Cognitive Lab Evaluation of Innovative Items in Mathematics and English Language Arts Assessment of Elementary, Middle, and High School Students

Research Report

Robert P. Dolan
Joshua Goodman
Ellen Strain-Seymour
Jeremy Adams
Sheela Sethuraman

March 2011

ALWAYS LEARNING

PEARSON

About Pearson
Pearson, the global leader in education and education technology, provides innovative print and digital education materials for preK through college, student information systems and learning management systems, teacher licensure testing, teacher professional development, career certification programs, and testing and assessment products that set the standard for the industry. Pearson's other primary businesses include the Financial Times Group and the Penguin Group. For more information about the Assessment & Information group of Pearson, visit http://www.pearsonassessments.com/.

About Pearson's Research Reports
Pearson's research report series provides preliminary dissemination of reports and articles prepared by TMRS staff, usually prior to formal publication. Pearson's publications in .pdf format may be obtained at: http://www.pearsonassessments.com/research.

Abstract

Traditional large-scale assessments, with their reliance on selected-response, static text- and image-based items, may be limited in their ability to support valid interpretations about students' knowledge, skills, and abilities that require applying critical thinking skills, involve complex processes, and/or are highly contextualized. As a result, current large-scale testing may not be well suited to meet the needs of more rigorous content standards that involve higher order thinking skills, collaboration, information literacy, etc., such as those emphasized by the Common Core State Standards initiative. A promising solution to improving our ability to assess students' deeper, more complex knowledge, skills, and abilities is likely to come through increased use of innovative computer-based items. Innovative items are delivered online and are designed to incorporate pedagogically and cognitively sophisticated features or functionalities beyond those of traditional text- and image-based, multiple-choice items. Innovative items can also allow for online administration of nontextual constructed-response items; for example, innovative items allow transitions from paper-and-pencil to online testing without losing functionality.

Prior to using innovative items in large-scale, K-12 testing programs, it is necessary to gather initial empirical data to evaluate claims that innovative items can improve measurement without introducing new barriers to accurate testing. In the current study, a set of prototype mathematics and English language arts (ELA) items at three different grade bands—elementary, middle, and high school—were developed to align with specific Common Core State Standards. These items were then administered to selected samples of students in a series of cognitive labs. Results suggest that the innovative item types included in this study align well to the intended constructs and provide flexible and authentic tasks that allow for the assessment of knowledge, skills, and

abilities that have traditionally been difficult to measure. The innovative items were generally well received by students, and students encountered few usability challenges when interacting with the items; however, students required more time to complete the innovative items in comparison to traditional, selected-response items. Implications for future research and recommendations for appropriate use of these item types based upon these initial findings are provided.

Keywords: Common Core State Standards, computer-based testing, innovative items, cognitive labs, large-scale assessment

Acknowledgements

The authors wish to thank Aaron Nance, Mollie Kilburg, Dee Heiligmann, Mary Veazey, Jay Larkin, Amy Hathorn, Mike Harms, and Denny Way, without whose expertise and diligence this study could not have been accomplished. We'd also like to thank Kimberly O'Malley, Katie McClarty, Ye Tong, and Shelley Ragland for their extensive feedback and valuable insights while preparing this report.

Cognitive Lab Evaluation of Innovative Items in Mathematics and English Language Arts Assessment of Elementary, Middle, and High School Students

Introduction

Text- and image-based multiple-choice questions are the bread and butter of large-scale educational testing. However, when measuring knowledge, skills, and abilities (KSAs) that require applying critical thinking skills and/or involve complex processes—or rely on tasks that involve multiple steps and/or are highly contextualized—some would argue that such traditional item types are limited in their ability to support valid interpretations due to construct underrepresentation (Messick, 1989; Messick & Hakel, 1998; Miller & Linn, 2000; Hambleton, 2000). Traditional performance or task-based assessments improve the situation by expanding the depth and complexity of what students are asked to do (Bachman, 2002; Chalhoub-Deville, 2001; Kuechler & Simkin, 2010; Lane, 2010; Darling-Hammond & Adamson, 2010; Messick 1994; Messick 1995), but may still provide inadequate domain coverage and are limited in the degree to which they can be standardized and consistently scored (Messick 1994; Linn, & Burton, 1994; Kane, Crooks, & Cohen, 1999). Furthermore, except for text-based constructed responses, such performance assessments are not easily administered within the online large-scale testing systems generally in use today. As a result of these factors, most current large-scale testing is not well suited to meet the needs of more rigorous, content standards involving higher order thinking skills, collaboration, information literacy, etc., such as those emphasized by the Common Core State Standards (CCSS) initiative[1] and the Partnership for 21st Century Skills[2].

[1] http://www.corestandards.org.
[2] http://www.p21.org.

One solution to improving our ability to assess students' deeper, more complex KSAs is likely to come through increased use of innovative computer-based items. Innovative items can be designed to incorporate pedagogically and cognitively sophisticated features or functionalities beyond those of traditional text- and image-based multiple-choice items (Bennett et al., 1999; Bennett, 1999; Parshall, Harmes, Davey, & Pashley, 2010). Innovative items can also allow for online administration of nontextual constructed-response items; for example, innovative items allow transitions from paper-and-pencil to online testing without losing functionality.

Examples of innovative items include the use of multimedia within item stimuli, tasks that would be natural and traditional on paper (e.g., geometric constructs that require a compass), or complex performance exercises that integrate multiple steps and skills. Innovative items have the potential to improve qualitative and quantitative properties of K-12 educational assessments (Kane, 1992; Zenisky & Sireci, 2002; Jodoin, 2003) and the inclusion of these items on tests have been met with enthusiastic support from students, teachers, and policy makers (e.g., Strain-Seymour, Way, & Dolan, 2009; Wendt & Harmes, 2009). Furthermore, use of technology can allow for more fair and inclusive testing by incorporating flexible interfaces—often already used instructionally—to reduce barriers to students with disabilities and English learners; in fact, well designed interfaces, especially when coupled with universal design principles, can be beneficial for all students by providing enhanced opportunities to demonstrate their construct-relevant KSAs (Almond et al., 2010; Dolan & Hall, 2001; Ketterlin-Geller, 2005; Thompson, Johnstone, & Thurlow, 2002).

There are multiple mechanisms by which innovative items can potentially improve large-scale assessment:

- Innovative items allow for more explicit measurement of aspects of the domain of skills, such as through use of complex content (e.g., polytomously scored, multistep mathematics problems in which students build off their own responses), application of higher order thinking skills (e.g., ability to design a scientific experiment), and aspects of the existing construct that elude capture by traditional selected response items, including speaking or listening in the English language arts (ELA) curriculum.
- Innovative items can dually represent authentic, real-world tasks while aligning more closely with classroom instructions (Bennett, 1993, 1999) than traditional item types.
- By using multimedia to provide flexible representations of information and flexible modes of response, innovative items can reduce the impact of construct-irrelevant factors, such as reading and writing abilities in non-ELA tests.
- Innovative items can facilitate administration of both text-based and nontext-based constructed-response items during online testing, and thus can reduce the impact of successful guessing by students.
- Innovative items are likely more engaging to students than traditional test items and may improve student affect to the point where they are better able to access their construct-relevant KSAs (Kumar, White & Helgeson, 1993; Huff & Sireci, 2001; Scalise & Gifford, 2006).
- Innovative items that allow for automated scoring of constructed-response items can decrease both the expense and time constraints of human-scored constructed responses while offering richer measurement than text-based multiple-choice items (Scalise & Gifford, 2006).

Given the potential advantages and benefits associated with innovative items, it comes as no surprise that state and national testing programs are expressing interest in developing these items for use in their large-scale assessment programs. The aforementioned CCSS initiative has increased the momentum to develop a pool of innovative items that are aligned to new educational standards in mathematics and ELA. The two state consortia that were awarded Race to the Top Assessment funds to develop CCSS assessments—the Partnership for the Assessment of Readiness for College and Careers (PARCC) and the SMARTER Balanced Assessment Consortium (SBAC)—emphasize the need for a wide array of innovative items. The overwhelming message from consortia, state testing organizations, school administrators, and curriculum experts is a call for a new generation of assessments. These assessments must engage students, cover a wide variety of performance and content standards, be authentic, provide rich descriptions of student performance, and give useful and timely information to teachers and schools.

Though innovative items show promise for creating the new generation of assessments, they come with several sets of challenges that must be considered before relying on innovative items in large-scale K-12 testing programs. Technical challenges include how to develop items in timely and cost effective ways, how to represent them in a manner that allows interoperability across systems, how to deliver them consistently and accurately, and how to develop user interface designs to match students' instructional and study practices. Psychometric challenges include building equivalent test forms and developing scoring and equating models that can handle multipart items in which latter responses build upon former responses. Most critically at this early stage, however, is the challenge of producing empirical data to support claims that innovative items can improve measurement. Although research has demonstrated the value of

traditional performance-based assessment and text-based constructed-response items, only recently has limited evidence emerged providing empirical evidence supporting the use of innovative technologies to create items with pedagogically and cognitively sophisticated features or functionality (e.g., Bennett, Persky, Weiss, & Jenkins, 2010; Quellmalz & Pellegrino, 2009; Rupp, Gushta, Mislevy, & Shaffer, 2010). In fact, there is reason to believe that unless properly designed, increases in task complexity that accompany many innovative item designs could compromise validity by increasing construct-irrelevant variance (Dolan, Rose, Burling, Harms, & Way, 2007). Thus, a program of research is necessary so that innovative items can be evaluated, first for suggestions of efficacy, then efficacy itself through pilot testing, and finally value in large-scale testing programs as proven through field testing.

Purpose

This research was part of a long-term innovative item development and evaluation effort that includes a cross-functional team of content specialists, usability experts, instructional and user experience designers, cognitive scientists, and psychometricians. This preliminary phase of research, at its broadest level, was intended to produce initial evidence for the effectiveness and applicability of innovative items when testing students on CCSS constructs and to greater depths of KSAs not readily assessable using traditional items. To this end, a set of prototype mathematics and ELA items at three different grade bands—elementary, middle, and high school—were developed to align with specific CCSS[3]. These items were then administered to selected samples of students in a series of cognitive labs, and the qualitative results were analyzed to answer the following specific questions:

1. What are the students' perceptions of and reactions to these innovative items?
2. What are the cognitive processes employed by students while responding to innovative items, and do they align with the intended constructs?
3. How do the cognitive processes employed by students when responding to innovative item compare to the processes used while responding to corresponding traditional items?
4. Can evidence be gathered suggestive that innovative items are better at evaluating constructs that are traditionally difficult to assess in standardized formats (e.g., critical thinking skills, mathematical construction, speaking, and collaboration)?
5. Are there usability issues that impact how students interact with innovative items and to what degree are usability issues a potential source of construct-irrelevant variance?

It should be emphasized that this study represented only an initial phase of the longer term innovative item research initiative. As such, only a limited sample of items and students were used in the current study, since the intent was neither to assess the statistical performance of these specific items nor to strongly generalize to the K-12 student population. Instead, this study examined the successes and limitations of the item types by using a carefully selected, small student sample. These first investigations will, in turn, help foster a better understanding of the cognitive and design schema that should be used when producing greater numbers of innovative items for large-scale deployment. The process will be iterative and interactive and will build in scope and scale as we learn more about designing, delivering, and scoring items with new and innovative formats. Thus, we hope this will lead to the development and application of a systematic approach for developing families of innovative items that behave in a predictable manner and align closely with a predefined set of content standards.

---

[3] Additional prototype innovative items have been developed to assess science but are not included in this phase of the research.

Methods

Cognitive labs, also referred to in practice as think-aloud protocols or protocol analysis, are frequently used to investigate the processes employed by examinees when responding to test items. These methods typically present examinees with an item or task and ask them to work through the task while speaking aloud their thought process (Ericsson & Simon, 1993). The collected data are used to evaluate the steps and processes that underlie an examinee's response and to investigate how different features of test items impact examinees' response processes. Cognitive labs have been successfully used to explore the cognitive processes in various educational assessment contexts that include support for diagnostic models (Gierl, Wang, & Zhou, 2008), differential item functioning (Ercikan et al., 2010), and differential processing in English learners (Johnstone, Bottsford-Miller, & Thompson, 2006).

A two-step process is recommended when conducting cognitive labs (Branch, 2000). In step one, a researcher collects data in real time by simply asking students to think aloud while completing the task. Researchers should probe students as infrequently as possible during this step to prevent any distractions during the completion of the task or inadvertently leading a student to a particular problem solving approach (Ericsson & Simon, 1993). If, during the real-time step, the participant is silent for several seconds, researchers can prompt with neutral cues like "keep talking" or "what are you thinking now?" In reality, the real-time step of a cognitive lab is often challenging because the produced verbalizations are often incoherent and lack context, and/or because the cognitive load of problem solving and speaking the process aloud may be too much for participants (Branch, 2000). For this reason, the second step in a cognitive lab is often a retrospective interview completed after a student has completed the task. The follow-up questions in this second step can be used to complement the primary data source by

clarifying data from student verbalizations and supplementing the real-time data by asking students directly about particular areas of interest in the study that may not have been covered during the first step.

In the context of this research, cognitive lab protocols assist in revealing (1) the cognitive schema students actually employ when responding to innovative items as compared to expected and preestablished cognitive pathways and (2) the degree to which the enhanced functionality of different item types impacts student responses.

Certain types of usability testing also commonly rely on the think-aloud protocol with very similar guidelines regarding minimizing facilitator interference. The qualitative data produced by such a protocol generally allow for differentiation between struggles with the interface and construct-related challenges. Thus, the potential for the cognitive lab protocol to generate both types of data—usability-related and construct-related—was regarded as a benefit rather than a confounding factor. The inclusion of usability findings was also appropriate to this validity study because usability failings and hefty requirements for computer familiarity constitute one of the greatest risks in terms of introducing construct irrelevant variance within innovative items.

For this study, a sample catalog of innovative items were developed at the elementary, middle, and high school levels in the two CCSS content areas (ELA and mathematics). The six resulting clusters of items (three grade bands fully crossed with the two content areas) were then administered to six appropriately aged participants in a series of hour-long cognitive lab sessions, for a total of 36 cognitive labs.

Items

The innovative items in this study were developed with the intent of expanding the boundaries of standardized measurement. This was done by designing items that (1) measure traditionally hard-to-assess content standards and skills (e.g., research, speaking, and listening skills), (2) capture activities and tasks that students engage in instructionally but have traditionally been difficult to capture within a large-scale assessment (e.g., geometric constructions, graphing, plotting, creating nets, constructing an equation), (3) integrate several skills (e.g., watch and read a story, place graphical depictions of the story's events in sequence, and write a summary), and/or (4) more directly probe a student's degree of understanding of specific content (e.g., correct sequential errors in expository or narrative text). Each innovative item used in the current study was developed to specifically align to one or more CCSS.

A total of 20 mathematics items were used in the mathematics cognitive labs. Fifteen of these items were innovative with the remainder serving as "matched" traditional multiple-choice items that covered the same content standard(s) as one of the innovative items. These traditional items (two elementary school, one middle school, and two high school) were administered just before the corresponding innovative items were administered to support comparisons of the cognitive processes used to answer the innovative items. Participants were also asked to directly compare the traditional and innovative items in the retrospective portion of the cognitive lab. This provided direct feedback on the perceived differences between the items' format in terms of overall difficulty, format preference, usability, and general level of comfort with the item formats.

A total of 16 items were used in the ELA cognitive labs. Unlike the mathematics cognitive labs, the ELA cognitive lab sessions included only innovative items. Although this

limited the ability to more directly compare cognitive processing between traditional and innovative items, it allowed adequate time for students to complete the innovative ELA items, which took longer than the mathematics items due to increased amounts of reading, listening, and/or writing. Nonetheless, when time allowed, participants were asked to reflect on their experience with traditional multiple-choice ELA items as part of the retrospective interview.

Tables 1a and 1b provide an overview of the mathematics and ELA items used in the study along with a general description of the content covered in the items and the type of item for each grade band and content area.

Table 1a: Items included in mathematics cognitive labs, along with the CCSS standard or standards to which they align. Within each grade band, items are listed in the order they were administered to students. Traditional items are shown in italics.

Content Assessed in Item	Aligned CCSS Domain(s)	Grade	Standard
Elementary School			
Place value [traditional item]	*Number-Base Ten*	*2*	*1*
Place value	Number-Base Ten	2	1
Measurement (length)	Measurement and Data	4	2
Measurement (money) [traditional item]	*Measurement and Data*	*4*	*2*
Measurement (money)	Measurement and Data	4	2
Middle School			
Construct nets using rectangles and triangles	Geometry	6	4
Use a number line (negative numbers)	Expressions and Equations The Number System	6 6	8 6
Use a number line (real-world application)	Expressions and Equations The Number System	6 6	8 6
Dilations on the coordinate plane	Geometry Geometry	7 8	1 4
Unit rates/Number line diagrams (distance-rate-time)	Proportional Relationships The Number System	6 6	3 6
Scatter plots [traditional item]	*Statistics and Probability*	*8*	*1*
Scatter plots	Statistics and Probability	8	1
High School			
Simplify an algebraic expression [traditional item]	*Algebra*	*9-12*	*1*
Simplify an algebraic expression	Algebra	9-12	1
Geometric proof (quadrilaterals)	Geometry	9-12	11
Geometric construction (perpendicular bisector)	Geometry	9-12	12
Graph on the coordinate plane (exponential function)	Algebra Functions	9-12 9-12	2 1
Geometric construction (equilateral triangle)	Geometry	9-12	13
Write an equation (quadratic function) [traditional item]	*Algebra* *Functions*	*9-12* *9-12*	*10* *7*
Write an equation (quadratic function)	Algebra Functions	9-12 9-12	10 7

Table 1b: Items included in ELA cognitive labs, along with the CCSS standard or standards to which they align. Within each grade band, items are listed in the order they were administered to students.

Content Assessed in Item	Aligned CCSS Domain(s)	Grade	Standard
Elementary School			
Determine main idea and supporting details	Reading: Informational Texts	5	2
Revise for logical consistency	Writing Writing	5 5	1a 1b
Oral narrative with accompanying animations	Speaking and Listening Writing	5 5	5 3b
Middle School			
Identify and sequence important actions in plot	Reading: Literature	6	3
Identify conflict and resolution in plot	Speaking and Listening Reading: Literature	6 6	4 3
Interpret information presented orally	Speaking and Listening	7	2
Evaluate argument presented orally	Speaking and Listening Speaking and Listening	7 7	3 2
Evaluate factual accuracy in text	Reading: Informational Texts Writing	8 8	9 5
High School			
Evaluate validity of evidence in an argument	Reading: Informational Texts	9-10	8
Determine meaning of multiple meaning words in context	Career and College Readiness Career and College Readiness	ALL	9
Analyze language to infer meaning	Reading: Literature Reading: Literature	9-10 9-10	4 1
Evaluate speaker's point of view and validity of evidence	Speaking and Listening Reading: Informational Texts	9-10 9-10	3 7
Revise for organization and clarity	Writing	9-10	3c
Demonstrate command of English language conventions	Language Language	9-10 9-10	1 2
Revise and edit for accuracy of English language conventions	Writing Language	9-10 9-10	2 1
Create a balanced report with valid evidence	Speaking and Listening Speaking and Listening	9-10 9-10	4 5

Participants

To create the sample of 36 students (two content areas by three grade bands by six students each) for this study, the school-age children of Pearson employees were asked to volunteer for participation. (Pearson has a large and diverse workforce.) The extremely time-intensive nature of data collection and analysis in cognitive lab methods prohibit the use of large samples. The relatively small sample size in this study is typical of those in other research that use similar methods (see Ercikan et al, 2010; Gierl, Wang, & Zhou, 2008). A total of 457 potential participants responded to the call for volunteers and each provided basic demographic information (gender, race/ethnicity, last grade completed, geographic location, computer skills). Before proceeding with the selection for participation, each student was assigned a unique number and identifying information was removed to promote student anonymity[4].To increase the probability that students had been given the opportunity to learn the content covered on the tests, only students who had completed the 4th or 5th grade were considered for the elementary school sample, 7th or 8th grade for the middle school sample, and 10th or 11t^{h} grade for the high school sample. To ensure a diverse sample of students was represented in the study, a stratified sampling method was employed to select a sample that came as close as possible to having half female and half male participants, at least two non-Caucasian students, and equal number of students with medium and high computer skills[5]. The elementary and middle school respondents were large and diverse, and yielded a sample that was well balanced across the demographic

[4] Student names, addresses, and contact information were retained to perform administrative tasks such as scheduling the think-aloud session and mailing materials for the study.

[5] The computer skills variable was self-reported or parent-reported as high, medium, or low. No students with low computer skills were included as we wanted students who had a reasonable chance of interacting well with the interfaces for this initial study. (Note that very few students were reported as having low skills.). The researchers acknowledge that future studies should systematically include students with low computer skills as the enhanced technology used in the innovative items would likely differentially impact their performance.

combinations and deviated only slightly from the desired targets[6]. The high school grade band was more difficult as there were few 11th graders, very few non-Caucasian students, and very few students reporting that they had "medium" computer literacy. Overall, the marginal proportions for all content areas by grade level samples were satisfactory in terms of demographic representation. Table 2 displays the demographic breakdown of the overall sample.

[6] Deviations were due to the inability to schedule cognitive labs for some students who met the needed criteria during the time frame of this research project.

Table 2: Sampled Participants for Cognitive Labs

Content Area	Grade Band	Ethnic/ Racial Minority	Gender	Last Grade Completed	Reported Computer Skills
Mathematics	Elementary	No	Male	5	Medium
		Yes	Female	5	Medium
		No	Male	4	High
		No	Female	5	Medium
		No	Female	4	High
		Yes	Male	5	Medium
	Middle	Yes	Male	7	High
		No	Male	7	Medium
		No	Female	8	Medium
		Yes	Female	8	Medium
		No	Female	8	High
		No	Male	7	High
	High	Yes	Female	10	High
		Yes	Male	10	High
		Yes	Female	10	Medium
		No	Female	10	Medium
		No	Female	10	High
		No	Male	11	High
ELA	Elementary	No	Male	5	High
		No	Female	5	High
		Yes	Male	5	High
		No	Male	4	Medium
		Yes	Male	5	Medium
		No	Female	4	High
	Middle	No	Female	7	Medium
		No	Male	8	High
		Yes	Male	8	Medium
		No	Female	8	High
		Yes	Female	7	High
		No	Male	7	Medium
	High	No	Female	10	High
		No	Male	10	Medium
		No	Male	10	High
		No	Female	11	Medium
		Yes	Female	10	High
		No	Female	11	Medium

Cognitive Lab Procedures

The cognitive labs were conducted online. Two members of the research team acted as session moderators and initiated a simultaneous web conference and telephone conference call at a prescheduled time in which the student could speak with the moderator and view and control what was on the moderator's computer screen (via Cisco® WebEx®). The moderator first gave a brief introduction to the purposes of the project and assured the student that the session was on the test items and how they functioned and not the correctness or quality of their answers. Each student was also given an overview of the cognitive lab method. After the student confirmed that he or she understood what was expected, the moderator provided a brief overview of the test environment and then turned control of his or her computer over to the student participant. The two-step cognitive lab process, as described earlier, was used for the session. The student was asked to complete each item, thinking aloud as he or she worked through the task. The moderator used neutral prompts in cases when the student stopped verbalizing, and in cases when the student asked questions, the session moderator responded by asking the student "What do you think you should do?" or "What would you do if this were a real test?" When participants reported that they had completed an item to their satisfaction, a retrospective interview specific to that item took place while the student could view and interact with the item. The moderator asked specific questions related to the student's experiences or interactions with specific features of the item (e.g., if an aspect of an item was challenging in terms of content or usability, the student was encouraged to elaborate on the obstacles). In addition, each student was asked a series of questions related to the focal points of the study. These questions included:

- Did you know how to answer the question? Were you clear what you were being asked?
- What features of the item made it easy to use or difficult to use?

- Have you ever seen an item like this?
- Did you like working on this item? Why or why not? What would make this item easier to use?

If the item was an innovative item, the questions also included:

- How does this item compare to items that you typically see on a test?
- Which item would you rather answer—this one or a multiple-choice item? Why?

After the retrospective interview, a participant was instructed to proceed to the next question.

Sessions were scheduled for, and generally completed within, one hour. In some cases, students were not able to complete all the items. The audio and video were recorded for each cognitive lab and retained for offline analysis. Due to technical limitations, video sessions recorded with WebEx® (approximately half) did not include the students' mouse movements[7].

Data Analysis

The cognitive lab sessions were viewed by one or two observers. Each observer took detailed notes of the participant's interactions with the items. To assist in focusing the analysis on the research questions, two different types of coding templates were applied for each grade band/content area, one focusing on content and the other on interface.

The first coding template was specific to the content and the cognitive processes used to respond to the content in each question. Prior to the study, an assessment content specialist identified the strategies—correct and incorrect—that a student could employ in responding to the

[7] This likely had little impact on study findings as the focus of this study was equally divided on the path students took to produce a response (which did not require mouse movement tracking) and an overall check of usability of item features. In the latter, mouse movement may have supplemented some of the observations, such as for those students who happened to use the mouse to track their own reading. Ultimately, however, we focused on the end result of "did they interact?" or "how much trouble, if any, did they have interacting?" These observations were accomplished without knowing where the mouse was by using observable action of the item features, student utterances, and their final responses.

item. In addition, one or more strategies were identified that could lead to a correct response, incorrect steps, common mistakes, and common misconceptions that might prevent a correct solution. During the cognitive lab, the observers recorded which steps were taken and the order these steps took place. In cases where the student employed a strategy that did not appear on the content coding sheet, the coder took specific notes describing the actions that were taken and when they took place.

The second coding template was specific to the usability of the item features and general feedback from students. The different user interface elements (e.g., drag-and-drop tool, video snapshot capturing tool, compass tool) were listed for each item, and the coders indicated the degree to which a student could interact with the elements as intended, detailing each student-element interaction when appropriate. Certain features of the user interface were identified as essential to the use of the items, while other features merely added convenience, provided additional user feedback, or enabled multiple ways to achieve the same results. The usability coding template also provided coders an area to record a student's responses to the retrospective interview questions—specifically focusing on aspects of the items that were particularly successful and features of items that were confusing or difficult. The coding template provided an area to record observations and/or comments about a student's familiarity or specific experience with the content contained in an item.

When all observations were completed, the observations were coded by a single researcher. Using the notes, the videos of the cognitive lab sessions, and the templates, the coder identified instances of interest in each session. These instances related to either the preestablished research questions of the study or reflected trends and patterns that emerged from

the set of observations. The instances were classified into themes, which in turn helped to identify patterns across and within the various subsets of cognitive lab sessions.

Results

The results of this study are presented here in six parts, each of which describe the usability and content-related observations for each of the six items sets (each grade level fully crossed with the two content areas) separately. Where appropriate, item level descriptions and observations are presented; otherwise, observations are presented for the item sets in general.

Elementary School Mathematics Items

The elementary school students had little trouble with the user interface or the functionality of the three innovative mathematics items included in the cognitive lab. A description of these items is provided in Table 3. Table 4 summarizes general observations about student interactions with each item. All six students determined how to use the capabilities included in the three innovative items. In most cases, the students accomplished this independently and immediately. The measurement (money) item was the most challenging item due to the art associated with the item, the overall design, and certain aspects of the item functionality (as explained in more detail below). Out of the total 18 observations made of elementary students completing innovative items, three students spent more than 10 minutes exploring the functionality of the innovative items (two spent extra time on the measurement [money] item, and one student spent extra time on the place-value and measurement [length] item) before being able to complete the problem. Often these students had to reread the directions to understand how to interact with the items in the intended manner.

Students typically reported that the basic mathematics concepts included in the items were topics that they recognized from prior mathematics coursework. The students consistently

followed one or more of the projected pathways suggested in the templates to produce correct answers. There was no evidence that correctness of a student's responses was related to anything other than the possession (or lack thereof) the requisite content knowledge. However, for some items, issues with the user interface may have increased the associated cognitive load. For instance, the user interface in the measurement (money) item undid any drag-and-drop operation in which a student placed money atop other money. This "fail-safe" mechanism designed to avoid coins hidden beneath other coins or bills confused some students. While all students did complete the item correctly, some did not have the intended answer after their initial attempt, resulting in them re-creating and verifying their answer multiple times.

Table 3: Description of Items in the Elementary Mathematics Problem Set

Question Stem	User Interface Instructions	Stimulus	Student Response
Place-value item			
Use base-10 blocks to model the value of digit 6 in the number 468. Each unit block represents a value of 1.	Adding blocks: Use your mouse to drag-and-drop a block into a column. Removing blocks: Use your mouse to drag-and-drop a block back to the top.	Blocks representing units, tens, hundreds are lined up in three columns.	Drag-and-drop blocks to specific place value location.
Measurement (length) item			
A science class planted seeds and then measured how fast the seedlings grew. Show the seedling at the height described below the box.	Use the slider below to view one seedling's growth. Use the camera to take snapshots of the seedling. Place a snapshot in each of the three boxes to show the seedling at the height described below the box.	Animation of a seedling growing to a certain height is shown near an upright ruler.	Capture a snapshot of the animation with camera and drag-and-drop snapshots to one of three boxes corresponding to the heights required.
Measurement (money) item			
Look at the two pictures below and answer the question that follows. What is the total cost of Armin's two orders?	Drag-and-drop bills and coins over to the tray to represent the cost of Armin's two orders.	Two images are shown: In one image Armin orders two enchiladas, and in the other image Armin orders one burrito.	Drag-and-drop coins and bills to tray.

Table 4: Summary of Item Interactions in the Elementary Mathematics Problem Set

Successful Outcomes	Challenges	Preliminary Conclusions
Place-value item		
Intuitive interface. Engaging and easy to use.	Students represented entire number. It is unclear if they misunderstood the task or were unclear on the mathematical concept	Familiar drag-and-drop interaction successfully applied to base-10 blocks. Item type easily applicable to other items.
Measurement (length) item		
Snapshot interface with camera was intuitive and easy to use.	Students had a difficult time capturing the exact height of the seedling. In some cases it was unclear if this was a content limitation or if the animation used for this item could be improved for clearer heights to be captured.	Snapshot interface has enormous potential to be used with various constructs, particularly when the context involves time-based processes. Tutorial material should address using the slider and the single frame arrows for greater precision than can be achieved with the Play button.
Measurement (money) item		
The tray metaphor worked along with the idea of dragging money over.	The screen was densely populated. Some students found the automatic dragger adjustment to avoid bills or coins covering up other coins to be unexpected. A question was raised about whether students might expect an unlimited number of units in each denomination rather than expecting that the payee would have a limited amount of available cash in his wallet to cover the bill.	This particular drag-and-drop interface (i.e., free-form drag-and-drop) needs to be re-conceived to address both usability and the danger of draggers blocking the view of other draggers. Other minor design issues could be addressed to improve this item but do not necessarily point to deficiencies in this general item type.

Middle School Mathematics Items

The middle school students had few challenges with the user interface or navigating the functionality of the five[8] mathematics items included in the cognitive labs. Students used interface elements (e.g., drag-and-drop, playing videos) intuitively, including those that required significant exploration (e.g., number line, dilation, scatter plot). The content of the middle school items appeared to be appropriate for the students in term of students' prior opportunity to learn. The strategies—both correct and incorrect—employed by students in producing their responses matched those identified in advance by the content experts. Table 5 gives a general description of the content and functionality of each item included in the middle school mathematics cognitive labs. The middle school items as a set are far more specific in content than the elementary items, and hence required more specialized and complicated features and functions within each item. Table 6 describes some of the general observations about the performance of each item. Of the five items included, two (items 1 and 4) were administered with no noted usability issues, and thus students were presumably able to produce responses based on their understanding of the item content. The scatter plot item functioned as expected even though the user interface was complex and the item required a fair amount of time to complete. Only the dilation problem, where most students required ample time and sometimes produced unexpected results, was observed as having any usability obstacles. On items with many interactive features, most middle school students gave only a cursory glance at item directions before interacting with the items features. In just over half of the observed student-item interactions, students only read directions as a final troubleshooting step. The interface and functionality also did not prevent any student

[8] A total of six innovative items were administered to the sample of middle school students, but any single participant only viewed five of the six innovative items (see Table 1).

from ultimately producing a correct answer, though the functionality may have differentially affected some students in terms of the time required to answer the question.

Table 5: Descriptions of Items in the Middle School Mathematics Problem Set

Question Stem	User Interface Instructions	Stimulus	Student Response
Construct nets using rectangles and triangles			
Create the net for a square pyramid.	By dragging the shapes on the left to the grid.	A three by three grid is on the right. A square, right triangles, and equilateral triangles are on the left.	Drag-and-drop the appropriate shape into the grid.
Use a number line (real-world application)			
Graph the set of numbers that is less than –3.5.	None	A number line from –10 to +10, set choices, and ray choices are shown.	Click to select a set or a ray. Move points to stretch or shrink the chosen set or ray.
Dilations on a coordinate plane			
Quadrilateral STUV *is similar to quadrilateral* ABCD. *The segment* AB *is provided. Complete the drawing of the quadrilateral* ABCD *on the grid and label its vertices.*	No specific instructions. Reset and Close the Shape buttons provided.	A coordinate grid containing the graphed quadrilateral *STUV* and line segment *AB* are presented.	Plot point on the grid. A line is automatically created between two plotted points.

Table 5 (continued): Descriptions of Items in the Middle School Mathematics Problem Set

Question Stem	User Interface Instructions	Stimulus	Student Response
Unit rates/Number line diagrams (distance-rate-time)			
Billy and Sally both went for a bicycle ride at noon. Watch the videos and show the distance each child rode by dragging the bicycle below.	Drag-and-drop each bicycle to show how many miles Billy and Sally each rode.	Two rulers are shown, one below the other. A bicycle is at the left end of each ruler. A triangle points to the location on the ruler based on where the bicycle is positioned.	Drag-and-drop the bicycle on the ruler. A blue line appears on the ruler to show the distance that Billy rode, and a red line appears on the ruler to show the distance that Sally rode.
Scatter plots			
Zach surveyed a group of people visiting state parks to determine if the distance they lived from the park affected how frequently they visited the park. Create a scatter plot that correctly represents Zach's data.	Choose a title on the graph for both axes. Select an appropriate scale for each axis. Pick the points.	A graph with drop-down options for the title and labels for *x* and *y* axes are shown. Blank space is included to specify the scale and +/– signs to increase or decrease graph lines.	Click on a graph line to plot each point. . Use drop-down options for titles and axes labels. Expand or reduce graph lines. Insert number to set appropriate scale for *x* and *y* axes.

Table 6: Summary of Item Interactions in the Middle School Mathematics Problem Set

Successful Outcomes	Challenges	Preliminary Conclusions
Construct nets using rectangles and triangles		
The interface was easy to use. It is a well-designed template for other types of nets.	None	The user interface has potential to be used for similar types of mathematics problems.
Use a number line (real-world application)		
Those who knew the concept of the number line were able to solve the problem.	The interface design was not as intuitive as seen in other items. Stretching, shrinking, and removing the points and lines could be easier.	Add specific instruction on how to remove the points and lines. Perhaps include visual feedback with highlights as the points are expanded or shrunk.
Dilations on a coordinate plane		
None	Some experimentation was necessary for a student to understand how to create and close a shape. The meaning and function of the "Close the shape" button was not immediately apparent to all students. Some students did not anticipate that consecutive points would be connected.	Better user interface instructions should be provided to instruct the student how to edit the points, how points get connected (i.e., in the order in which they are added), and what the "Close the shape" button does. More investigation on whether a Reset button is the best route or if supporting greater editability of any added point would suffice.
Unit rates/Number line diagrams (distance-rate-time)		
The interface was very intuitive and easy to use.	None	Students found the interface easy to use.
Scatter plots		
The interface shows great promise. The ability to expand and reduce graph lines and set the scale of the *x* and *y* axes are elegant.	Many steps were needed to choose a title and add points. This item was time intensive like creating an actual graph. Students took far less time to choose a multiple-choice option.	Prior exposure to the interface will most likely reduce the time investment involved with first-time exposure to the interactivity. The item has tremendous potential, especially when compared with its traditional response counterpart.

High School Mathematics Items

The innovative mathematics items in the high school cognitive labs included some concepts that required even more specialized and unfamiliar formats or functions than the middle school items. Tables 7 and 8 give a general description of the content of the items and some general observations about the performance of the items during the cognitive labs. In particular, the geometric proof item, the construction of a perpendicular bisector, and the quadratic function item were administered with no usability issues and successfully captured student responses at different levels of mastery (i.e., while students had varying degrees of content mastery, all still successfully generated a response). Only two students, however, recorded any response for the algebraic expression item. In general, the students who failed to respond spent considerable amounts of time trying to work through the problem, but they made no attempt to record their incorrect answers. Students were not probed adequately enough to determine whether their failure to respond was due to a lack of content mastery (e.g., students were aware their answer was incorrect), usability problems (e.g., students could not determine how to record their answer), or some combination of these factors.

Across the set of high school mathematics items, partial or incomplete answers were frequently submitted as a final response. In these cases, one or more of the tasks needed for a complete answer was omitted. For example, a few students selected an equation type in the write-an-equation (quadratic) item without filling in values. Likewise, they selected a graph type in the graph-on-the-coordinate-plane (exponential function) item, but did not adjust the points to correctly position the graph. Students tended to use appropriate strategies in responding to the items when they had adequate content knowledge and skills. Additionally, most of the high school students showed evidence of advanced computer skills (e.g., easily and intuitively

navigating all but the most complex of item features) but could not produce a correct response to some of the items aligned to advanced mathematical concepts (e.g., geometric construction). Responding to the innovative items, especially those with multiple tasks to complete, did typically require considerable amounts of time (in many case more than 10 minutes was spent on a single response).

Table 7: Descriptions of Items in the High School Mathematics Problem Set

Question Stem	User Interface Instructions	Stimulus	Student Response
Simplify an algebraic expression			
Simplify the expression below completely. Use the area at the bottom of the screen to represent your answer as a polynomial.	Build a polynomial by dragging the appropriate pieces to the area below. Then complete the polynomial by typing the correct coefficients and the exponents into the blanks.	An expression, followed by a collection of boxes with coefficients, exponents, and operators are shown. Blank boxes are shown below.	Drag-and-drop the coefficients, exponents, and operators to the blank boxes.
Geometric proof (quadrilateral)			
Construct a proof for the theorem – If polygon ABCD *is a square, then* ABCD *is a quadrilateral*	Arrange the statements and reasons in the table below.	Statement and Reason columns are shown with only the first Statement cell filled in. All other cells are blank. Below, a set of statements and reasons are presented in their columns.	Drag-and-drop the statement and reasons to the blank cells.
Geometric construction (bisector)			
Watch the animations below and choose the two that demonstrate proper technique for creating a perpendicular bisector for line MN.	Drag-and-drop the animations to the two slots on the filmstrip. Use the Play All button to observe the result.	Eight animations, with four in each row are shown. The top row presents choices for step 1 and bottom row of animations presents choices for step 2.	Drag-and-drop animations to represent step 1 and step 2 for constructing a perpendicular bisector

Table 7 (continued): Descriptions of Items in the High School Mathematics Problem Set

Question Stem	User Interface Instructions	Stimulus	Student Response
Graph on a coordinate plane (exponential function)			
A bacteria is cultured and its rate of growth observed. The observation begins with three bacteria cells. The number of cells doubles every hour. Graph the function that models this bacteria growth where x *is the number of hours and* y *is the number of bacteria cells.*	1. Click a button to choose the graph type. 2. Drag-and-drop the two points to the correct position.	An animation of growing bacteria cells, including a microscopic view is shown. Students are provided four graph types to choose from and an empty graph with *x* and *y* coordinates.	Click on a graph type. Drag-and-drop points on the graph.
Geometric construction (equilateral triangle)			
Use compass and line tool to create an equilateral triangle using line AB.	Detailed instructions are provided about using the compass and line tools.	A construction space with line *AB* and tools on top that include compass, line tool, and eraser are shown.	Draw with the compass and line tool.
Write an equation (quadratic)			
Write an equation for this graph.	1. Select an equation. 2. Fill in the correct values.	Five equation choices are shown.	Click on an equation type. Fill in coefficients and exponents.

Table 8: Summary of Item Interactions in the High School Mathematics Problem Set

Successful Outcomes	Challenges	Preliminary Conclusions
Simplify an algebraic expression		
None observed, though this may be due to content limitations.	The interface may have initially looked complicated to students and required time investment to construct the polynomial.	More research is needed to compare building a polynomial using this method (drag-and-drop combined with typing numbers) with other methods, such as using an equation editor and writing by hand. Compare time on task, availability of automated scoring methods, and interface challenges.
Geometric proof (quadrilaterals)		
The task was easy and the interface was intuitive.	None	Drag-and-drop construction can be applied to geometric proofs for automated scoring. Other variations might involve other types of automated scoring (latent semantic analysis), more statements/reasons, or limiting where statements and reasons can be placed.
Geometric construction (perpendicular bisector)		
Easy to use.	None	For solving multistep mathematics problems in this way, further exploration could be dedicated to how/if steps are labeled. How does increasing or decreasing the scaffolding in this way affect item performance? This interface could be used for different kinds of mathematics problems.
Graph on a coordinate plane (exponential function)		
Task and interface instructions were clear. Choosing from the graph types made the task somewhat easier. The graph interface was simple and easy to use.	Not all students spent time with animation—did some student not consider it important to the task?	Prior exposure to the interface would be expected to reduce time on task. Research with more motivated students to see if there is greater engagement with all aspects of the item.

Table 8 (continued): Summary of Item Interactions in the High School Mathematics Problem Set

Successful Outcomes	Challenges	Preliminary Conclusions
Geometric construction (equilateral triangle)		
This item type has the potential to mimic real-world geometric construction.	Students did not read instructions for using the compass. Greater referencing of the eraser feature in the instructions may or may not have led to greater usage. Overall, it was difficult to translate this physical task to an online environment without prior exposure to the compass tool.	Prior exposure to the interface is critical. While the instructions were clear, the learning curve to master the compass tool was steep. While the computer-based version allows for automated scoring, students will be more comfortable with the physical compass unless such digital tools become more widely used.
Write an equation (quadratic function)		
Well-designed interface.	None	Interface has the potential to be used for similar types of mathematics problems.

Elementary School ELA Items

Elementary students participating in the ELA cognitive labs were able to competently navigate the user interface and the item features with only a few instances of difficulty. Tables 9 and 10 provide a general description of the content and format of each item included in the elementary sessions as well as a summary of the student interactions with the items. In several cases, a specific item feature caused some trouble initially for a subset of examinees (e.g., the paragraph reorder feature, finding and using the audio recording feature on the storytelling item), but ultimately students figured out the interface with little or no external assistance. When initially dealing with the interface, students explored the item features, trying different actions, until they succeeded. After an initial success with the item interface, all students successfully replicated outcomes on subsequent steps. Although specific directions explained user interface and/or the novel features included in each item, these instructions were mostly ignored or read

incompletely at the onset of an item. In cases where the directions were read initially or reread after an unsuccessful attempt on an item, students were always able to quickly and efficiently interact with the user interface elements contained in an item. Despite any initial struggles with understanding the task or how to use the interface, the six elementary students were able to produce responses for all three items within the hour allotted for the cognitive lab session. In no case did issues of usability prohibit the production of an answer. The items successfully elicited answers of different quality and correctness across the sample of six students. In most cases, students' strategies matched those identified in advance by content experts; in only a few cases did students employ an unexpected incorrect strategy.

Table 9: Descriptions of Items in the Elementary School ELA Problem Set

Question Stem	User Interface Instructions	Stimulus	Student Response
Determine main idea and supporting details (sentence extractor)			
Dr. Castigo is a dentist whose patients often ask for his recommendations regarding chewing gum. He decides to make an informational poster for his office by pulling two positive research findings and two statements of caution from a research summary. Read the research report on the left and drag-and-drop the most appropriate pieces of information to Dr. Castigo's poster.	Instructions are contained in the question stem.	Research summary is on the left. Poster with blank spaces for statements on the right. Poster states: Is Chewing Gum Good or Bad for your Teeth? If you enjoy gum, here's some good news…___. But don't forget about these cautionary words…___.	Drag and drop sentences from the research summary text area to the blank spaces in the poster.
Revise for logical consistency (paragraph reordering)			
A student has written about a class experiment, but the paragraphs are out of order. Rearrange the paragraphs in a logical order.	To move a paragraph, click it and drag-and-drop it to a new location. The paragraph will turn yellow when it is selected. While the paragraph is being dragged, a blue line will show where it would go if released.	Several paragraphs are shown with a scroll bar to view the entire passage.	Drag-and-drop paragraphs within the passage area. Ability to reset with "Reset paragraph order."
Oral narrative with accompanying animations (storytelling)			
Create a story around four of the images below.	Drag-and-drop the images you have chosen to the filmstrip. Click the area beneath each image on the filmstrip to record narration for that image. Use the Play All button to see the end result.	Nine images are arranged in a grid. A film strip is on the right side with four image areas and a "Click here to record" link is below each. A Play all button is below.	Drag-and-drop images in the film strip. Record the narration using standard audio controls to record, playback, etc.

Table 10: Summary of Item Interactions in the Elementary School ELA Problem Set

Successful Outcomes	Challenges	Preliminary Conclusions
Determine main idea and supporting details (sentence extractor)		
None	Student engagement with instructions was minimal.	Need to better understand student expectations regarding items that may have multiple correct answers, which differs from a traditional large-scale assessment using multiple-choice items. Explore other ways of presenting instructions such as with an animation that precedes the item. Item type could be used for various CCSS that require citing evidence from a text, but further research should explore whether content issues with this particular item made it difficult for this grade band.
Revise for logical consistency (paragraph reordering)		
Clear interface instructions and elegant design. Task was also clear and easy to execute.	Due to the way a spot in between two paragraphs is indicated in comparison to how the space above the first paragraph is indicated, it was more difficult for students to drag a paragraph to the beginning of the passage than to a spot between two paragraphs.	Ability to "Undo last move" may be useful. Minor change to better support dragging paragraph to the beginning of the passage could be made.
Oral narrative with accompanying animations (storytelling)		
Potential for use with different types of ELA activities.	The task in its on-screen form was unfamiliar. Not all students understood that a single story was to be created. Students had not encountered audio recording through this type of interface within an assessment before.	Prior exposure to the interface and/or an animated introduction will increase comfort with audio recording and with sequencing content in this way. Interesting potential for use with CCSS that emphasizes electronic publishing in addition to writing.

Middle School ELA Items

The tasks included in the middle school innovative ELA items were more specific and specialized than the elementary school items, resulting in item features and functionalities that were more involved and a more complex user interface. Despite the increased complexity, middle school students typically encountered no usability obstacles when interacting with the item set. In addition to new and complicated item features, the middle school ELA items required more information to be consumed by the student by reading or listening in order to respond. Most middle school items, especially those that contained text-based stimuli, took far longer to complete than the either the middle school mathematics items or the elementary ELA items. Tables 11 and 12 describe the content and features of each item and some general observations about the successes and challenges associated with the student-item interactions. Overall, students engaged the interface immediately (a “click and go” mentality) after only a cursory glance at the text associated with the task or the interface, and the quality and correctness of the responses students produced for the item set was highly variable. In almost one-third of the student-items interactions that were observed, the students produced only partially complete items before indicating they were ready to move to the next item.

Multiple students expressed that the ELA content in some of the items was unfamiliar to them (e.g., not familiar with “faulty reasoning” or “advancing the plot”), even though these concepts are found in the CCSS ELA standards for the middle school grades. In most cases, coding sheets produced by content experts describing the expected the steps and missteps for each item successfully captured the processes students followed in responding to the set of items, though the exact order and combination of these steps were highly variable across the sample of students.

Table 11: Item Description for Middle School ELA Problem Set

Question Stem	User Interface Instructions	Stimulus	Student Response
Identify and sequence important actions in plot (*Someday*)			
These four animations show objects used in Someday.. *Which three are the most important in advancing the plot of the story? Select the three animations and drag them to the filmstrip area. Use the note card space to explain how each object is important in advancing the plot of the story. Use "Play All" button to view the completed filmstrip.*	Instructions are contained in the question stem.	The passage opens and closes when clicked. Four animations are on the left. A film strip is to the right with "click to edit text" displayed under each strip segment.	Drag-and-drop animation to the film strip. Click on link to open up text area and enter text information.
Identify conflict and resolution in plot (*Lion and Shepherd* snapshot)			
Use the Play button below to view the story of The Lion and the Shepherd. *Use the slider and the camera button to take snapshots. The snapshots should show three parts of the story: Problem, Climax, and Resolution. Drag-and-drop the snapshots to the correct boxes on the right.*	Instructions are contained in the question stem	Animation of *The Lion and The Shepherd* story is shown with subtitles.	Capture a snapshot of the animation with the camera and place a snapshot in each of the three boxes corresponding to the story parts required.
Interpret information presented orally (*Lion and Shepherd* listening)			
After watching The Lion and the Shepherd, *three students describe what they believe to be the moral of the story. Listen to the students and decide who you most agree with. Support your answer with details from the story.*	Use the area on the right to explain your choice.	Animation of *the Lion and the Shepherd* story is shown with subtitles. On the right side, three static image sketches of students with play/pause buttons to hear their opinions is shown.	Type the response into the text area.

Table 11 (continued): Item Description for Middle School ELA Problem Set

Question Stem	User Interface Instructions	Stimulus	Student Response
Evaluate argument presented orally (teen debate)			
Sherri and Rick are speaking about the topic "Should schools monitor student Internet activity?" Which claims made by these speakers show faulty reasoning?	Use the audio controls to listen to the students, and then answer the questions that follow.	On the left, image sketches are shown of Sherri and Rick with play/pause buttons and sliders to hear their opinions. A text area is shown on the right with the question displayed above.	Type the response into the text area.
Evaluate factual accuracy in text (penguins research)			
The entry within a user-edited encyclopedia has four content errors. Use the link to the National Geographic website to research the emperor penguin.	Correct the errors by clicking the Edit buttons and making these small corrections to eliminate the four errors.	Four different information regions are shown with pictures and text. An Edit button is under each.	Click the Edit button, position the cursor, and type text change.

Table 12: Item Interaction for Middle School ELA Problem Set

Successful Outcomes	Challenges	Preliminary Conclusions
Identify and sequence important actions in plot (*Someday*)		
Good layout of elements given the amount and density of information	Not all students immediately noticed the note card for text entry.	Make note card feature more obvious. Provide practice items or animated introduction to build familiarity with this item type.
Identify conflict and resolution in plot (*Lion and Shepherd* snapshot)		
Well-designed interface. Snapshot feature has potential to be used with different types of items.	More than one right answer made the task a little more complex.	Create more items using this item type to determine the effect of interface versus the effect of multiple correct answers. Prior exposure to interface will enable students to see the relative uses of the play button, the slider, and the frame-by-frame moving arrows.
Interpret information presented orally (*Lion and Shepherd* listening)		
Easy and intuitive user interface.	The student responses were not very in-depth or well conceived. (This was a content-relevant issue not an interface or task-comprehension issue.)	Item type was successful. As we move into listening assessments, explore effects of different voice qualities and ideal pitch ranges for students with possible hearing impairments.
Evaluate argument presented orally (teen debate)		
Easy and intuitive user interface.	The student responses were not very in-depth or well conceived. (This was a content-relevant issue not an interface or task-comprehension issue.)	Overall the item was successful.
Evaluate factual accuracy in text (penguins research)		
Good concept. Edit feature was intuitive. Interface has potential to be used for different kinds of research items.	Toggling between item and website could be tedious (though not difficult). The website – National Geographic – was chosen for search features and recognizable and reputable content, yet still had advertisements.	Students used different search techniques to get to the same webpage/information in nearly the same amount of time. Do students need computer experience to recognize that blue underlined text indicates a web link? How do we assess 21st century skills such as Internet research skills without penalizing students with less access to computers?

High School ELA Items

A total of eight innovative ELA items were developed for use in the high school cognitive labs (see Table 1b for descriptions of content). In the administration of the cognitive labs, students were encouraged to work through problems at their own pace and were not encouraged by the moderator to continue on unless they indicated they were finished or wanted to skip an item. The high school ELA items were often multistep tasks that required considerable amounts of reading or listening. As a result, very few students completed more than the first four ELA items in the hour time slot allotted for the cognitive lab. For this reason, the results presented here are limited to only the first four items in the set. Tables 13 and 14 provide an overview of the content and format of each item and a summary of the interactions observed for each item. High school students, like most of the elementary and middle school students in this study, tended to pay little attention to text-based instructions. While it was not unusual for students to return to the direction set or the item stem if they encountered trouble, it was far more common for students to experiment with different approaches to making the item "work" until they were successful. While the "click and go" mentality did tend to add additional time to each task, the interactions with the item features (e.g., dragging-and-dropping objects, flagging or highlighting sentences) rarely were impediments to producing an answer.

The content of the items and the degree to which the tasks aligned with intended standards and skills was assessed using the content coding templates. The expected processes described in the coding template by content experts were observed in practice, particularly in terms of the steps involved in producing a complete and correct answer. In fact, no student produced a complete response for the entire set of items. Incomplete responses were most often produced on the "Parachute" item (see Tables 13 and 14 for a more complete description of this

item), which had a large amount of text-based information that the student had to consume, a passage to read, and multiple tasks that had to be completed (including the production of several written responses).

Table 13: Item Description for High School ELA Problem Set

Question Stem	User Interface Instructions	Stimulus	Student Response
Evaluate validity of evidence in an argument (multitag)			
Read this fictional letter based on the events that occurred in the Republic of Panama in the early 1900s. Find the letter-writer's statements about the significance of the Panama Canal to the United States' status and national identity. Using these statements, identify the best examples of each of these: fact, opinion, and reasoned judgment.	Drag-and-drop the flags below to the bar at the right of the letter to attach a flag to the correct sentence.	Text of fictional letter is on the left. The question stem is on the right with the three flags below.	Drag-and-drop flags to the bar to tag each sentence.
Determine meaning of multiple-meaning words in context (connections)			
Locate the two uses of the word plot *in the excerpt on the left.*	Use the colored blocks to link each use of the word *plot* with the correct definition on the right. Dragging the block highlights a sentence.	A passage is on the left. Several definitions of the word *plot* on the right. Question stem and colored blocks in the middle.	Drag-and-drop two pairs of colored blocks on the dark gray bar to link the word with its correct definition and usage.

Table 13 (continued): Item Description for High School ELA Problem Set

Question Stem	User Interface Instructions	Stimulus	Student Response
Analyze language to infer meaning (Parachute)			
Reread paragraph 3 of Jack London's "An Adventure in the Upper Sea." Choose four animations that accurately represent this description. Place them in sequence on the filmstrip. Click beneath each animation the filmstrip to open a note card. Use these note cards to describe what is happening in the animation, using the story as a reference.	Click the area below each animation to add text. Use the Play All button to view the completed sequence.	The passage opens and closes when clicked. Six animations are shown. A film strip with space for four animations is below. Link to add text and Play All button.	Drag-and-drop animation to film strip. Click on link to enter text into note card.
Evaluate speaker's point-of-view and validity of evidence (graffiti snapshot)			
Use the Play button to view the newscast below. Then use the slider and the camera icon to take snapshots of certain moments in the newscast. Drag-and-drop the appropriate snapshots to the boxes on the right to best illustrate the captions.	Instructions are contained in the question stem.	Animation of the newscast with subtitles is shown.	Capture a snapshot of the animation with the camera and place a snapshot in each of the three boxes to illustrate the captions below each.

Table 14: Item Interaction for High School ELA Problem Set

Successful Outcomes	Challenges	Preliminary Conclusions
Evaluate validity of evidence in an argument (multitag)		
Good concept and elegant design. This has potential to be used with different types of items.	Many students had trouble with flagging sentences initially. Students tried to place tags over text rather than on the docking bar.	Provide a stronger introduction to the notion of a docking bar to the right of the passage where flags are to be placed.
Determine meaning of multiple-meaning words in context (connections)		
This is well designed and has potential to be used with different types of items.	None.	Provide a stronger introduction to the notion of a docking bar to the right of the passage. If flag-based annotations attached to a docking bar are made available as a non-scored helper tool, this will provide greater exposure to this feature for when flag use is scored.
Analyze language to infer meaning (parachute)		
Interactive interface has good visual presentation of information, excellent example of multistep task.	The task took a long time to complete. Many students were frustrated by the need to toggle between text and animations. Most students did not produce a complete answer potentially due to unmotivated context and presence of an observer.	Research should be repeated with a more conventional passage presentation. Normally a student would be exposed to the passage on its own and then provided a series of items. In this abbreviated version, the passage was not presented independently first and some students missed it.
Evaluate speaker's point-of-view and validity of evidence (graffiti snapshot)		
Well-designed interface has a snapshot feature that has potential to be used with different types of items.	Students were unclear about concepts such as “bias.” (This was a content-relevant issue, not an interface or task-comprehension issue.)	The interface worked well. The snapshot feature has potential to be used with different types of items as a multimedia equivalent of pulling out a sentence from a passage as evidence.

Discussion

Student Perceptions of and Reactions to Innovative Items

Overall, student reactions to the innovative items were positive with regard to item format and content. Most students across grades and content areas commented on how much more they enjoyed working with these items. A few of the high school students who completed the ELA items were very reflective on how these new types of items could change the dynamic of the testing environment to which they are accustomed; they reacted enthusiastically to the presentation of the items, often commenting on how engaging or interactive the tasks were. This sentiment parallels the high task engagement of most students observed during the real-time cognitive lab sessions. Exemplar comments from students across all grade levels and both content areas include:

> "In a test, people think, 'oh, tests are boring', but if you add something a little fun like that, you really have to hit the play button and take snapshots to get your answer, then those are, you still show that you know the answer, but it is also fun for the kids who are taking the test."
>
> "I would definitely be able to focus more, and I know my friends would focus more."
>
> "I like that you can interact with it. If it's a piece of paper, like you read it and it's there, it doesn't feel interesting while taking the test. But if you can hear it and interact with it, you feel motivated."

Students in the higher grade bands of both content areas were also highly aware of the amount of time required to complete an innovative item and frequently commented on this. High school mathematics students, especially students with strong mathematics skills, expressed a general concern about how long it took them to complete the innovative items. The high school ELA sample of students, on the other hand, had a much more tempered view of the time required for each task. They more often embraced the quality of the item while noting the increased time

on task as a consequence of the interactive and engaging nature of the task. For example, one high school ELA student noted:

> "I think this is a great question, but it's time consuming and I spent a lot of time trying to make sure I got it right."

Students in all grade levels were also asked to reflect on their experiences on traditional test items as compared to innovative items. In nearly all cases, students freely commented on the interactivity and engagement of the new items types, but expressed their comfort and security while responding to a multiple-choice item. Some also expressed the perceived ease and benefits of selected-response items. Students, especially in the middle and high school grades, clearly stated that having answer options from which to select a correct response was easier than constructing a response:

> "[I'd] probably [prefer] multiple choice. It's easier and I know I'll have a twenty-five percent chance of getting it right. In this one it's hard to just find an answer by guessing."

> "Interactive things make it kind of difficult, but actually help you come up with an answer. You can't just leave it blank. You actually have to try and think."

> "…it would depend on the person. Some people would prefer #5 [traditional multiple-choice item], but like people who are more hands-on, they learn by doing things will like # 6 [innovative counterpart to #5] better."

In addition to the comfort of employing clear guessing or option elimination strategies, most students have experienced assessments through multiple-choice paper-based tests for many years. As a result, students have had ample opportunity to develop a secondary set of test-taking skills targeted at responding to multiple-choice items and might be reluctant to endorse item types where application of these skills do not benefit them.

Measurement Properties of Innovative Items

Overall, the cognitive labs provided some preliminary evidence that these types of innovative items can produce high quality measurements aligned with the intended skills. The high correspondence between expected and observed student response pathways—correct and incorrect—provides some evidence that the items elicit behaviors that are consistent with subject matter experts' expectations, providing preliminary, but promising, evidence that an item aligns with the intended skills. A second consideration when assessing alignment of intended processes is whether the item adequately allows students to approach the task as they would outside of the testing environment. Overall, the items in this study allow examinees to produce answers using a wide array of combinations of the expected steps and missteps. For example, the middle school ELA item in which students evaluate factual accuracy in text (penguins research; see Figure 1 asks students to assess and correct factual inaccuracies in an article by using information available on an external website. Outside of a hyperlink to the external website, the item provides no other guidelines that tell a student how to proceed. This item allowed students to take multiple paths to constructing their response. For example, a student might choose any of these as a valid and correct first step:

- Reading the entire article without making edits or conducting research.
- Correcting general knowledge errors in the article (e.g., indicating that penguins cannot fly) without confirmation from the website.
- Visiting the website and reading the article on emperor penguins prior to reading the article.
- Reading, researching, and editing one portion of the article at a time.

Different student strategies were most noticeable in the research aspect of the question. In all but one case, the students ended up at a given appropriate informational page on the website in approximately the same amount of time, regardless of how they approached the task. Research approaches included:

- Typing “penguins” or “emperor penguins” from the item into the external website search bar
- Misspelling “penguins” or “emperor penguins” in the search bar and then using “suggested search” to find the appropriate page.
- Copy and pasting “penguins” or “emperor penguins” from the item into the external website search bar.
- Using the “Animals” index on the site rather than searching the external webpage.

Figure 1: Penguins item (evaluate factual accuracy in text). Students are instructed to read the article and follow the external link to a website to find and correct factual errors.

Students also took notably different approaches and demonstrated contradictory degrees of mastery when responding to the traditional multiple-choice items versus the corresponding innovative items. This difference is exemplified by comparing how students interacted with the two scatter plot items in the middle school cognitive labs (Figure 2 and Figure 3 show the traditional and innovative scatter plot items, respectively). In the multiple-choice scatter plot item, many students answered correctly by eliminating implausible choices; no student actually constructed a scatter plot using the provided data and compared it with the answer choices. The innovative item, on the other hand, required students to generate an entire scatter plot, a task that requires a greater depth of knowledge and skill than just identifying a correct one; as such, the innovative item task better matches the intent of the CCSS standard to which it aligns[9], as well as more closely resembles what a student will do in the real world. The quality and completeness of the responses were highly variable in the innovative version, presumably better representing students' true KSAs.

[9] "Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association." (CCSS Math 8.SP.1)

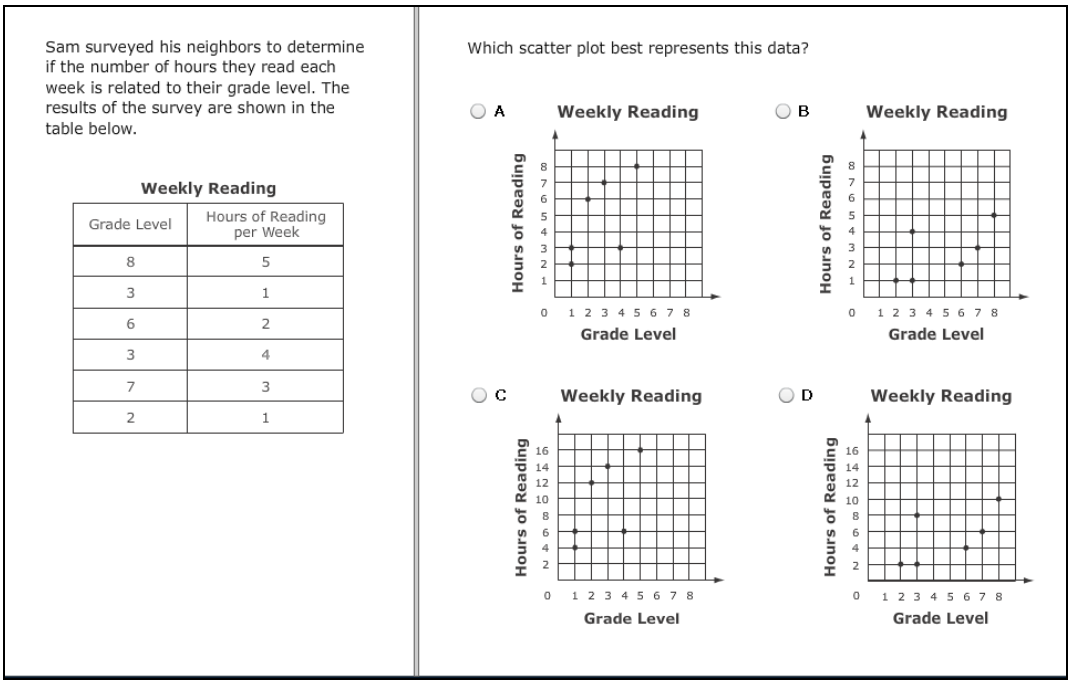


Figure 2: Traditional scatter plot item

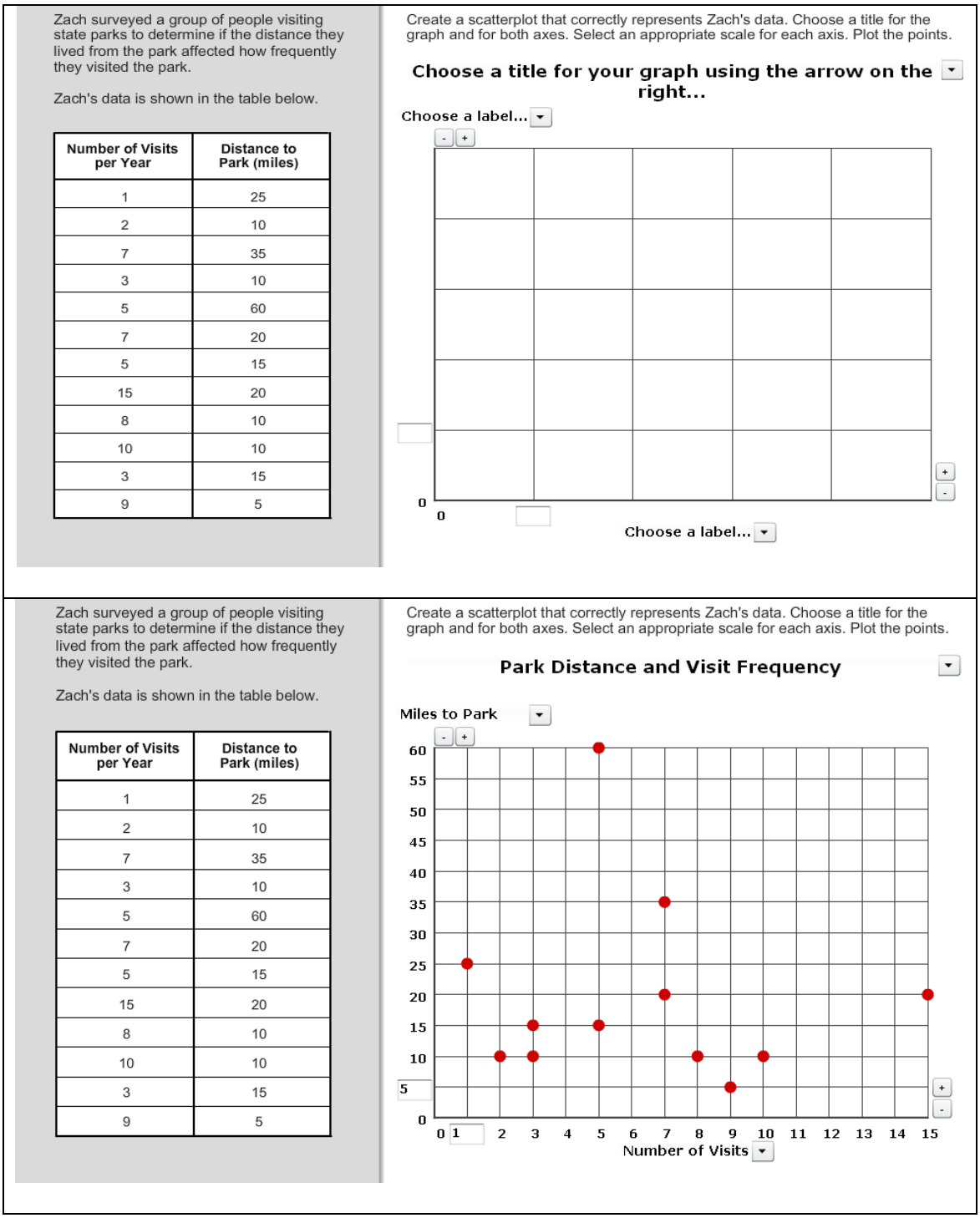
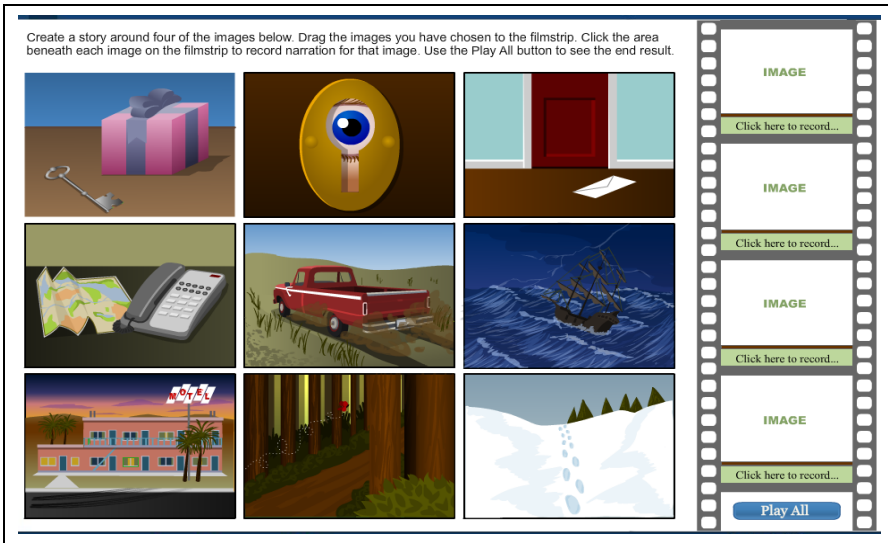


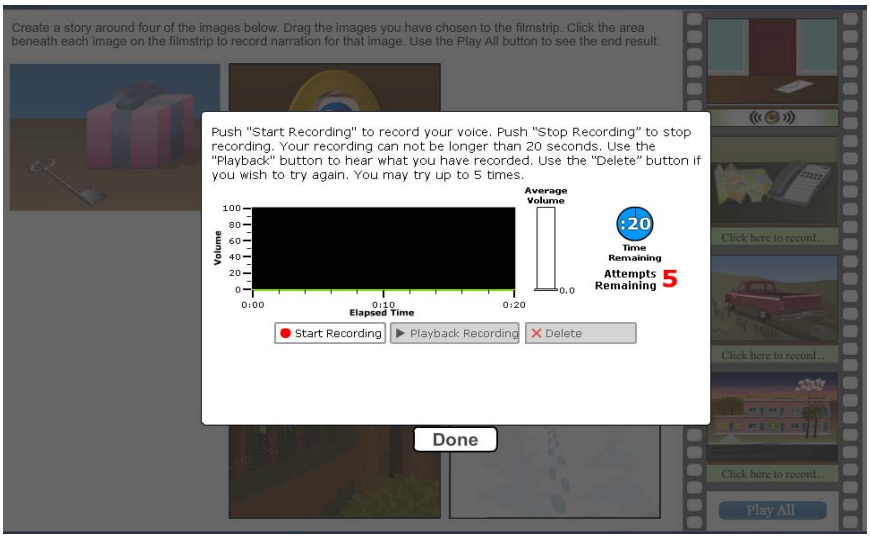
Figure 3: Innovative scatter plot item. The upper image shows the item as first presented to students, while the lower image shows a completed (and correct) response.

In addition to increasing the depth of KSAs that can be measured, study results suggest innovative items can open new and important aspects of the constructs we already measure in both mathematics and ELA across all three grade levels. Results from ELA items at all three grade levels demonstrated that the CCSS speaking and listening standards can be readily assessed in a standardized way. The process skills needed in upper levels of mathematics (e.g., geometric constructions, creating charts and graphs) that traditionally are captured only topically in multiple-choice assessment were directly assessed by items in this study. It is not only the addition of the tools needed to measure new standards that we gain with these item types, but also our ability to create better situations for measurement. Most of the items in this study were designed to integrate multiple skills and required students to complete multiple steps, creating a contextually rich environment that more closely resembles the types of problem a student would encounter in the classroom or the world at large.

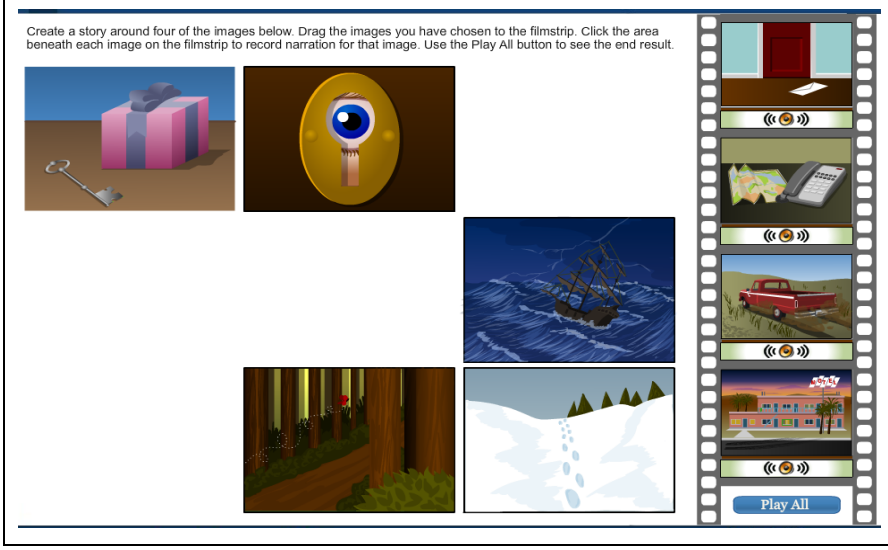
For example, the storytelling item (oral narrative with accompanying animations) asks a student to tell a story by first selecting four out of nine available pictures. Students are instructed to record narration for each picture to communicate that story (see Figure 4). Overall, this task—creating a story—was an activity students were familiar with, but few had completed a task like this on a test or by using audio recordings. This item presents a situation where a student must use multiple sets of skills to create a response and integrates some familiar skills (create a story) with skills that have traditionally been difficult to include. The construction of an equilateral triangle item also showed great promise in its ability to tap into a set of skills that would be difficult to assess in traditional large-scale testing.



a) Students are instructed to select four pictures and use them to tell a story.



b) Students are instructed to record narration for each of the images.



c) The final story (all four pictures, each with audio narration) can be reviewed by clicking the play all button below the filmstrip.

Figure 4: Storytelling item (oral narrative with accompanying animations).

Usability Issues and Construct-Irrelevant Variance

The introduction of new formats and methods during testing, especially those that may rely on additional sets of skills, increases the potential for unintended constructs to influence test performance and hence contaminate measurement. In the case of innovative items, a primary concern is the impact of the user interface and item functionality on a student's ability to understand the task and produce an answer. There were a small number of items spread across the grade levels and content areas where features of the items or the interface did present an obstacle to successful or easy completion of an item even when student had the requisite KSAs (see Tables 4, 6 and 8 for specific examples). While interface modifications to these items have been suggested and may be likely to have a positive impact, such a process must be iterative and include follow-up research to assure that usability issues have been adequately addressed. In the case that usability challenges remain even after repeated modifications, it is possible that some item types or interactive tasks may be determined to not be viable candidates for continued development.

The innovative items in this study did often require large amounts of time for students to complete. While much of the increased time on task was spent on response-related tasks, students' lack of familiarity with the functionality of some items' features and their desire to explore the items' interactive features may have also contributed to the amount of time students spent on the item. In general, students demonstrating strong computing skills quite readily navigated new features. Other students engaged in extended exploration and testing of the item features before producing their response. Given this, it is likely that increased computer skill may reduce the overall time spent on task when students are introduced to new item types and unfamiliar interactivity. This may result in a test that is differentially speeded or fatiguing due to

factors other than the constructs of interest. Further exploration of whether a speed/fatigue factor is introduced by unfamiliar innovative item types needs to be fully explored outside the laboratory setting (i.e., in an environment that more closely resembles an actual testing environment). In any case, these cognitive labs sessions suggest that with appropriate preparation, students will generally navigate through the innovative items more quickly than was observed in this research.

Though unrelated to the technological skills associated with the innovative items, a second trend was also observed in mathematics and ELA at all grade levels. As mentioned above, most students had little trouble with the interface and generally were relaxed and comfortable with the computer interactions required by the items. However, they often failed to engage effectively with the text, including text describing the required tasks or actions, presented in the items. Observations indicate that the students who read the directions initially or who returned to the directions at some point in the response process were able to complete the task as intended. If the failure to read required information is systemically associated with interactive items, a factor other than content knowledge could impact a student's score. Before determining the impact of this failure to read directions, additional research in a context where students are motivated to answer completely and correctly should be conducted.

Summary and Conclusions

The cognitive labs provided initial evidence regarding the design and administration of innovative items developed to align with targeted CCSS constructs not easily measured using traditional items. This evidence included aspects of student interactions with the items, feedback regarding usability issues associated with new functionalities, and information about students'

reactions to the items. This section explores some general patterns and themes across both mathematics and ELA, as well as presents the implications for future research.

Measuring the Intended Constructs

The primary goal of this study was to provide initial evidence that innovative items could be developed to measure specific CCSS constructs that are difficult to assess using traditional items. This is, at its core, a question of validity at the item level. This study began with content specialists and item developers creating items tied to specific CCSS and developing detailed, presupposed cognitive schema for each item (as captured in the content coding templates). Overall, the proposed schema successfully provided blueprints for the multiple strategies students could employ to arrive at a correct, partially correct, or incorrect response. The ability to predict the steps in producing correct responses and types of missteps students tend to make is an important first step in demonstrating the alignment between items and standards. In keeping with much of the contemporary thought on the technical quality of items and/or test instruments (e.g., Messick, 1989; Benson, 1998; Kane, 2006), validation is a process of collecting evidence that justifies the inferential links of test scores (and the intended uses of these scores) with the intended domain of knowledge and skills (i.e., constructs). In this sense, the results of this study provide an initial piece of validity evidence by showing that the cognitive pathways (leading to both correct and incorrect answers) the students engaged in while responding to given items corresponded to the pathways predicted by content experts. However, inferences made based on these studies cannot directly link the standard to the item using direct and quantifiable methods because the relationship between standard and item is based solely on the expert opinion of the content specialist. Thus, this is at best only a start to the process of establishing links between items, scores, and the construct. As with all validation efforts, further studies must confirm the

finding of this study to extrapolate the links between item and skill. Traditional methods using large sample sizes (e.g., confirmatory factor analytic methods, multitrait/multimethod techniques, studies with criterion variables) should be conducted to corroborate and extend these initial findings.

Taking this inferential process one step further, a goal is to not just validate individual items but item types. Each sample item included in this study was intended to evince an item type that could be used to create multiple items in the same way that a drag-and-drop interaction could be deployed in a variety of items. Further studies would be involved to analyze the extent to which validity and usability results for an item type could be applied to new item instances using the same functionality to assess a similar construct. Knowledge around best uses of a given item type and expected psychometric properties would aggregate over time and would not replace item field testing but would promote higher confidence around the use of certain item types, particularly when some innovative items can be more costly to produce than multiple-choice items.

Usability

A major goal of this study was to explore the hypothesis that innovative items capture differences in student performances that relate to their knowledge and understanding of item content (i.e., construct-relevant variance), and that the features and functionalities of these items do not negatively impact the ability of participants to produce acceptable[10] responses (i.e., do not introduce construct-irrelevant variance). However, the results of the cognitive labs provided some evidence that complex and unfamiliar functionality and/or tools impacted the amount of

[10] Acceptable in this context only implies that examinees were able to produce a response using the item interface that could be used to produce a score. For example, a student with a limited understanding of the item may produce a product that is incomplete or simply incorrect—both of these would reflect an acceptable response.

time students needed to respond to the innovative items. Students possessing construct-relevant KSAs and who were either familiar with the item format or computer savvy tended to move more quickly through an item and with less fatigue. Those possessing construct-relevant KSAs but who were either unfamiliar with the format of the items or less computer savvy tended to need more time for complex innovative items. Generalizing from the cognitive labs to the use of innovative items in high-stakes testing situations, it is important to consider that these examinees may be pressed for time and that the extended effort may cause fatigue and/or additional cognitive loading. This could impact their overall performance when students have not been adequately introduced to new item formats.

While a speededness factor related to the functionality of the items likely introduced construct-irrelevant variance in this study, the observations made in the course of this study suggest several potential solutions that could be implemented to lessen any impact of item functionality on student scores. Students must be given adequate practice time with the types of tools and item formats they may encounter in a test. This could be accomplished in a number of ways, such as through practice items or tests, tutorials before or during test administration, and increased alignment of instruction and assessment practices. Items that have many complex features could be deployed first as instructional tools or items for supporting formative assessment; once students become familiar and facile with these item types, the items can be reconfigured for summative use while minimizing usability-based construct-irrelevant variance.

Student Engagement

Increased engagement in task completion and higher examinee motivation may result in test scores that more accurately reflect a student's understanding of intended constructs. However, the observed enthusiasm for the innovative items tended to decrease as students got

older. Elementary students were observed to be highly engaged in each task and most enjoyed the interactive nature of the items. High school students, however, noted that the items were more interactive and visually appealing as regularly as they noted that the innovative items required greater time and effort to complete when compared to the "efficiency" of interpreting and responding to traditional, selected-response items. This sentiment was often expressed much more strongly in the area of mathematics and by high-performing high school students. It is important to note that these students did not seem less motivated to complete the innovative items, only that the interactive nature did little to encourage active engagement above and beyond what might be expected for a traditional item. Within this grade band, greater attention was directed at the implications of including more demanding item types within high-stakes assessments than at the "fun" factor of such items. Students in the lower grade levels, on the other hand, exhibited less self-awareness regarding these implications.

Curbing the Impact of Guessing

A dominant theme across all grade levels was a clear recognition by students that the constructed-response innovative items require them to produce their own answer; for many participants this was markedly more challenging than answering a multiple-choice item. When asked to compare innovative constructed-response and traditional multiple-choice items, participants in the mathematics cognitive labs commented on the perceived ease and confidence in responding to the latter. Several students indicated that they used test-taking strategies in combination with some degree of content knowledge to answer the multiple-choice items included in the mathematics cognitive labs. As with traditional constructed-response items, these innovative items require students to respond primarily on the basis of their content knowledge and guessing plays little or no role. For several items included in this study, students could and

did produce a partially correct answer. Unlike multiple-choice items, where a student who has some knowledge is either scored as knowing or not knowing the content, constructed-response innovative items allow for partial knowledge to be captured in the response, though much work still remains on how to best harvest and express such partial knowledge through a score. That said, many of the designs employed for innovative items in this study intentionally constrained student constructed responses to improve and standardize applicability of rubrics for either human or automated scoring.

Study Limitations

It is important to consider the context of the current research when interpreting the qualitative data produced from the cognitive labs and making inferences based on student verbalizations. Specifically, the cognitive labs were conducted in a low-stakes environment where the interaction between the moderator of the session and participant was intentionally friendly and conversational. The informal and relaxed tone facilitated such verbalization actions and processes and encouraged open dialogue. Inadvertently, it may also have affected the way that students interacted with items in the cognitive lab sessions. The low stakes associated with the session may have lowered the motivation that students felt to put forth their most concerted effort and may have resulted in some situations where students did not feel compelled to completely answer a question or fully read the item or directions. Additionally, participants were openly asked about their perceptions of the items and their attention repeatedly drawn to features of the items that made them unique, thus encouraging interface exploration beyond what would be expected in a normal testing environment—as well as potentially distracting them. Variations on the cognitive lab procedures used in this study, such as retrospective think-alouds in which

students only engage with researchers once they have completed an item or items, might also be considered in the future.

Implications and Recommendations

It must be emphasized that the data and evidence gathered in this research study provide a starting point for further investigations. This process must be iterative and ongoing as the field expands and explores new item types and the benefits these new item types might provide. In fact, approaching items in terms of item types with a focus on common interactions, features, and task types—as was done in this study—provides an opportunity to refine an interface and eliminate usability problems for the item type before using those features in multiple items of the same item type. The inferences drawn in this initial set of cognitive labs represent one piece in what must be a larger pool of evidence. As such, the results of this study can be used to plan the next iterations of data collection and to consider what other sources of data (e.g., recording of student interactions with interfaces in realistic testing situations; tracking of student eye movements), might be mined for additional information about item performance.

The intersection of item content and the user interface features is complex. The way these two vital aspects of computer-delivered items interact can impact how students demonstrate their KSAs in unpredictable ways. An increased understanding of interface options by the content developers will assist in creating families of items that are able to better measure hard to capture KSAs, facilitate the measurement of high-order thinking skills, provide a richer array of scoring options for the assessment of the degree of concept mastery, and allow construction of tests tailored for specific uses or specific students.

A theme that consistently arose in this study was the role that text-based instructions and written descriptions of the task(s) play in responding to innovative items. Across content area

and grades, there was great variability in the participants' awareness of and willingness to attend to text-based instructions. Additional research to better understand how examinees consume the information contained in an innovative item and determining how to facilitate examinees' use of this information (especially understanding the required response) will be helpful in informing the development of new items.

Although this preliminary study did not address accessibility issues directly, test fairness and reduction of construct-irrelevant variance for students with disabilities and English learners is of great concern and can be thought of as an extension of general validity and usability concerns for all students. Whereas our definition of innovative items did not include it, technology can be used to administer traditional items solely to reduce sources of construct-irrelevant variance, such as through use magnification, read-aloud, and keyboard-only navigation and response. Ideally, such expanded modes of administration should be considered from the start of the item development process, as per the principles of universal design (Almond et al., 2010; Dolan & Hall, 2001; Ketterlin-Geller, 2005; Thompson, Johnstone, & Thurlow, 2002). In doing so, we can move away from a post-hoc retrofit approach in which accommodations are relied upon for achieving usability and accessibility. In the current study we chose to focus our investigation first on general usability; in parallel we have been developing items designed to be usable by and accessible to students with disabilities and English learners using guidelines (Dolan et al., 2006) based upon the principles of Universal Design for Learning (Rose & Meyer, 2002). Future studies are needed to most closely examine how innovative items perform with students with disabilities, English learners, students with limited computer skills, and diverse students in general.

Though not explicitly explored in this study, advanced approaches toward scoring innovative item responses will likely provide richer information about what students know or do not know. For example, multistep items allow for processes to be explored and partially correct scores may be able to tell in greater detail what a student knows and does not know. Other items may allow students to demonstrate how deeply they understand a concept. At the extreme, we are currently developing prototype innovative items that require collaborative investigation in puzzlelike scenarios. While the potential for these new types of scoring is exciting and promising, many of the logistic (e.g., automated scoring, creating data structures to hold and deploy items) and psychometric (e.g., statistical dependencies) challenges remain to be explored.

Lastly, it is clear that the use of innovative items will likely be accompanied by an increase in testing time. The added time observed in this study has a number of possible explanations. First, the innovative items require more of the student cognitively and physically when producing an answer. Second, the cognitive labs by their nature also encourage exploration of the interface, and the moderators routinely asked students about their experience with the interface and features of the items during these sessions. Finally, the innovative items were presented to students with no opportunity for practice, and thus the unfamiliarity and novelty of the item may have required more initial exploration of the interface. Further studies are needed to more fully explore how much administration time the items require relative to the amount of measurement information and the quality of measurement that is obtained from these items.

References

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke, J., Torres, C., Haertel, G., Dolan, R. P., Beddow, P. & Lazarus, S. (2010). Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research. *Journal of Technology, Learning, and Assessment, 10*(5), 1-52.

Bachman, L. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues & Practice, 21*(3), 5-18.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & . C. Ward (Eds.), *Construction vs. Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment* (pp. 1-27). Hillsdale, NJ: Erlbaum.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*, 5–12.

Bennett, R.E., Goodman, M., Hessinger, J., Kahn, H., Ligget, J., Marshall, G., (1999). Using multimedia in large-scale computer-based testing programs. *Computers in Human Behavior, 15*(3), 283-294.

Bennett, R.E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *Journal of Technology, Learning, and Assessment, 8*(8). Retrieved 2/1/2011 from http://www.jtla.org.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17,* 10-17.

Branch, J. L. (2000). Investigating the information-seeking processes of adolescents: The value of using think-alouds and think afters. *Library and Information Science Research, 22*(4), 371–392.

Chalhoub-Deville, M. 2001: Task-based assessments: characteristics and validity evidence. In Bygate, M., Skehan, P. and Swain, M., editors, *Researching pedagogic tasks: second language learning, teaching and testing*. Harlow, England: Longman, 210–28.

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

Dolan, R. P., Burling, K. S., Harms, M., Beck, R., Hanna, E., Jude, J., Murray, E. A., Rose, D. H., & Way, W. (2006). *Universal Design for Computer-Based Testing Guidelines*. Retrieved May 4, 2009, from http://www.pearsonassessments.com/udcbt.

Dolan, R. P., & Hall, T. E. (2001). Universal Design for Learning: Implications for large-scale assessment. *IDA Perspectives*, *27*(4), 22-25.

Dolan, R. P., Rose, D. H., Burling, K. S., Harms, M., & Way, W. (2007). The Universal Design for Computer-Based Testing Framework: A structure for developing guidelines for constructing innovative computer-administered tests. Paper presented at the *National Council on Measurement in Education Annual Meeting*, April 10, Chicago, IL.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*(2), 24-35.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: Massachusetts Institute of Technology.

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Algebra on the SAT. *Journal of Technology, Learning, and Assessment, 6*(6).

Hambleton, R. K. (2000). Advances in performance assessment methodology. *Applied Psychological Measurement, 24*(4), 291-293.

Huff, K. L, & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20,* 16-25.

Jodoin, M.G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement, 40*(1), 1-15.

Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the Think Aloud Method (Cognitive Labs) to Evaluate Test Design for Students with Disabilities and English Language Learners* (Technical Report No. 44). Minneapolis, MN: University of Minnesota, National Center on Education Outcomes.

Kane, M.T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Washington, D.C.: American Council on Education/Praeger Publishers.

Kane, M.T. (1992). The assessment of professional competence. *Evaluation and the Health Professions, 15,* 163-182.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal of Technology, Learning, and Assessment, 4*(2), 1-23.

Kuechler, W.L., & Simkin, M.G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education, 8*(1), 55-75.

Kumar, D. D., White, A.L. & Helgeson, S.L. (1994). A study of the effect of HyperCard and pen-paper performance assessment methods on expert-novice chemistry problem solving. *Journal of Science Education and Technology*, *3*(3), 187-200.

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

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues & Practice, 13*(1).

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp.12-103). New York: American Council on Education and Macmillan.

Messick , S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5.

Messick, S., & Hakel, M. D. (1998). Alternative modes of assessment, uniform standards of validity. In *Beyond multiple-choice: Evaluating alternatives to traditional testing for selection.* (pp. 59-74). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessment. *Applied Psychological Measurement, 24*(4), 367-378.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2010). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 215-230). New York: Springer.

Quellmalz, E. S., & Pellegrino, J. (2009). Technology and testing. *Science, 323*(5910), 75-79.

Rose, D. H., & Meyer, A. (2002). *Teaching Every Student in the Digital Age: Universal Design for Learning*. Alexandria, VA: ASCD Press.

Rupp, A.A., Gushta, M., Mislevy, R.J., & Shaffer, D.W. (2010). Evidence-centered Design of Epistemic Games: Measurement Principles for Complex Learning Environments. *Journal of Technology, Learning, and Assessment, 8*(4). Retrieved 2/1/2011 from http://www.jtla.org.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing “Intermediate Constraint” questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6).

Strain-Seymour, E., Way, W. D., & Dolan, R. P. (2009). *Strategies and Processes for Developing Innovative Items in Large-Scale Assessments*. Iowa City, IA: Pearson Education.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (No. NCEO Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Education Outcomes.

Wendt, A., & Harmes, J. C. (2009). Evaluating innovative items for the NCLEX, Part 1: Usability and pilot testing. *Nurse Educator, 34*(2), 56-59.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337-362.