an event).

This question was designed to assess whether students' movement of one or more sentences to different positions provides evidence of students' ability to demonstrate consideration of chronology, coherence, transitions, or author's craft. Six ELA items were included in a test form. The forms were administered to five students: two in grade 5, two in grade 6, and one in grade 10. Because there is little difference in the pattern of responses and because the sample sizes are small, the results will be reported for the sample as a whole.

Results

It was hypothesized that students who do well on these items would recognize the need to revise for chronology, coherence, transitions, or author's craft. Table 12 shows the percentage of students who recognize the need to revise for chronology, coherence, transitions, or author's craft for students who performed well on the items and those who performed poorly. The results show that students

who performed well are more likely to consider chronology, coherence, transitions, or author's craft in their revisions than students who do not. Among the four writing skills examined, author's craft was considered less often than the other three writing skills.

Table 12. Percentage of Students Considering Targeted Writing Skills When Revising, by Those Students Who Performed Well and Those Who Performed Poorly

Characteristic	Students Who Perform Well	Students Who Perform Poorly
Chronology	100%	33%
Coherence	100%	33%
Transitions	100%	33%
Author's Craft	50%	0

Also of interest was whether students referenced organization, coherence, transitions, or author's craft when moving sentences. Table 13 shows the percentage of students who considered each of the targeted writing skills relative to the number of appropriate and inappropriate sentence moves. The results suggest that students who make more appropriate sentence moves (and fewer inappropriate sentence moves) are more likely to consider the writing skills of chronology, coherence, and transitions; however, the pattern is less clear for consideration of author's craft.

Table 13. Percentage of Students Who Considered Chronology, Coherence, Transitions, and Author's Craft at Each Number of Appropriate and Inappropriate Sentence Moves

% Students Who Recognized Need For	*N* Appropriate Sentences Moved								*N* Inappropriate Sentences Moved							
	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
Chronology	38	50	50	67	75		100	100	100	100	40	50	33	0		0
Coherence	38	43	67	67	75		100	100	100	100	40	50	33	0		0
Transitions	38	50	50	67	75		100	100	100	100	40	50	33	0		0
Anthor's craft	13	13	0	33	25		67	0	29	17	20	50	33	0		0

Table 14 shows the percentage of students who considered chronology, coherence, transitions, and author's craft when answering the items as observed by the interviewers. Students did express consideration of chronology ("I moved the first sentence because it goes at the top," "This seems to be in order," "This should be the second to last sentence"); coherence ("This seems like something you'd say," "I don't need to take out more phrases, it sounds OK," "I removed the two sentences because they did not make sense and were irrelevant"); and transitions ("This would sound better here") when answering the questions; however, fewer took author's craft ("I think there is a flow to the story," "Some sentences are awkward and need to be moved") into account when answering these questions.

Table 14. Percentage of Students Who Considered Chronology, Coherence, Transitions, and Author's Craft When Answering, Across Items

Writing Skills	Chronology	Coherence	Transitions	Author's Craft
Percentage	68	68	57	18

Summary

Students who performed well on the items were more likely to consider the targeted writing skills (chronology, coherence, transitions, and author's craft) when answering the questions. Also, students who made appropriate sentence moves were more likely to consider the targeted writing skills than those who made inappropriate sentence moves. A high percentage of students considered chronology, coherence, and transitions; however, they were less likely to consider author's craft.

Research Question 5: Do Students Who Construct Text Reveal More Understanding of Targeted Writing Skills Than Students Who Manipulate Writing Through the Manipulation of Text (MT) Tasks?

Many believe that the best way to measure writing is to have students write. However, in a testing environment, it is often difficult to adequately sample the writing content domain with an assessment composed exclusively of CR items. There is an ongoing effort to find items that are efficient, but that can adequately measure the components of the writing domain, thus allowing for a broader selection and greater number of items to be delivered. Examples of the types of questions used can be seen in Exhibit 2. The question examines whether comparable understanding of the targeted writing skills (see Table 15) can be achieved using a set of MT tasks in comparison to comparable CR tasks.

Four pairs of ELA items were developed. Each pair contained one MT item and one CR version of the same item. Two forms were created, with each form containing a single version of an item. Each form contained two MT items and two CR items. The MT items were almost exclusively "select and order" items, though two of the items, one in grade band 3 and one in grade band 11, were "reorder text" items.

Forms were constructed in ELA at three grade bands: grades 3–5 (referred to as grade band 3), grades 6 and 7 (referred to as grade band 6), and grades 10 and 11 (referred to as grade band 11). In grade band 3 two forms were administered; in grade bands 6 and 11, only a single form was administered. All forms assessed claim 1, target 1.

The sample consisted of seven students in grade band 3, two students in grade band 6, and one student in grade band 11.

Exhibit 2. Sample Items Used in this Research Question

Stem

A student wrote the first draft of a story about a girl who eats nine berries for an afternoon snack every day. Read the story. Then complete the task that follows.

Every day after school, Kim eats nine red, juicy raspberries. One day, Kim sits down at the big kitchen table and has a surprise. She notices that one of her berries is missing! "[ ]," she says.
"I counted nine just a minute ago," Dad says.
"[ ]," Kim says. "[ ]."
Kim begins her search in the garage. "[ ]?" Kim asks.

Dialogue
Oh no! There are only eight raspberries in my bowl

I wonder what happened to the ninth berry

Grandma, why are your mouth and lips red

It looks like I have a mystery to solve

Revise the story to include dialogue that introduces the plot. Place each piece of dialogue in the correct place in the story.

The dialogue will go in the brackets.

CR Prompt

A student wrote the first draft of a story about a girl who eats nine berries for an afternoon snack every day. Read the story. Then complete the task that follows.

> Every day after school, Kim eats nine red, juicy raspberries. One day, Kim sits down at the big kitchen table and has a surprise. She notices that one of her berries is missing!
>
> Her dad had counted nine just a few minutes ago.
>
> Kim knew she had a mystery to solve.
>
> Kim began her search in the garage. She found her grandmother in the garage with bright red lips.

Revise the story to include dialogue. Use dialogue to introduce the plot.
Type your response in the space provided.

Table 15. Targeted Writing Skills with Examples of Representative Statements

Target	Types of Statements
Chronology	- I knew it was telling a story, so looked for the beginning then moved the rest around to make sense. - I knew what the end was, so worked backwards from there. - I knew the youngest son went last, so put him at the end, then put the two older ones before him. Then picked the beginning and put it first. - Some spots didn't sound quite right, so added the sentences in. - Read the sentences, then looked for related sentences in the passage that they'd go with. - I used transitions to cue position of sentences. - I need to revise the order of the sentences so that they more clearly support the main idea of the article. I do not need to move the first or last sentence.
Coherence	- Sentence is like a preview of the rest of the essay, so it should go first. - This sentence sounds professional and it also connects to the facts that follow. This is the best thesis statement. - This sentence wraps up the author's argument/point of view and finishes the essay by restating the main point. - The conclusion often just rephrases the thesis, which this sentence does, but it also talks about other things from the passage, so it should be the conclusion. - I have to choose the two sentences that shouldn't be part of the paragraph. - I have to take the sentence at the top and drag it to best spot in the paragraph below.
Transitions	- The word "next" tells him it comes after something else. - The word "first" is a clue that it goes at the beginning. - "Finally" usually tells you you're at the end. - A transition like "therefore" at the start of a sentence connects it to the sentence before. They have the same topic but this one comes second. - I have to use transitions words to make the paragraph clearer. - I looked at the transition words to see what should come before them, then put in a sentence if needed.
Author's Craft	- I found the parts that didn't give me a really

	clear picture in her mind and changed them. - I looked for the parts that weren't as descriptive as the rest and made them more descriptive. - I looked for the parts that sounded a little boring and made them more exciting. - I read the topic sentences and looked for the sentence that didn't go with it. - If a sentence makes the argument weaker, then it should be taken out, so these two need to be removed.

Results

It was hypothesized that student think-alouds on MT items would reference the appropriate writing skills reflected in the assessment target at a level comparable with CR items. Table 16 shows the percentage of students who referenced the targeted writing skills, by item format and grade band. In grade band 3, chronology was more likely to be considered during revision when the item format was MT ("First, next, last order of events") than when the item format was CR ("Historically probably comes first, having trouble ending story"). Similar patterns, but less pronounced, were seen with coherence, transitions ("This is a cause…as a result (an effect) should be here"), and author's craft. Grade band 3 students only considered author's craft during revision for about one-third of the items regardless of item format. Grade band 6 students always considered chronology and coherence during revision, but transitions and author's craft were only considered about half the time. In grade band 11 chronology, coherence, and transitions were always considered in both formats. Author's craft was only considered about half the time in the CR format and not mentioned at all in the MT format. One interviewer commented, "Student made no comment about author's craft."

Table 16. Percentage of Items in Which Students Considered Target Characteristics When Responding to the Item, by Item Format

		Grade Band		
Target Characteristics	Item Format	3	6	11
Chronology	CR	31	100	100
	MT	94	100	100
Coherence	CR	63	100	100
	MT	75	100	100
Transitions	CR	44	50	100
	MT	69	50	100
Author's Craft	CR	31	50	50
	MT	43	100	0

Table 17 shows the counts for item scores received for the two item formats, by grade band. Comparable scores were achieved for the two item formats.

Table 17. What Score (Across Items) Would the Student Receive on this Type of Item?

		Grade Band								
		3				6		11		
	Item Format	0	1	2	0	1	2	0	1	2
	CR	8	5	3	0	0	2	0	0	2
	MT	7	6	2	1	0	1	0	2	0

Table 18 provides information about whether the students who construct text through writing reveal comparable or greater understanding of targeted writing skills than students who manipulate text. The grade band 3 and grade band 6 students were either more effective in applying the targeted writing skills when the items were in a MT format or no differences were observed in effectiveness between item formats. For the grade band 11 students the results were mixed, but students tended to be more effective in applying the targeted writing skills in the CR format, particularly for transitions and author's craft.

Table 18. Effectiveness of Applying Targeted Writing Skills by Item Format (Percentage of Students as Assessed by Interviewer)

	Grade Band								
	3			6			11		
Target Characteristics	MT More Effective	No Difference	CR More Effective	MT More Effective	No Difference	CR More Effective	MT More Effective	No Difference	CR More Effective
Chronology	38	63	0	0	100	0	0	100	0
Coherence	38	63	0	0	100	0	100	0	0
Transitions	25	75	0	50	50	0	0	0	100
Author's Craft	38	63	0	50	50	0	0	0	100

Summary

The results showed that the targeted writing skills are considered by students who manipulate text at a level comparable to (or greater than) that encountered when they are constructing text. The grade band 3 and 6 students showed comparable (or greater) levels of understanding when the items were in an MT format. For the grade band 11 students the results were mixed, but students tended to be more effective in applying the targeted writing skills in the CR format, particularly for transitions and author's craft. Score distributions were comparable for MT and CR item formats.

Research Question 6: Do different types of directions (minimal, concise or extensive) have an effect on the performance of different item types in ELA and Mathematics?

The optimal amount of direction that should be given to students for some item types is unclear. With minimal directions students may not know how to approach the item; with extensive directions students may be distracted or slowed to a point where the item becomes inefficient. This may be particularly true with elementary school students, who may take longer to process text. This question examined these issues for ELA and mathematics items. Three types of directions (minimal, concise, and extensive) were examined for different item types.

Forms were constructed in ELA at five grade bands: grade 3 (referred to as grade band 3), grades 4 and 5 (referred to as grade band 4), grades 6 and 7 (referred to as grade band 6), grades 7 and 8 (referred to as grade band 7), and grade 11 (referred to as grade band 11) with a single form administered in each grade band. In mathematics, forms were constructed at four grade bands: grades 3 and 4 (referred to as grade band 3), grades 4 and 5 (referred to as grade band 4), grades 6 and 7 (referred to as grade band 6), and grade 11 (referred to as grade band 11).

Parallel items were created with minimal, concise, or extensive directions in ELA and for most item types in mathematics. However, not all direction types appeared with all item types in all grades in mathematics. Four items in one of the three formats (SR, TE, and CR) appeared in each mathematics form. Two items in one of the three formats appeared in each ELA form. Multiple forms were administered, each one to a different sample of students. An example of the different direction types for an ELA item and a mathematics item is presented in Exhibit 3.

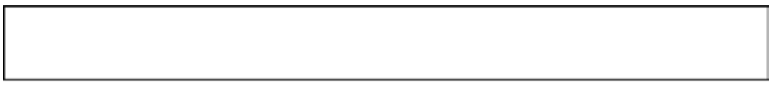
Exhibit 3. Example of the Types of Instructions Under the Minimal, Concise, and Extensive Instruction Condition for the Item That Follows

ELA Example

Minimal Directions
Drag the **best** transition word to each blank in the paragraph.

Concise Directions
Complete the paragraph by selecting the **best** transition word that fits in each blank. Drag each transition word you selected to the correct blank in the paragraph.

Extensive Directions
There are six transition words in the text box. Complete the paragraph correctly by choosing a transition word that **best** fits each blank. Drag the transition word you selected from the text box to the correct blank in the paragraph.



It was winter. The cold wind was blowing and snow was covering the ground. Sarah gazed out the window and saw a bird trying to find food. She wanted to help the bird. After thinking for a while, Sarah decided to make a pinecone bird feeder. First, she tied a string to the top of a pinecone. _____, she covered the pinecone with peanut butter. After this, she placed the pinecone in the freezer. Later, she rolled the pinecone in birdseed. _____, she placed the pinecone bird feeder on a tree for the birds.

Mathematics Example

Minimal Directions

Drag numbers to make the equations true.

Concise Directions

Move numbers to make the equations true.

Drag the numbers to the answer space.

Extensive Directions

Drag numbers to make the equations true.

Each number can be used only once. To use a number, drag it to the appropriate box in an equation.

$\sqrt{\square} = \square$

$\sqrt[3]{\square} = \square$

Results

Table 19 provides a count of the students in a grade band, by content area and direction type.

Table 19. Sample Sizes by Content Area, Direction Type, and Grade Band

Content	Direction Type	Grade 3	Grade 4	Grade 6	Grade 7	Grade 11
ELA	Minimal	14	12	14	14	10
ELA	Concise	12	15	12	12	14
ELA	Extensive	18	17	15	7	6
Mathematics	Minimal	4	11	8	-	18
Mathematics	Concise	20	4	27	-	27
Mathematics	Extensive	19	4	27	-	16

Table 20a shows the percentage of students receiving full credit for the ELA items by direction type, item type, and grade band. In grade band 3, "select text" items were more challenging than "reorder text" items. This was especially true when the directions were "concise." With the "reorder text" items the grade band 3 students did less well with minimal directions. The grade band 11 students also had some difficulty with the "reorder text" items when the directions were "extensive." For the other grade bands, neither the level of instruction nor the item type showed a differential effect.

Table 20a. Percentage of Students Who Received Full Credit on ELA Items by Direction Type and Grade Band

ELA		Grade Band				
Direction Type	Item Type	3	4	6	7	11
Minimal	Reorder Text	40				71
Concise	Reorder Text	100				59
Extensive	Reorder Text	67				33
Minimal	Select and Order		69			
Concise	Select and Order		75			
Extensive	Select and Order		53			
Minimal	Select Text	33		100	41	
Concise	Select Text	0		100	60	
Extensive	Select Text	38		100	50	

In mathematics, a low percentage of students received full credit for "placing points" under the minimal and concise directions in grade band 11 (Table 20b). However, under extensive directions all students received full credit. With "placing points and tiling" items a higher percentage of students received full credit as the amount of instructions were reduced (grade band 6). "Select and order" items were difficult (grade bands 6 and 11) regardless of the direction type; however, no direction type proved better than another. The "select defined partition" items and the "straight lines" items showed high percentages of students receiving the maximum score, but the direction type did not make a difference. "Vertex-based quadrilateral" items seemed to benefit from minimal directions in grade band 11. Finally, "tiling" items were generally difficult, but no benefit was shown for different types of directions. The incompleteness of the data limits other comparisons.

Table 20b. Percentage of Students Who Received Full Credit on Different Types of Mathematics Items, by Direction Type and Grade Band

Direction	Template	Grade Band			
		3	4	6	11
Minimal	Placing Points				21
Concise	Placing Points				21
Extensive	Placing Points				100
Minimal	Placing Points and Tiling			67	
Concise	Placing Points and Tiling			57	
Extensive	Placing Points and Tiling			38	
Minimal	Select and Order				44
Concise	Select and Order			32	43
Extensive	Select and Order			33	0
Minimal	Select Defined Partitions	100	70		
Concise	Select Defined Partitions	76	100		
Extensive	Select Defined Partitions	71	83		
Extensive	Single Ray			15	
Minimal	Straight Lines		100		100
Concise	Straight Lines	100	100		100
Extensive	Straight Lines	100			
Extensive	Straight Line and Tiling			29	
Concise	Tiling	19			
Extensive	Tiling	20	20		
Minimal	Vertex-Based Quadrilaterals				69
Concise	Vertex-Based Quadrilaterals			64	88
Extensive	Vertex-Based Quadrilaterals	30			

Understanding instructions

In ELA (Table 21a), for most item type/direction type/grade band combinations few students had difficulty understanding instructions. Cases in which difficulties were mentioned included about 50 percent of the students in grade band 4 with both minimal and extensive instructions for the "select and order" items. This was also true in grade band 3 for the "reorder text" items with extensive instructions and for the "select test" items with concise and extensive instructions. Finally, in grade band 11 the "reorder text" items with minimal and concise instructions elicited more comments.

In mathematics (Table 21b), the cases in which more comments were made about the instructions included "placing points" with minimal and concise instructions (grade band 11), "single ray" items with extensive instructions (grade band 6), "straight lines" items with extensive instructions, and "vertex-based quadrilateral" items with extensive instructions (grade band 3). The single ray item with extensive instructions in grade band 6 stood out as an item in which instructions were not well understood. ("Weren't totally sure how instructions were to be completed.") The percentage of students getting the maximum score on this item type was also low.

Table 21a . Percentage of Students Who Express the Difficulties in Understanding Each Type of Instruction for Each TE Type in Their Think-Alouds (ELA)

ELA		Grade Band									
		3		4		6		7		11	
Direction Type	Item Type	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit
Minimal	Reorder Text	0	0							0	20
Concise	Reorder Text		25							0	20
Extensive	Reorder Text	33	12							0	0
Minimal	Select and Order			50	11						
Concise	Select and Order			0	6						
Extensive	Select and Order			44	0						
Minimal	Select Text	0	0				0	6	0		
Concise	Select Text	33					14	0	22		
Extensive	Select Text	25	40				19	10	10		

Table 21b. Percentage of Students Who Express the Difficulties in Understanding Each Type of Instructions for Each TE Type in Their Think-Alouds (Mathematics)

Math		Grade Band							
		3		4		6		11	
Direction Type	TE type	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit
Minimal	Placing Points							55	33
Concise	Placing Points							67	0
Extensive	Placing Points								20
Minimal	Placing Points and Tiling					0	0		
Concise	Placing Points and Tiling					33	25		
Extensive	Placing Points and Tiling					7	33		
Minimal	Select and Order							14	9
Concise	Select and Order					13	8	4	10
Extensive	Select and Order					12	8	12	
Minimal	Select Defined Partitions		0	0	0				
Concise	Select Defined Partitions	14	9		0				
Extensive	Select Defined Partitions	25	13	0	20				
Extensive	Single Ray					82	33		
Minimal	Straight Lines				25				
Concise	Straight Lines				0			100	0
Extensive	Straight Lines		50						
Extensive	Straight Lines and Tiling					0	0		
Concise	Tiling	15	0						
Extensive	Tiling	6	0	0	0				
Minimal	Vertex-Based Quadrilaterals							25	0
Concise	Vertex-Based Quadrilaterals	30	0			67	12	33	14
Extensive	Vertex-Based Quadrilaterals	43	17						

Difficulty Using the Computer

The results for ELA related to difficulty using the computer were mixed (Table 22). In grade band 3 under minimal directions for both "select text" and "reorder text" items, the students seemed to have difficulty using the computer. The grade band 11 students seemed to have some difficulty with the "reorder text" items.

Table 22. Percentage of Students Who Said They Had Trouble Using the Computer (ELA)

		Grade				
Direction Type	Item Characteristic	3	4	6	7	11
Minimal	Select Text	43		4	0	
Concise	Select Text	25		0	4	
Extensive	Select Text	19		8	0	
Minimal	Select and Order		22			
Concise	Select and Order		25			
Extensive	Select and Order		16			
Minimal	Reorder Text	31				25
Concise	Reorder Text	11				48
Extensive	Reorder Text	24				30

Most students in mathematics had little trouble using the computer with mathematics items.

Summary

In most cases in ELA the level of instruction did not have an influence. For most grade bands and item types, neither the level of instruction nor the item type had a differential effect in ELA. Cases in which differences were observed included "select text" items when the directions were "concise" (grade band 3). With the reorder text items the grade band 3 students did less well with minimal directions. The grade band 11 students also have some difficulty with the "reorder text" items when the directions were "extensive."

In mathematics, the level of instruction also did not make a difference for many of the item types and grade bands. "Select and order" items were difficult (grade bands 6 and 11) regardless of the direction type; however, no direction type proved better than another. High percentages of students received full credit on the "select defined partition" items and the "straight lines" items; however, the direction type did not make a difference. Finally, "tiling" items were generally difficult, but no benefit was shown for different types of directions. Places where differences were observed included "placing points" under the minimal and concise directions in grade band 11; however, under extensive directions all students received the maximum score. In working with "placing points and tiling" items, a higher percentage of students received full credit with fewer instructions (grade band 6). Finally, "vertex-based quadrilateral" items seemed to benefit from minimal directions in grade band 11.

The results for ELA related to trouble using the computer were mixed. In grade band 3 under minimal directions with both select text and reorder text items the students seemed to have difficulty using the computer. The grade band 11 students seemed to have some difficulty with the "reorder text" items. Mathematics students did not seem to have any problems using the computer.

ELA Questions, Passage Processing

Research Question 7: Smarter currently intends to administer the passage first, and then administer the items one item at a time. Does this affect student performance?

Smarter Balanced is interested in the possibility of administering items adaptively within a passage. This would require administering items sequentially so that the ability estimate could be updated after each item. Presenting items one at a time may take longer, and students may object to not knowing what is coming next. This question is designed to assess whether administering an item set takes longer when the items are presented sequentially and whether there is a difference in confusion or frustration level when students are presented a passage and all the items together or are presented a passage with the items then being presented one at a time. The item sets were not administered adaptively.

Two sets of items were created for a given test form. Both sets contained passages of equivalent length and difficulty as well as items of equivalent difficulty.[2] The first set in a form presented the passage with all the items together. The second set presented the passage with the items presented one at a time.

The forms were administered, within grade band, to different samples of students. Each sample contained both a general education group (Gen Ed) and a group that received ELL students. One sample was timed without thinking aloud during the administration. Each item set in these forms was separately timed. This sample provided timing information only. The second sample involved thinking aloud while responding to the questions and was not timed. Forms were constructed in ELA at three grade bands: grades 3–5 (referred to as grade band 3), grades 6–8 (referred to as grade band 6), and grades 10 and 11 (referred to as grade band 11).

The primary questions of interest were:

1. Does presenting the items individually after the passage appear to take longer (timed condition)?
2. Does presenting the items individually after the passage increase the student's negative emotional states (e.g., frustration, confusion; think-aloud condition)?
3. Do students prefer one approach or another (think-aloud condition)?

[2] Comparable passage difficulty was achieved through the use of readability and lexile measures. Comparable item difficulty was achieved through depth of knowledge (DOK) measures.

Results

Table 23 shows the sample sizes taking each form of the tests, by grade band, for the ELL and Gen Ed samples. Sample sizes are smaller for the ELL sample in grade band 11.

Table 23. Student Counts by Grade Band, Testing Population, and Testing Condition

		Grade Band		
		3	6	11
Timed	Gen Ed	9	6	8
	ELL	8	4	1
Think-Aloud	Gen Ed	6	6	7
	ELL	8	7	2

Table 24 shows the time (in seconds) it took to complete the item sets when all items were presented together or items were presented one at a time, by grade band and sample. For the grade band 3 and grade band 11 samples, timing differed little whether the items were presented in one block or one at a time. However, for grade band 6, presenting the items one at a time took substantially longer. While there is some variability between the ELL and the Gen Ed samples, the differences are not large and show a similar pattern. Note that the grade band 11 ELL sample was a single student and is not presented to avoid misleading results.

Table 24. Average Time to Complete the Passage and Items, by Administration Format, Grade Band, and Sample

Grade Band	Sample	*N*	Passage + All Items	Passage + One Item at a Time	Difference (All—One at a Time)
3	Gen Ed	9	250	239	11
3	ELL	8	263	239	24
6	Gen Ed	6	401	462	–61
6	ELL	5	336	465	–129
11	Gen Ed	8	270	285	–15

Tables 25 and 26 show whether the ELL or Gen Ed sample students expressed confusion (Table 25) or frustration (Table 26) with the passages or items. There appears to be slightly more confusion for both the Gen Ed and the ELL sample students in grade band 3 when all the items are presented together. However, similar frustration levels were observed under the two formats for the grade band 3 students. The grade band 6 ELL sample, showed similar patterns of frustration and confusion for the two presentation formats. However, the Gen Ed grade band 6 students showed slightly more confusion when the items were presented one at a time. The grade band 11 Gen Ed students showed similar levels of confusion and frustration under both administrative formats. The grade band 11 ELL sample included only two students and is not reported.

Table 25. Percentage of Students Expressing Confusion with the Different Components of the Test by Administration Format, Grade Band, and Sample

		All Items			One at a Time		
		Grade Band			Grade Band		
Sample	Test Component	3	6	11	3	6	11
Gen Ed	Passage	33	29	17	0	43	14
	Items	25	30	17	9	36	18
ELL	Passage	50	50		25	50	
	Items	32	50		16	50	

Table 26. Percentage of the Students Expressing Frustration with the Different Components of the Test, by Administration Format, Grade Band, and Sample

		All Items			One at a Time		
		Grade Band			Grade Band		
Sample	Test Component	3	6	11	3	6	11
Gen Ed	Passage	0	29	17	0	29	14
	Items	13	18	17	13	11	14
ELL	Passage	13	38		13	38	
	Items	3	41		3	50	

Table 27 presents the average score students obtained for the think-aloud protocols. The grade band 6 students tended to score higher when the items were presented all at one time (for both the Gen Ed students and the ELL students). The grade band 3 students scored higher when the items were presented one at a time, regardless of sample or testing condition. The grade band 11, Gen Ed students scored higher when the items were presented one at a time, while the grade band 11, ELL sample students scored higher when the items were presented all at one time, though the latter sample size is small.

Table 27. Average Score, by Administration Format, Grade Band, and Sample

	All Items at Once			One Item at a Time		
	Grade Band			Grade Band		
Gen Ed	2.2	3.0	1.8	2.5	2.3	2.5
ELL	2.4	2.9	2.0	2.5	1.7	1.5

Table 28 shows the preference for a presentation format. Both the ELL and Gen Ed grade band 3 students preferred to have the items presented one at a time. ("I preferred one at a time—less confusing than seeing too many questions," "One at a time made me less nervous about how many more there were," "I liked one at a time because it did not seem overwhelming.") Grade band 11 students (Gen Ed and ELL) had a slight bias toward having the items presented one at a time ("Let's me focus on that one question"). Conversely, grade band 6 Gen Ed students preferred to have the items presented together ("I liked them altogether," "This way I know I was on the same passage," "All together, you can refer to the questions while you read the passage," "I liked everything on one page because it was more easy," "With all together, I was able to refer back and I could see where I was going," "I liked altogether, though it was more confusing and distracting.") The grade band 6 ELL students were equally divided between the two formats.

Table 28. We Presented the Questions to You in Two Different Ways. Which Way Did You Prefer: All Together or One at a Time (Percent Responding)?

	Grade Band								
	3			6			11		
Sample	All Together	No Preference	One at a Time	All together	No Preference	One at a Time	All Together	No Preference	One at a Time
Gen Ed	33		67	57	29	14	29	14	57
ELL	14		86	43	14	43		50	50

Summary

We were interested in assessing whether there is a difference in timing and increased negative emotional states (confusion, frustration) when students are presented a passage with all the items or are presented a passage with the items presented one at a time. Forms were administered to two groups of students: a group that received English language accommodations and a Gen Ed group.

The time it took to complete the sets when all items were presented together or one at a time varied by grade band and sample. For the grade band 3 and grade band 11 samples, timing differed little whether the items are presented in one block or one at a time. However, for grade band 6, presenting the items one at a time took substantially longer for both the Gen Ed and ELL samples. While there is some variability between the ELL and the Gen Ed samples, the differences are not large and show the same pattern within grade band.

There appeared to be slightly more *confusion* for both the Gen Ed and the ELL samples in grade band 3 when all the items were presented together. However, similar *frustration* levels were observed under the two formats for the grade band 3 students. The grade band 6 ELL sample students showed similar patterns of frustration and confusion for the two presentation formats. However, the Gen Ed grade band 6 students showed slightly more confusion when the items were presented one at a time.

The grade band 6 students tended to score higher when the items were presented all at one time (for both the Gen Ed students and the ELL students). The grade band 3 students showed similar results, regardless of sample or administration format. The grade band 11, Gen Ed students scored higher when the items were presented one at a time, while the grade band 11 ELL sample students scored higher when the items were presented altogether.

Both the ELL and Gen Ed grade band 3 students preferred to have the items presented one at a time. Grade band 11 students had a slight bias toward having the items presented one at a time. Conversely, grade band 6 students preferred to have the items presented together.

Research Question 8: Smarter intends to present relatively long passages. Do longer passages reduce student engagement?

Smarter Balanced is interested in using passages that are longer than those presently used. The Smarter Balanced recommended passage lengths are: for grades 3–5: 450–562 words for short passages and 563–750 words for long passages; for grades 6–8: 650–712 words for short passages and 713–950 words for long passages; and for high school, 800–825 words for short passages and 826–1100 words for long passages. There is concern that the longer passages may tax the processing abilities of ELL and SWD students.

This question is designed to assess whether longer passages reduce student engagement, hamper the completion of the longer passages, or affect the depth of processing of the passage. Two sets of items were created. Both sets contained passages of equivalent difficulty with four items of equivalent difficulty attached to each passage. Both sets present the passage and all the items together. Each form contained a standard-length and an extended-length passage. The first set contained a passage of standard length. The second set contained a passage that is longer than

standard length (extended-length, the length equivalent to that intended for use by Smarter Balanced).

Forms were constructed in ELA at three grade bands: grade band 3–5 (referred to as grade band 3), grade band 6–8 (referred to as grade band 6), and grade band 10 and 11 (referred to as grade band 11). The design was intended to compare the performance of two groups of students—ELL/SWD and Gen Ed students—across three grade bands (3, 6, and 11). Thirteen students took the forms. Of these, nine were grade band 3 Gen Ed students. One grade band 3 student was classified ELL/SWD. The single grade band 6 student was an ELL/SWD student. The two grade band 11 students were Gen Ed students.

Results

Table 29 shows the percentage of students whose engagement was improved or unaffected by the longer passage, by subgroup. All the ELL/SWD students were unaffected by the use of the longer passage. Gen Ed students did appear to be affected by the longer passage in grade bands 3 and 11. All the ELL/SWD students were able to read the entire passage regardless of passage length. Only about 25 percent of the grade band 3 Gen Ed students and none of the grade band 11 Gen Ed students were unaffected by the use of the longer passage (see Table 29; "I have to read the whole passage?"). The ELL/SWD students all demonstrated that the longer passage was processed at a deep level ("It was a good story"). However, only 43 percent of the Grade band 3, Gen Ed, students demonstrated a level of deep processing ("I learned many new things") and only 50 percent of the grade band 11 Gen Ed students demonstrated a level of deep processing (Table 31). The ELL/SWD students were not bored or distracted while reading either passage; however, some percentage of the Gen Ed students were bored regardless of the length of the passage.

Table 29. Percentage of Students Whose Engagement Is Improved or not Affected by the Longer Passage

	Grade Band		
Subgroup	**3**	**6**	**11**
GE	25		0
ELL/SWD	100	100	

Table 30. Percentage of Students Who Appear to Read the Entire Passage

Standard Length	Grade Band		
Subgroup	3	6	11
GE	88		100
ELL/SWD	100	100	
Extended Length	Grade Band		
Subgroup	3	6	11
GE	88		50
ELL/SWD	100	100	

Table 31. Percentage of Students Whose Think-Aloud Demonstrate Deep Processing as Assessed by the Interviewer

Standard Length	Grade Band		
Subgroup	3	6	11
GE	43		100
ELL/SWD	100	100	
Extended Length	Grade Band		
Subgroup	3	6	11
GE	43		50
ELL/SWD	100	100	

Table 32. Percentage of Students Who do not Appear Bored or Distracted

Standard Length	Grade Band		
Subgroup	3	6	11
GE	63		100
ELL/SWD	100	100	
Extended Length	Grade Band		
Subgroup	3	6	11
GE	88		50
ELL/SWD	100	100	

Summary

Smarter Balanced is interested in using passages that are longer than those presently used. There is concern that the longer passages may tax the processing abilities of ELL and SWD) students. This question is designed to assess whether longer passages reduce student engagement, hamper the completion of the longer passages, or affect the depth of processing of the passage. The design was intended to compare the performance of two groups of students—ELL/SWD and Gen Ed students—across three grade bands (3, 6, and 11). Two sets of items were created. Both sets contained passages of equivalent difficulty with four items of equivalent difficulty attached to each passage. Both sets present the passage and all the items together. Both the standard-length and the extended-length passage were included in a given form and administered to the same student.

All the ELL/SWD students were unaffected by the use of the longer passage. They were able to read the entire passage regardless of passage length and demonstrated that the longer passage was processed at a deep level. The ELL/SWD students also were not bored or distracted while reading either passage.

On the contrary, Gen Ed students did appear to be affected by the longer passage in grade bands 3 and 11. About 75 percent of the grade band 3 students and all of the grade band 11 students were affected by the use of the longer passage. Only 43 percent of the Grade band 3 Gen Ed students demonstrated a level of deep processing and only 50 percent of the grade band 11 Gen Ed students demonstrated a level of deep processing. Also, some percentage of the Gen Ed students were bored, regardless of the length of the passage

Research Question 9: How long does it take for students to read through complex texts, performance tasks, etc.? Is timing affected by the way students are presented the passage and items?

One way of making items more difficult is to increase their complexity. Complex items often take longer to solve or answer. In computer adaptive tests, added complexity may decrease the time a high ability student has to complete the test if the items are made more difficult through increased complexity. This potentially creates some fairness issues in an adaptive test if there is a time limit on the test. This question was designed to assess the time it takes for students to answer complex and simpler items. Complexity was defined as a function of the DOK demanded by the test question. It was hypothesized that more complex tasks would take more time.

Each ELA form had six items. These items varied in item complexity (simple or complex) and item format (SR, TE, or CR). The TE items were all "hot text" items. These items require the student to either highlight the text or drag the text to answer the item.

Forms were constructed in ELA at two grade bands: grade band 3–5 (referred to as grade band 3) and grade band 6 and 7 (referred to as grade band 6). Two forms were administered in grade band 3. One form was administered in grade band 6.

Results

Eight students took the grade band 3 forms with four students taking each form, and two students took the grade band 6 form.

Table 33 presents the average time (in seconds) a student took to answer an item. SR items were answered in the shortest time. HT items took about one minute longer than the SR items. CR items took the most time to answer, about 75 seconds longer than the "hot text" items. With the exception of the complex CR item administered to grade band 6 students, item complexity did not seem to have an impact on item performance. (An interviewer commented, "Student took about the same time for complex and easy items.")

Table 33. Average Time (in seconds) to Answer an Item by Grade Band, Item Type, and Item Complexity

			Grade Band	
			3	6
Item Format	Difficulty	Item	Avg. Time	Avg. Time
SR	Simple	1	49	52
SR	Complex	2	29	59
TE (HT)	Simple	3	83	126
TE (HT)	Complex	4	96	123
CR	Simple	5	182	168
CR	Complex	6	158	185

Table 34 presents a summary of the average time students took to complete complex and simple items across item types by grade band. Complex items seemed to have more impact in grade band 6, but there is no evidence that complex items, as defined here, take longer than simpler items.

Table 34. Interviewer's Summary of Item Timing by Grade Band and Item Difficulty

	Grade Band	
	3	6
Difficulty	Avg. Time	Avg. Time
Simple	104	115
Complex	94	126

Summary

It was hypothesized that more complex items would take longer to complete than simpler items. No evidence was found to support this hypothesis. In terms of the time spent on an item, SR items were answered in the shortest time. "Hot text" items took about one minute longer than SR items. CR items took the most time to answer, about 75 seconds longer then the "hot text" items.

Effective Communication of Mathematics

Research Question 10: Working mathematics problems on computer: Communicating mathematics on computer—feasibility of measuring student understanding of items for Claims 2–4 on computer.

With paper tests some students write in their test books while working out mathematics problems. When mathematics items are presented on computer, scratch paper is often provided if students want to transfer the problem to paper and work it out there. Because scratch paper is often destroyed after an online testing session, the degree to which scratch paper is used is not known; neither is the importance of scratch paper in working out a problem (or potentially for use in scoring). This research question examines the need for paper when solving mathematics problems. Forms were constructed at four grade bands: grade band 3 and 4 (referred to as grade band 3), grade band 6 and 7 (referred to as grade band 6), grade band 7 and 8 (referred to as grade band 7), and grade band 11 (referred to as grade band 11) to investigate whether the scratch paper usage was uniform or varied by educational level.

Each student was presented with three grade-appropriate items. The interviewer recorded whether the student made a comment, and the nature of the comment, while working the mathematics problems. The students first tried to work the problem without paper. Scratch paper was then offered to the student to rework the problem, if desired. The interviewer noted whether students chose to add anything additional and noted the nature of the addition (more text, equations, graphics). Note that there were only three comments for the third item in the lowest grade band, 3.

Results

Twenty students were administered the grade band 3 form, 37 students were administered the grade band 6 form, 21 students were administered the grade band 7 form, and 19 students were administered the grade band 11 form.

Table 35 shows the percentage of comments made for an item and the type of comment made. Two types of comments were of interest: did the students who wanted paper draw a picture or write an equation or did they find the online system difficult to use. The lowest grade band students (grade band 3) did not need paper to solve any of the problems (Table 635. Some students in the highest grade band (grade band 11) commented that they would like to draw a picture for the items they were administered (15–30 percent). (“I wanted to graph the area.”) There was also one item (Item 2) for which about 15 percent of students wanted paper to write equations. About 5–10 percent of students in each grade band found the online system difficult to use. (“Confused me, I didn’t know how to write an equation,” “Tried the keypad, but it wouldn’t work,” “It was much easier with paper.”) The strongest result came from the grade band 6 and grade band 7 groups, where 30 to 42 percent of the sample, respectively, indicated that they wanted to write an equation. Between 3 and 23 percent of the grade band 6 and 7 groups also indicated that they wanted to draw a picture. This may be a function of newly introduced algebra concepts for this group.

Table 35. Percentage of Comments for an Item, by Question Type and Grade Band

		Grade Band			
Question	Item	3	6	7	11
Picture	1	5	0	23	32
	2	15	12	3	16
	3	0	4	6	16
System Difficulty	1	5	9	10	5
	2	11	3	10	5
	3	0	4	7	5
Equation	1	0	31	45	6
	2	0	32	34	16
	3	0	29	43	6

Table 36 shows the nature of the student comments made on paper and whether the additional information recorded on the paper improved the response according to the rubric. For all grade bands the additional information recorded on the paper included a graphic. In grade bands 6, 7, and 11, the additional information recorded on paper included an equation. The grade band 6, 7, and 11 groups provided additional information on paper that improved the response according to the rubric. For example, one administrator noted, “When given paper, she was able to do the proper equation and solve for x. She was more confident with paper and pencil.” The number of cases in which improvement was observed varied by item. For grade band 6, item 2, about 11 percent of the responses were improved when scratch paper information was taken into account during scoring. For grade band 11, item 3, about 16 percent of the responses were improved when scratch paper information was taken into account during scoring. Responses to all items in grade band 7 were improved when scratch paper information was taken into account. The improvement for this group ranged between 10 and 20 percent across items.

Table 36. Percentage of Changes Made When Paper Was Introduced

		Grade Band			
Nature of Students' Changes	Item	3	6	7	11
No Additions Made	1	80	57	71	53
	2	60	65	67	63
	3	10	32	52	58
Addition Included Graphic	1	5	3	33	37
	2	15	5		11
	3			10	32
Addition Included Equation	1		22	19	5
	2	20	16	38	16
	3		19	38	11
Addition Improved Response According to Rubric	1		11	14	
	2		3	29	
	3			24	16

The interviewer's comments suggested that most students in grade band 3 (75 percent) and grade band 11 (63 percent) were able to accurately respond to the mathematics items they saw only using the online text editor. However, fewer than half of the students in grade band 6 (45 percent) could accurately respond to questions using only the text editor and only 13 percent of the students in grade band 7 were observed to be able to accurately respond to questions using only the text editor. One student commented, "It's much easier with paper."

Summary

The general conclusion is that a subset of students benefit from being able to work mathematics problems on paper. It appears to be especially important when students are beginning to learn algebra concepts.

Grade band 3 students did not need paper to work the problems. However, in the grade band 6 and grade band 7 groups, 30–42 percent of students indicated that they wanted to write an equation. In grade bands 6, 7, and 11, the additional information recorded on paper would have improved the

response according to the rubric. Responses for specific items in grade bands 6 and 11 were improved by 15 percent of the students and responses for all items in grade band 7 were improved when information on the scratch paper was taken into account. Improvement for this group ranged between 10 and 20 percent of the responses. This was supported by interviewer observations. About 5–10 percent in each grade band found the online system difficult to use, but few specifics were recorded.

Research Question 11: Usability of equation editor tool—can students use the tool the way it is meant to be used?

Although students begin to use technology at a very early age, it is prudent to verify that young students are able to use the assessment interface to be used during testing. This question sought to evaluate the ability of grade 3–5 students to use the equation editor tool to be included in the Smarter Balanced delivery system. Three mathematics items were presented to the students (*N*=33). The first item only required the student to copy his or her response. The second item was a simple mathematics item and the third item was a more challenging mathematics item. The first item would demonstrate whether the student could use the equation editor tool. The second and third items would provide evidence of whether the ability to use the tool interacted with item difficulty.

Results

Between 15 and 30 percent of the students indicated that they had difficulty using the equation editor. About 30 percent had trouble just copying the answer, as required by item 1. The examiners assessed that 35 percent had difficulty using the equation editor and that only 40–57 percent of the students would get a given item correct. Students had more difficulty with the more challenging items. A summary of representative comments made by students about the equation editor during the administration of the think-aloud protocol is presented below:

1. Clicked on the + sign, but it didn't work, twice.
2. How do I choose the numbers?
3. I needed paper to make a picture.
4. How do I use the number pad?
5. I tried to use the numbers on the keyboard, but wouldn't work.
6. Some symbols didn't respond to first click.
7. I had trouble getting bottom half of fraction to record.
8. Unclear what possible value meant.
9. I didn't see decimal point down there [due to scrolling].
10. Couldn't find x symbol.
11. Unclear whether to click and drag or type.
12. Would rather type than use a mouse.
13. Difficult to use fraction tool.

Summary

Elementary students had some difficulty using the equation editor. Between 15 and 30 percent of the students indicated that they had difficulty using the equation editor. The examiner's assessment

concurred that about 35 percent had difficulty using the equation editor and that about 50 percent of the students would get a given item correct.

Research Question 12: Intuitive understanding of the relationships in multiplying fractions.

This question is designed to assess whether students with a strong understanding of fractions and the multiplication and division of fractions complete the items without performing the indicated multiplication. The task asked students to compare the size of a product to the size of one factor, on the basis of the size of the other factor, without performing the indicated multiplication. Also of interest was whether students who complete an item as intended (without using multiplication) spent less time on an item than those who did not. To investigate this question a single form was administered for grades 3–5.

Results

The form was administered to 33 students at the elementary level. Table 37 compares those with a strong understanding of fractions with those who do not have a strong understanding of fractions and whether they completed the task with or without using multiplication. There does not appear to be a relationship between strength of understanding of fractions (multiplication and division) and whether they used multiplication to solve the problems.

Table 37. Strength of Understanding of Fractions and Whether Multiplication was Performed

	Not Strong Understanding of Fractions		Strong Understanding of Fractions	
Item Number	Performed Multiplication	Did not Perform Multiplication	Performed Multiplication	Did not Perform Multiplication
1	9	7	8	1
2	9	8	9	1
3	10	6	6	1
4	6	7	10	1
5	7	7	9	1
6	4	6	15	0

Table 38 presents descriptive statistics for the timing of each item (in seconds). In addition to means, medians are reported because timing distributions tend to be highly skewed. On average, those who did not have to perform the multiplication completed the items in less time. The results for item 6 were comparable for the two groups.

Table 38. Comparison of the Time to Complete the Item for Those Who Did not Use Multiplication to Solve the Item and Those Who Did

	Performed Multiplication				Did not Perform Multiplication			
Item Number	Mean	Std Dev	Median	Range	Mean	Std Dev	Median	Range
1	210	136	179	59–543	136	90	114	53–360
2	145	119	106	36–420	126	110	89	30–336
3	75	104	42	10–480	34	28	25	3–90
1	123	111	70	21–480	88	69	57	25–195
2	133	130	95	28–480	79	67	68	9–185
3	69	118	32	4–540	65	63	51	3–170

Table 39 shows the percentage of students answering the item correctly. The students tested generally found the items to be difficult. ("Multiplying fractions was hard.") Some students did not understand the inequality signs, while others did not understand improper fractions or how to make a whole number into a fraction. One interviewer commented that the "student had little or no understanding of fractions."

Table 39. Percentage of Students Answering an Item Correctly.

Item Number	Percent
1	17
2	20
3	28
4	42
5	26
6	33

About 69 percent of the students used multiplication to solve the problems (Table 40). Student comments support this. "I multiplied… each box and put them in the correct boxes (columns)." "I timesed [sic] the numbers." "I looked at each number expression and multiplied it in my head and moved it to where I thought it was right." "Some numbers on the bottom depends on the top number which is bigger or smaller." Only about 40 percent of the students understood fractions or at least the multiplication of fractions. The examiner's comments (Table 41) concur with this conclusion.

Table 40. Percentage of Students Using Multiplication to Solve the Items

Item Number	Yes
1	68
2	70
3	75
4	72
5	73
6	81

Table 41. Interviewer's Assessment of: (1) Whether the Student Used Multiplication and (2) Whether the Student Had a Strong Understanding of Fractions

Summary	Percent
Did student use multiplication?	72
Did student have a strong understanding of fractions (multiplication/division)?	40

Summary

There seemed to be little relationship between whether a student has a strong understanding of the multiplication and division of fractions and whether he or she used multiplication to solve the items. However, students who did not have to perform the multiplication completed the items in less time than students who had to perform the multiplication. While most students said they understood the questions, 70 percent had to use multiplication to solve them. Only about 40 percent of the students had a firm understanding of the multiplication/division of fractions, according to the interviewers.

Special Populations

Research Question 13: Contextual glossaries are item-specific glossaries that provide a definition of a word that is targeted to, and appropriate for, the context in which the word is used in the item. Are these a fair and appropriate way to support students who need language support?

This question addressed the efficacy of the use of contextual glossaries with non-native (Spanish) speakers (see Exhibit 4 for an example of a contextual glossary item) when solving mathematics problems. A contextual glossary item contains highlighted words when presented online. Clicking any of these highlighted items produces a list of all highlighted words in the item with Spanish definitions for each. Two sets of items were created that were parallel in difficulty. The first set of items contained no contextual glossaries with only single words translated. The second set of items

contained contextual glossaries. The interviewer was asked to determine whether the student was having trouble understanding a word and whether the contextual glossary aided in the interpretation of the word or sentence.

Only three ELL students participated: one from grade 3 and two from grade 6.

Exhibit 4. Example of a Contextual Glossary Item

1. A roller coaster has a large rise and drop followed by a complete circle. The following diagram shows measurements for the track. An extra 20 feet are needed for cutting and welding. How many feet of track should be ordered? (Use π = 3.14)

A. 280 feet

B. 407 feet

C. 415.6 feet

D. 1,537.4 feet

Glossary Window

Roller coaster
montaña rusa

Rise
subida

Drop
bajada
caída

Complete
completo
entero

Diagram
diagrama
quema
gráfico

Track
vía
riel

Cutting
cortar

Welding
cortar

Results

The grade 3 student had trouble understanding a few items, but had few word confusions. For the second set of items, this student used the contextual glossaries for one item but not for the other items. The student said that there was not a problem understanding the items because the student used “sentence context” to answer them, or the words the student didn't know weren't in the glossary so the student stopped using it. In terms of scoring, this student answered two of the three “translated” items correctly, but did not answer any of the “contextual glossary” items correctly, so the results are difficult to interpret as to whether the use of contextual glossaries aided the students’ performance.

The two grade 6 students (one ELA form and one Math form) both had difficulty with the “translated” items in the first set with six or more word confusions each for most items. Both students found the contextual glossary useful to some degree, though not for all items. (“The words I don’t know aren’t in the glossary.”) However, the interviewers suggested that the use of the contextual glossary improved the performance for both grade 6 students. Though the ELA student got all questions incorrect, the interviewer believed that this was mainly due to careless mistakes and that the student used the glossary to help make sense of the key components of the questions and understood the procedures for answering the questions. The math student got two-thirds of the items correct when the items were translated, and one-third of the items correct when the contextual glossary was used. The student had difficulty understanding an essential word in one of the incorrect items. However, the interviewer commented that once he understood the words, he could confidently work on the problem and he knew how to proceed.

Summary

In summary, contextual glossaries appeared to be somewhat effective when they were used, but the impact was not always reflected in the score the student received for an item. The contextual glossaries appeared to be incomplete in that they did not include words that the students needed. This limited the use of the glossaries in these situations. Interviewer’s comments suggested that performance was improved when the students used the contextual glossaries.

Research Question 14: Under what conditions do students with lower reading ability use text-to-speech (TTS) to help focus on content in ELA and mathematics? Is this affected by the quality of the voice-pack?

TTS is a technology that can give students with low reading ability access to an assessment. For this technology to be effective the language produced from the voice-pack must be clear so that it can be understood. This is particularly true for non-native speakers of English.

This question is designed to assess whether students with lower reading ability and non-native speakers of English use TTS to help focus on content in ELA and mathematics. Only students familiar with TTS were included in the study. Overall, 77 students used TTS at least once. Among them, 58 students are LEP students, 13 students had reading difficulties (IEP), and six students were Gen Ed students.

Forms were constructed at three grade bands: grade band 3 (referred to as grade band 3), grade band 6 and 7 (referred to as grade band 6), and grade 11 (referred to as grade band 11). In ELA, four forms were administered with both high- and low-quality voice-packs. In mathematics, two forms were administered in grade bands 3 and 11. Only a single form was administered in grade band 6. For all mathematics forms only high-quality voice-packs were administered. In Tables 42–45, yellow shading denotes the use of high-quality voice-packs while a white background denotes the use of a low-quality voice-pack.

Results

For ELA (Table 42), for all groups and grade bands, a high percentage of students tended to make comments indicating an improved focus on the content when the voice-pack was of high quality. About one-third of the students (except the Gen Ed grade band students) indicated that TTS kept their focus on content even when low-quality voice-packs were used. For ELA, students in all groups tended to make greater use of TTS when the voice-pack was of high quality.

About 50 percent of the LEP students in mathematics in grade bands 3 and 11 made comments indicating that TTS helped them focus on content. All of the LEP grade band 6 group and the IEP students in grade band 3 found that TTS helped them focus on content. (“It made me think about the question.”) The Gen Ed students in grade band 3 found that TTS helped them focus on content; however, the Gen Ed grade band 6 students did not find TTS useful.

Table 42. Percentage of TTS Students Who Made Any Comment Indicating That He/She Is Mainly Focused on the Content of the Item, by Content, Voice-Pack Quality, Sample, and Grade Band

			LEP			IEP			Gen Ed		
Content	Voice Pack Quality	Grade Band	3	6	11	3	6	11	3	6	11
ELA	Low			32	39	35				0	
	High		36	67	100	100					100
Mathematics	Low										
	High		50	100	60	100			100	0	

Table 43 shows the percentage of students who answered the items correctly, averaged across items. In ELA, the grade band 6 and 11 LEP students and the grade band 3 IEP students found the items more difficult using a low-quality voice-pack. The Gen Ed grade band 6 ELA students were not administered a high quality voice-pack. In the LEP grade band 6 group, about half the students answered an item correctly using the high-quality voice-pack. The percentage answering an item correctly was close to 75 percent for the other LEP grade bands and the grade band 3 low-level reading students when the high-quality voice-pack was used.

In mathematics, in grade band 3, about 40 percent of the LEP students answered an item correctly. For the other grade bands, for the LEP and IEP samples, no items were answered correctly, even with the high-quality voice-packs. This was also true for the Gen Ed grade band 3 students. However, the general education students in grade band 6 answered all the items correctly.

Table 43. Percentage of TTS Students Who Answered the Items Correctly by Content, Voice-Pack Quality, Sample, and Grade Band

			LEP			IEP			Gen Ed		
Content	Voice Pack Quality	Grade Band	3	6	11	3	6	11	3	6	11
ELA	Low			14	0	50				75	
	High		77	50	80	75					0
Mathematics	Low										
	High		40	0	0	0			0	100	

Tables 44 and 45 summarize the interviewer's assessment for ELA and mathematics related to whether TTS improved access to the content or was a distraction. TTS improved access in ELA regardless of the quality of the voice-pack. Greater access was achieved when high-quality voice-packs were used in ELA except in grade band 11. This is probably an artifact of the very small sample size. The low-quality voice-pack appeared less effective at providing access and was distracting in ELA, where the high-quality voice-pack was not distracting at all. One student said, "[I] didn't like using TTS … the sound was robotic and would break my concentration."

In mathematics, TTS helped to improve access for some grade band 3 LEP students, but not for middle- and upper-level LEP students or the IEP or Gen Ed grade band 3 students. All the Gen Ed, IEP, and grade band 6 LEP students found the high-quality voice-pack distracting in mathematics. This was in part a function of trying to describe a table verbally. ("When TTS read the chart aloud, I got lost in the numbers and couldn't figure out what the question was asking.")

Table 44. Assessment by the Interviewer of the Percentage of TTS Students Whose Access to Content Was Improved by the Use of TTS by Content, Voice-Pack Quality, Sample, and Grade Band

			LEP			IEP			Gen Ed		
Content	Voice Pack Quality	Grade Band	3	6	11	3	6	11	3	6	11
ELA	Low			57	75	79				100	
	High		76	100	33	100					100
Mathematics	Low										
	High		43	0	0	0			0	0	

Table 45. Assessment by the Interviewer of the Percentage of TTS Students Who Were Distracted by TTS, by Content, Voice-Pack Quality, Sample, and Grade Band

			LEP			IEP			Gen Ed		
Content	Voice Pack Quality	Grade Band	3	6	11	3	6	11	3	6	11
ELA	Low			12	20	33				0	
	High		0	0	0	0					0
Mathematics	Low										
	High		44	100	40	0			100	100	

Summary

TTS improved access in ELA regardless of the quality of the voice-pack. Greater access was achieved when high-quality voice-packs were used. LEP students and students with reading difficulties tended to benefit more from the use of TTS. Using TTS with high-quality voice-packs improved focus on content in ELA. The use of TTS with low-quality voice-packs tended to distract students in ELA, whereas high-quality voice-packs did not. In mathematics, access was improved only for grade band 3 students. All the Gen Ed, IEP, and grade band 6 LEP students found the high-quality voice-pack distracting. This was in part a function of trying to describe a table verbally.

Final Summary

Smarter Balanced is moving toward an assessment model that is largely scored automatically and delivered adaptively on computer. The Smarter Balanced cognitive laboratories were conducted to investigate questions that arise from such an automated design. While think-aloud protocols are time consuming, they have the potential to provide a level of information not easily accessed through large-scale studies. However, the sample sizes are small. Therefore, should a more rigorous investigation of any of the research questions be of interest, specifically designed studies with large samples will be needed.

This report presents the results from 14 small think-aloud studies that addressed topics that pertain to an automated test delivery system.

1. Can non-constructed-response item formats assess components that have historically been believed to be measured only with CR items?
2. What is the optimal amount of direction to provide for TE items? Does this vary with grade level?
3. What is the appropriate degree of labeling to provide for MPSR items so that students know to complete all parts?
4. Does it matter whether items associated with a passage are presented in a single block or presented one item at a time? Are ELL students impacted by these different arrangements?
5. Do the longer passages favored by Smarter Balanced reduce student engagement?
6. How much time do items in different formats take to answer? Are ELL students affected more than general education students?
7. In mathematics, could information captured on scratch paper facilitate the working of a problem and benefit the performance and scoring of a student?
8. Do contextual glossaries help improve the performance of students with language disabilities?
9. Does TTS help focus students of low reading ability on the content of an item?
10. Can younger students effectively use the equation editor?
11. Mathematics intuition: Can students compare the size of a product to the size of one factor, on the basis of the other factor without multiplying?

On the whole, the cognitive laboratories were successful in providing answers to most of these questions. They provide a glimpse of issues that may exist and need to be investigated further. To investigate these issues more completely, larger-scale studies should be conducted.

Appendix A

Question 2. Full Claim Descriptions

Content	Content Grade	Claim	Claim Description
ELA	3–5	1	Students can read closely and analytically to comprehend a range of increasingly complex literary and informational text.
ELA	3–5	2	Students can produce effective writing for a range of purposes and audiences.
ELA	3–5	3	Students can employ effective speaking and listening skills for a range of purposes and audiences.
ELA	3–5	4	Students can engage in research/ inquiry to investigate topics and to analyze, integrate, and present information.
ELA	6–8	1	Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.
ELA	6–8	2	Students can produce effective writing for a range of purposes and audiences.
ELA	6–8	3	Students can employ effective speaking and listening skills for a range of purposes and audiences.
ELA	6–8	4	Students can engage in research/ inquiry to investigate topics and to analyze, integrate, and present information.
ELA	9–12	1	Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts.
ELA	9–12	2	Students can produce effective and well-grounded writing for a range of purposes and audiences.
ELA	9–12	3	Students can employ effective speaking and listening skills for a range of purposes and audiences.
ELA	9–12	4	Students can engage in research/inquiry to investigate topics, and to analyze, integrate, and present information.
Math	3–5	1	Students can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.

Content	Content Grade	Claim	Claim Description
Math	3–5	2	Students can solve a range of well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies.
Math	3–5	3	Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.
Math	3–5	4	Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.
Math	6–8	1	Students can explain and apply mathematical concepts and carry out mathematical procedures with precision and fluency.
Math	6–8	2	Students can solve a range of well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies.
Math	6–8	3	Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.
Math	6–8	4	Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.
Math	9–12	1	Students can explain and apply mathematical concepts and carry out mathematical procedures with precision and fluency.
Math	9–12	2	Students can solve a range of well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies.
Math	9–12	3	Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.
Math	9–12	4	Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.

Question 2. Full Target Descriptions

Content	Grade Content	Grade Band	DOK	Claim	Target	Target Description
ELA	3	3	2	2	3	Write or revise one or more informational/explanatory paragraphs demonstrating ability to organize ideas by stating a focus, including appropriate transitional strategies for coherence, or supporting details, or an appropriate conclusion.
ELA	6	6	2	1	11	Use supporting evidence to justify interpretations or analyses of information presented or how information is integrated within a text (point of view; interactions among events, concepts, people, or ideas; author's reasoning and evidence).
ELA	6	6	3	1	11	Use supporting evidence to justify interpretations or analyses of information presented or how information is integrated within a text (point of view; interactions among events, concepts, people, or ideas; author's reasoning and evidence).
ELA	7	6	2	2	3	Apply a variety of strategies when writing or revising one or more paragraphs of informational/explanatory text organizing ideas by stating and maintaining a focus/tone, providing appropriate transitional strategies for coherence, developing a topic including relevant supporting evidence/vocabulary and elaboration, or providing a conclusion appropriate to purpose and audience.
ELA	7	7	2	1	1	Identify explicit textual evidence to support inferences made or conclusions drawn.
ELA	8	7	2	1	1	Identify explicit textual evidence to support inferences made or conclusions drawn.
ELA	8	7	2	2	6	Apply a variety of strategies when writing or revising one or more paragraphs of informational/explanatory text organizing ideas by stating and maintaining a focus/tone, providing appropriate transitional strategies for coherence, developing a topic including relevant supporting evidence/vocabulary and elaboration, or providing a conclusion appropriate to purpose and audience.
ELA	11	11	2	1	1	Cite explicit textual evidence to support inferences made or conclusions drawn about texts.

Content	Grade Content	Grade Band	DOK	Claim	Target	Target Description
ELA	11	11	2	1	7	Determine or analyze the figurative (e.g., euphemism, oxymoron, hyperbole, paradox), or connotative meanings of words and phrases used in context and the impact of those word choices on meaning and tone.
MATH	3	3	2	1	F	Develop understanding of fractions as numbers.
MATH	4	3	2	1	F	Extend understanding of fraction equivalence and ordering.
MATH	4	4	2	1	L	Draw and identify lines and angles, and classify shapes by properties of their lines and angles.
MATH	4	4	3	4	A	Apply mathematics to solve problems arising in everyday life, society, and the workplace.
MATH	6	6	2	1	E	Apply and extend previous understandings of arithmetic to algebraic expressions.
MATH	6	6	3	2	A	Apply mathematics to solve well-posed problems arising in everyday life, society, and the workplace.
MATH	11	11	2	1	I	Solve equations and inequalities in one variable.
MATH	11	11	2	2	A	Apply mathematics to solve well-posed problems arising in everyday life, society, and the workplace.
MATH	11	11	3	2	A	Apply mathematics to solve well-posed problems arising in everyday life, society, and the workplace.

Appendix B

Demographic Information for Cognitive Laboratories

Total Number of Students: 774

By Cognitive Lab Location:
San Francisco, California: 80 (10%)
Monterey, California: 167 (22%)
Waterbury, Connecticut: 45 (6%)
Hartford, Connecticut: 26 (3%)
Pocatello, Idaho: 64 (8%)
District of Columbia: 31 (4%)
Honolulu, Hawaii: 43 (6%)
East Lansing, Michigan: 63 (8%)
Madison Heights, Michigan: 33 (4%)
Marquette, Michigan: 30 (4%)
Des Moines, Iowa: 52 (7%)
Pittsburgh, Pennsylvania: 76 (10%)
Columbia, South Carolina: 50 (6%)
Portland, Oregon: 14 (2%)

By School Location:
California: 243 (31%)
Connecticut: 71 (9%)
District of Columbia: 14 (2%)
Hawaii: 43 (6%)
Idaho: 64 (8%)
Iowa: 52 (7%)
Maryland: 12 (2%)
Michigan: 126 (16%)
Nevada: 4 (<1%)
Oregon: 12 (2%)
Pennsylvania: 76 (10%)
South Carolina: 50 (6%)
Virginia: 5 (<1%)
Washington: 2 (<1%)

By Grade:
Grade 3: 113 (15%)
Grade 4: 100 (13%)
Grade 5: 79 (10%)
Grade 6: 98 (13%)
Grade 7: 113 (15%)
Grade 8: 62 (8%)

Grade 9: 87 (11%)
Grade 10: 70 (9%)
Grade 11: 44 (6%)
Grade 12: 8 (1%)

By Gender:
Male: 393 (51%)
Female: 381 (49%)

Language(s) Spoken at Home:
English: 670 (87%)
Spanish: 100 (13%)
Chinese: 46 (6%)
Chaldean: 21 (3%)
Arabic: 18 (2%)
Albanian: 15 (2%)
Tagalog: 10 (1%)
German: 5 (<1%)
Vietnamese: 5 (<1%)
Hindi: 4 (<1%)
Korean: 4 (<1%)
Japanese: 3 (<1%)
Samoan: 3 (<1%)
Bengali: 2 (<1%)
Greek: 2 (<1%)
Ilocano: 2 (<1%)
Telegu: 2 (<1%)
Other: 14 (2%)
*Total percentage is more than 100% because more than one response could be selected.

Language(s) Most Frequently Spoken:
English: 707 (91%)
Arabic: 22 (3%)
Chinese: 18 (2%)
Chaldean: 16 (2%)
Spanish: 13 (2%)
Albanian: 3 (<1%)
Greek: 2 (<1%)
Tagalog: 2 (<1%)
Other: 7 (1%)
*Total percentage is slightly over 100% because some parents added an additional language in the comment section.

Type of School:
Public: 681 (88%)
Private: 42 (5%)
Charter: 18 (2%)
Home School: 14 (2%)
Parochial: 13 (2%)
Other: 4 (<1%)

Access to a Computer at Home:
Yes: 747 (97%)
No: 27 (3%)

Frequency of Computer Use:
Almost every day or every day: 438 (57%)
Three or four times per week: 175 (23%)
Once or twice per week: 146 (19%)
Never: 15 (2%)

Frequency of Internet Use:
Almost every day or every day: 401 (52%)
Three or four times per week: 189 (24%)
Once or twice per week: 166 (21%)
Never: 18 (2%)

Computer Classes:
Yes: 385 (50%)
No: 321 (41%)
Unsure: 68 (9%)

IEP:
Yes: 87 (11%) (e.g., ADHD, Dyslexia, Emotional Disturbance, Gifted, Hearing Loss, High Functioning Asperger's, Impaired/Slow Learning, Auditory Processing Disability, Orthopedic Impairment, Speech and Language, Speech Impairment)
No: 631 (82%)
Unsure: 56 (7%)

Testing Accommodations:
Yes: 83 (11%) (e.g., Paper Test, Printable Test, Student can take test in another language, ELD, Limited English Proficiency, Listen to questions on tape and use bilingual dictionary, Supervised breaks and additional time, Assessments can be read, Assessments one on one with administrator, Cantonese Bilingual Pathway Instruction, Extra time and modified questions, Extended response time, Separate room)
No: 647 (84%)
Unsure: 42 (5%)

There is no assessment program at this grade level: 1 (<1%)
Child does not participate in the school's testing or assessment program: 1 (<1%)

ELA Grades:
Above Average: 375 (48%)
Average: 324 (42%)
Below Average: 51 (7%)
Unsure: 20 (3%)
*Not all participants responded to this question.

Mathematics Grades:
Above Average: 392 (51%)
Average: 311 (40%)
Below Average: 55 (7%)
Unsure: 14 (2%)
*Not all participants responded to this question.

Ethnic/Cultural Breakdown:
White: 493 (64%)
Hispanic: 137 (18%)
Asian: 125 (16%)
Black/African American: 76 (10%)
Native Hawaiian or Other Pacific Islander: 28 (4%)
American Indian or Alaskan Native: 17 (2%)
Filipino: 12 (2%)
Asian Indian: 5 (<1%)
Other: 3 (<1%)
*Total percentage is over 100% because more than one response could be selected.

Household Income:
Under $25,000: 135 (17%)
Between $25,001 and $50,000: 170 (22%)
Between $50,001 and $75,000: 139 (18%)
Between $75,001 and $100,000: 145 (19%)
Between $100,001 and $150,000: 110 (14%)
Over $150,001: 54 (7%)