Chapter 13

Measuring 21st Century Reading Comprehension Through Scenario-Based Assessments

Jane R. Shore, Mikyung Kim Wolf, Tenaha O'Reilly, & John P. Sabatini

Educational Testing Service

This manuscript is an early draft of a paper published in English Language Proficiency Assessments and thus may differ slightly from the final published version. Please see below for the official paper:

Shore, J., Wolf, M. K., O'Reilly, T., & Sabatini, J. P. (2017). Measuring 21st century reading comprehension through scenario-based assessments. In M. K. Wolf, & Y. G. Butler (Eds.), *English Language Proficiency Assessments for Young Learners*, (pp. 234-252). New York, NY: Routledge.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305F100005 to the Educational Testing Service as part of the Reading for Understanding Research (RFU) Initiative. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

In the United States, educational thought leaders have called for higher expectations (Gordon Commission, 2010), more rigorous college and career readiness standards for K-12 education (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), and new constructs such as 21st century skills including collaborative problem solving in digital environments (Partnership for 21st Century Skills, Coiro, Castek, & Henry, 2013). They have also challenged assessment developers to address how best to provide information that is useful for instruction for various learners (Gordon Commission, 2013; Purpura & Turner, 2014; Watkins & Lindahl, 2010). Expanding the scope and variety of constructs (e.g., including elements like collaborative learning and digital literacy) will ensure measurement keeps pace with how people function and interact in various everyday reading activities.

This chapter describes a new assessment design approach called scenario-based assessment (SBA) and explains how it can be used to measure the reading ability of school-aged children in the current context of high standards. SBA combines a cluster of techniques for delivering a set of tasks and items that provide a platform for measuring the kinds of demanding reading skills, while simultaneously affording the potential to increase the instructional relevance of the assessment.

SBAs in reading typically include a range of principles and techniques that distinguish them from other types of assessments: (1) they provide an authentic purpose for reading, (2) they place reading in context for completing a set of interrelated activities that may move from more guided to independent performance, (3) items tend to require the integration and evaluation of a wide range of diverse sources and, (4) in many cases, items provide scaffolds (e.g., a graphic organizer for an analysis of text structures) and guidelines (e.g., tips for summary writing) to

help better understand and model the target performance in the assessment (O'Reilly & Sabatini, 2013). Some SBAs also include items that model the social aspects of literacy and learning, such as engaging with peers or a teacher to clarify understanding in reading, reviewing and evaluating peer writing. Using these principles, SBAs may broaden the range of interactions, perspectives, and information a test taker is exposed to on a topic. Ultimately, the key aims of scenario-based reading assessments are to *measure* 21st century reading ability while simultaneously *supporting* reading development and instructional usefulness.

In this chapter, we delineate two types of SBAs in reading, the Global, Integrated Scenario-Based Assessment (GISA) and English Language Formative Assessment (ELFA). These two assessments were part of two separate research projects. GISA was developed with a primary focus on benchmark or summative applications, across kindergarten through 12th grade. ELFA, on the other hand, was developed as a classroom-based, formative assessment of reading comprehension at the middle-school grade level. The GISA framework and design relied on computer delivery and principles from cognitive science, whereas ELFA was paper-based for its easier integration in daily instruction. Employing the SBA approach to developing reading assessments, both projects also aimed to build their SBAs to be feasible and practical, while maintaining adequate psychometric properties. Consequently, we also briefly describe some empirical evidence collected to date in support of these aims. We conclude this chapter with some considerations in designing SBA assessments based on the lessons we have learned from GISA and ELFA.

The Global, Integrated Scenario-Based Assessment

GISA Framework

The GISA (ETS, 2015) was developed under a federal research project called the Reading for Understanding (RfU) Initiative. The RfU initiative was funded with the overarching goal of improving reading comprehension though intervention and assessment research for K-12 students (Institute of Education Sciences, 2010). In the RfU project, the development of GISA began with the construction of a reading assessment framework designed to explain the purpose(s) of the assessment system, the constructs measured, the theoretical underpinnings, and the general design principles derived from a synthesis of the cognitive science literature. Consistent with evidence-centered design (Mislevy & Haertel, 2006), a series of framework papers was created to increase the transparency of the design *before* the assessments were created (O'Reilly & Sabatini, 2013, Sabatini & O'Reilly, 2013, Sabatini, O'Reilly, & Deane, 2013). With this documentation, potential users of the measures can make more informed decisions about whether to adopt the new assessments. The documentation also provides a partial road map for identifying and evaluating key claims underlying GISA's design.

To date, three installments of the reading framework have been developed for GISA. The first installment provides a set of general cognitive principles that guide the overall assessment design (Sabatini & O'Reilly, 2013). Some of these principles include the rationale for measuring the foundational components of reading, digital literacy, and purposeful reading.

The second installment provides a definition of the reading processes, the constructs to be measured, a position on reading development, and an overview of two types of assessments—component and higher-order skill measures (Sabatini, O'Reilly, & Deane, 2013). For the 21st century reading construct, the reading process is described as a set of purpose-driven activities, where one's goals serve as a standard for evaluating the quality and relevance of text sources (Linderholm, Virtue, Tzeng, & van den Broek, 2004; van den Broek, Lorch, Linderholm, &

Gustafson, 2001; McCrudden, Magliano, & Schraw, 2011). In modern reading environments, students are also expected to access and develop the language needed to comprehend a wide variety of texts (Bailey, 2007; Lesaux & Kieffer, 2010) build understanding within and across multiple sources (Britt & Rouet, 2012), engage in disciplinary reading (Goldman, 2012; Lee & Spratley, 2010; Shanahan, Shanahan, & Misischia, 2011), and evaluate (Graesser et al., 2007; Metzger, 2007) and integrate information in digital literacy environments (Coiro, 2009).

The third installment of the framework describes performance moderators and characteristics of SBA as applied to GISA. Performance moderators are factors that impact reading, but are not considered a direct part of the construct (for more information on performance moderators and their role in assessment, see O'Reilly & Sabatini, 2013). These include background knowledge (Shapiro, 2004), metacognition and self-regulation (Hacker, Dunlosky, & Graesser, 2009), motivation (Guthrie, & Davis, 2003), and reading strategies (McNamara, 2007).

The decision to include measures of performance moderators in the GISA design was twofold. First, the developers wanted to improve the interpretation of reading scores. For instance, if a student scores high on a reading test, does the score reflect high reading ability or high background knowledge? By having measures of performance moderators in the assessment, inferences about student performance can be enhanced. To further the example above, if the student has a lot of background knowledge on the topic, then they might not need to read the text deeply to answer the questions correctly. In this case, the reading test score might be questioned as it may reflect more about the students' background knowledge than their actual ability to read and comprehend text. Similarly, one might question the validity of a reading

score, if other performance moderator information was collected on the test that suggested the student was not motivated to try their best.

Second, GISA was designed to model and encourage good reading practices through the test itself. One might hypothesize that having students complete items that required metacognitive, self-regulatory, and other strategic reading behaviors may help promote their use in other academic contexts and could encourage more strategic reading. In this way, the use of the assessment itself serves as a tool *for* and *as* learning (Bennett & Gitomer, 2009).

Although GISA was designed to primarily measure higher-order reading skills (e.g., synthesis, evaluation, application), the reading framework also describes the need for measures of foundational reading skills. This is accomplished in a separate assessment battery called the Reading Inventory of Scholastic Evaluation (RISE). RISE consists of six computer-administered subtests that assess reading skills (e.g., decoding, morphology) identified in the literature as foundational to higher-order comprehension. Together, GISA and RISE are intended to provide a comprehensive picture of reading ability. As we describe later in the chapter, GISA and RISE can be used together to help determine whether a student has difficulties in higher-order reading comprehension or foundational reading skills (for more on foundational skills and the RISE battery, see Sabatini, Bruce, Steinberg, & Weeks, 2015).

SBA Approach to GISA

GISA measures higher-order reading comprehension by using SBA techniques to deliver a set of sequenced and thematically interrelated items. In GISA, students are presented with a purpose for reading a collection of related sources (e.g., to decide if a community garden is a good idea for their neighborhood). Sources include traditional forms of print such as a news

article, and more modern digital forms of communication such as web pages, e-mails, or simulated students' responses.

However, these higher-order skills are difficult, and a variety of students have not mastered them. For example, providing a test that asks students to write an essay that requires them to integrate a variety of perspectives from a range of sources is likely to reveal that many students cannot even begin to do this task. One might question the value of such unsupported assessment because the test does not provide any information on what parts of the complex task lower ability students can or cannot do. To this end, many of the tasks and activities within the GISA forms are sequenced both to *model* skilled performance and to *gather evidence* on what parts of a more complex task students can or cannot do. This sequencing is, in part, possible because the assessment is computer delivered and the order of items and tasks can be controlled. For instance, before students read any texts, their level of background knowledge is measured to help determine what they already know about a topic.

As mentioned previously, this performance moderator can be used to contextualize the reading score (e.g., did students already know the topic; thus the reading score is potentially compromised). However, the GISA assessments are also structured to *build up* students' understanding over the course of the assessment. For example, the first text in a GISA assessment typically describes the general issue (e.g., whether or not to create a community garden), subsequent texts then dig deeper into the issue (e.g., pros and cons of creating a community garden), and the final section requires the student to complete more complex tasks (e.g., integrate the information, make a decision, and communicate your understanding in a flyer). This way, the assessment design probes into progressively deeper, more complex literacy skills over time, while sampling what students can and cannot do along the way. This is not to

say that all GISA tasks are supported with modeling and scaffolding. Such an approach could result in complexity of the tasks always being reduced, and higher-order thinking would, in effect, not be assessed. However, when appropriate, the goal is to also model and elicit information on what parts of a complex task, students can, or cannot do. Thus, the assessment is designed to both *measure*, and *support* complex thinking.

To illustrate these issues, a short sequence from the community garden assessment intended for fifth and sixth graders is described. To measure independent performance, students are asked to write a summary about an article. Even though GISA is a reading assessment, the tasks are designed to measure integrated skills. In this example, a summary-writing task was designed to focus students' attention on constructing a more global representation of the text. To model desired performance, guidelines for writing a summary are provided. These guidelines contain suggestions such as to include only the main ideas, avoid adding one's own opinions, to paraphrase, etc.

Writing a summary is a difficult task; so even with the guidelines, many student's responses provide minimal evidence of their capabilities. However, this does not necessarily mean that students do not have some of the component skills needed to create summaries. Therefore, in addition to providing guidelines, other techniques are used to elicit desired evidence of partial skills. For instance, GISA assessments also include simulated peer and teacher interactions that facilitate the elicitation of test takers skills within the assessment. Continuing with the community garden example, after the test takers write their summary of the article, simulated peers show their written examples of text summaries. The peer summaries contain violations of the provided guidelines and the test taker is asked to *identify* the particular guideline that was not followed. In a subsequent task, the test taker is provided with the same

peer summary, but is now told which guideline was violated, and asked to highlight *where* the violation occurred. In the following task, the violation is highlighted and the test taker is asked to *fix* the error. Thus, in this four-part sequence, information is collected on whether a student can write a summary independently, identify if a given summary contains a violation, locate the violation, and correct the error.

Such sequencing and scaffolding techniques are not only useful for gathering more information about what students can or cannot do, but also help to model strategic reading behaviors (strategy use, metacognition, and self-regulation). Again, technology and computer delivery is critical to these aims, by allowing the test designer to control the sequence and flow of the tasks.

To date, the RfU team has developed over 20 GISA forms that are appropriate for students in kindergarten through 12th grade. Some forms contain the same structure and item types, but the content addresses different topics. The parallel structure can be useful in intervention evaluation designs, as well as in instructional programs. For instance, assessments with a parallel structure[1] can be used to evaluate the effectiveness of a reading intervention or alternatively, to measure changes in reading ability over time. While many of the skills measured in the assessments overlap, each assessment may emphasize some skill sets more than others (e.g., summary writing or disciplinary reading or error detection and repair). With a range of skill foci, educators can choose the particular assessment that best fit their needs – a system level feature anticipated in the framework (Sabatini, O'Reilly, & Deane, 2013)[2].

[1] Comparability of test forms requires that the forms are on the same scale or equated.

[2] For more information on GISA including some released screen shots of the assessment please visit the ETS website at: http://www.ets.org/research/topics/reading_for_understanding/.

Empirical Validity Evidence for GISA

During the development stage of GISA, evidence was collected to evaluate key claims that would support valid inferences about GISA use and scores. In a recently published study, O'Reilly, Weeks, Sabatini, Halderman, & Steinberg (2014) discuss the use of GISA forms as an outcome measure in a large-scale reading intervention evaluation. The intervention designers and evaluation team chose to use GISA because its theoretical foundation aligned well to the disciplinary-focused reading constructs that their intervention targeted. The report documents how the items, scores, and scales were evaluated to ensure that they met the research aims for this application of the tests.

In terms of the psychometric quality of the assessments, data have been collected on over 250,000 administrations across 28 states, sampled from urban, suburban and rural populations including both public and denominational schools. This work has shown that despite the novel interface and skills tested, the prototype forms are reliable (α=.80 or higher), and the range of scores shows wide variability. In other words, SBA seems to have adequate internal properties and is feasible to implement on computers in real school settings.

Moving beyond the basic psychometric properties of the test, other data indicate that certain features of the scenario could be useful for understanding more about students reading ability. For instance, O'Reilly and Sabatini (2015) found evidence to support the usefulness of the scenario-based sequencing technique. In the summary example mentioned earlier, items were sequenced to reveal what parts of the more complex task students could or could not handle. Data analyses revealed that, although the majority of students had difficulty writing a summary without support, many of them were able to complete the tasks that measure important

summary writing subskills (O'Reilly & Sabatini, 2015). For high scoring students, the subsequent "diagnostic" tasks serve as confirmation that their independent summary writing was undergirded by a solid understanding of the skills that enter into writing a strong summary of the article provided. On the other hand, for low scoring students, there was evidence to suggest the test takers could do some of the sub skills that fed into summarization skill (e.g., locate an error in a peer summary). Similarly, in the same study, when students were given support such as scaffolding, they were able to demonstrate evidence of complex thinking in a range of task types dealing with some components of argumentation (O'Reilly et al., 2015). While more research is needed to uncover the potential value of using SBA, there is preliminary evidence to suggest that it can both help elicit complex thinking and help identify what parts of a more complex task students can or cannot do.

While we are still exploring evidence to support the validity arguments for GISA, it is important to note the added value of the RISE components battery, which is a computer delivered non-scenario based assessment. The RISE may be used in conjunction with GISA. If a student scores low on GISA, the reading components measured by the RISE may be helpful in identifying foundational skill weaknesses that are impacting higher level comprehension.[3]

Higher-order comprehension skills as defined here are complex and require thinking, manipulating, synthesis, analysis, evaluation and applying concepts, facts and information. While these skills may be complex, they draw upon foundational reading skills such as accurate and efficient decoding, word recognition, and fluent text reading skills. Although more research is needed, we hypothesize that including assessments that measure foundational and higher-order comprehension may be particularly useful for teachers to identify students' underlying reading

[3] For more information on RISE, please see: http://rise.serpmedia.org.

difficulties while simultaneously engaging them in the kinds of complex reading tasks they are likely to encounter in classroom settings.

Using an SBA for English Learners: English Learner Formative Assessment (ELFA)

Now we turn to a second SBA example, ELFA. We also discuss English learner (EL)-specific design features of ELFA (e.g., activating background knowledge, scaffolding, and including tasks for both foundational and higher-order reading skills) integrating specific SBA design features.

ELFA Constructs

K-12 reading curricula place great importance on higher-order reading skills such as a close reading of complex texts, citing evidence from the texts to support a main idea, analyzing a text structure, or evaluating an argument (Bunch, Kibler, & Pimentel, 2012). Yet teachers of ELs in middle schools also need to constantly assess and monitor the progress of students' foundational reading skills (e.g., vocabulary knowledge, sentence-level understanding) as EL students' English language proficiency profiles vary greatly.

Addressing the need to engage EL students in rigorous reading tasks as well as to provide teachers with a classroom-based assessment tool for formative purposes, the ELFA assessment design framework (Wolf, Shore, & Blood, 2014) was developed to delineate ELFA constructs and task design features of the performance indicators and moderators. ELFA focuses on the measurement of ELs' basic and higher-order reading skills. The overall construct of ELFA is reading comprehension of persuasive texts at the middle-school level. It encompasses an array of skills that are based on an analysis of K-12 academic standards (e.g., Common Core State

Standards), academic language characteristics (Bailey, 2007; Schleppegrell, 2004), and subskills found to be differentially influential in EL reading comprehension (August, Francis, Hsu, & Snow, 2006; Carlo et al., 2008; Gottardo & Mueller, 2009; Lesaux & Kieffer, 2010; Proctor, Carlo, August, & Snow, 2005; Wong-Filmore & Snow, 2000). Figure 1 summarizes the constructs and skills covered in the ELFA assessment.

[Figure 1 here]

Design of ELFA Assessment Structure

The current version of ELFA includes nine assessment forms, three forms in each difficulty category (*Developing*, *Intermediate*, and *Experienced*). The intent of developing multiple forms was to provide a system of ongoing classroom assessment. The reading passages for each form were purposefully developed for the three levels of English reading proficiency. They vary in linguistic complexity, academic orientation, topic, and argument structure (for more information see Wolf, Blood & Shore, 2014). ELFA developers utilized readability software called *e-rater* and *TextEvaluator* to measure dimensions of the linguistic complexity of the passages (Sheehan, 2012; Sheehan, Kostin, & Napolitano, 2012). These tools provided developers with a profile of the linguistic complexity of each reading passage (e.g., the total number of words, lexical density, number of academic words, complexity of sentence structures, grade-level difficulty indices). All reading passages were also rated by ESL teachers at the middle-school level for appropriateness of topic, interest, relevance, and language complexity for their students and feedback was provided on which were most relevant, engaging and appropriate for each level.

In designing ELFA assessment forms, aside from the construct, two major design factors were taken into consideration. First, the assessment needed to be easily integrated into daily instruction for formative purposes. Second, it needed to provide opportunities for ELs to collaborate with peers while engaging in the assessment tasks. To support these two design characteristics, each form of the ELFA assessment was made up of two parts, Parts 1 and 2, both based upon two reading passages. The two reading passages are referenced in both Parts and present opposing viewpoints from two authors regarding one topic. Part 1 of each assessment form is designed to be completed with a peer, and to provide scaffolding activities to help ELs unpack a given passage and sequentially utilize basic to high-order reading comprehension skills. Teachers are also encouraged to observe, take notes, and participate in student discussions during Part 1 tasks. Since Part 1 is completed in pairs and with teachers' engagement, it does not provide individual students' reading ability. Hence, Part 2 was added to each assessment form in order to measure students' individual reading ability. In Part 2, students completed the tasks independently.

Scenario-Based Task Design in ELFA

SBA features were applied in developing Part 1 tasks, which include both selected-response and constructed-response tasks. The warm-up activities and main tasks follow a sequence of authentic reading activities (Wolf, Shore & Blood, 2014). All activities were designed to engage students in a realistic reading context by providing a purpose for reading, an authentic sequence of reading activities that move from general to specific while progressing to more challenging skills that require students to synthesize information, evaluate reasoning, and gather supporting evidence to support an argument.

Similar to GISA, ELFA's Part 1 begins with a scenario to establish a purpose for reading, like the one shown in Figure 2. For example, an authentic reading situation is provided for students in the beginning of each assessment (e.g., to prepare for a class discussion, to find specific information, to agree/disagree with the author, to evaluate the adequacy of arguments and evidence).

[Figure 2 here]

As students progress to higher-order reading tasks in Part 1, they also encounter scaffolding tasks that focus on foundational skills. In this way, the SBA-based ELFA forms evaluate not only a student's higher-order reading skills, but also foundational reading skills, identifying subskill challenges that could impede higher-order reading comprehension. The tasks are also designed to provide formative information that identifies which reading subskills might require more instructional attention.

ELFA uses the SBA approach to embed scaffolded tasks in an authentic and meaningful sequence. ELFA scaffolding and sequencing are designed to support: (1) a reading process that would engage readers to accomplish a given reading purpose (Linderholm et al., 2004); (2) tasks that would help EL students unpack the passage to build comprehension (Beck, McKeown, & Kucan, 2002; Biemiller, 2005; Francis, Rivera, Lesaus, Kieffer, & Rivera, 2006; Graves, 2000, 2006; Taboada, 2009); (3) tasks that reinforce students' close reading of the text (Silverman & Hines, 2009; Stahl & Fairbanks, 1986); and (4) tasks that would foster students' use of reading strategies (Carlo et al., 2008, 2009; DeLuca, 2010; Taboada, 2009). A guiding principle for designing the sequence of assessment tasks is to mirror actual stages in the negotiation of textual meaning that a typical EL middle-school student might experience. Figure 3 displays a high-level description of the Part 1 task sequencing.

[Figure 3 here]

This sequence is incorporated into the scenario for each assessment form. One of the intents in this sequence is that students build comprehension of texts, as they move through the purposefully-ordered tasks. To serve the role of scaffolding (particularly important for EL students to complete the given tasks), the tasks are designed with the following principles. First, the tasks are completed based on students' comprehension of the text, not on their test-taking strategies. Second, the tasks provide explicit strategies that the EL students can use to help them complete the tasks successfully. Third, in some cases, the task questions can provide essential information that a student needs in order to begin. By designing the tasks with scaffolding in mind, it is anticipated that teachers can also use the tasks selectively depending on their students' abilities and learning goals.

To illustrate a portion of the sequence, three sample tasks are provided below. Figure 4 presents the first activity that students see in Part 1, a warm-up activity to activate students' background knowledge and increase their interest in a given reading topic.

[Figure 4 here]

Then, students read the first passage and Part 1 main tasks begin by asking students to identify the main idea of the passage they read. Subsequent tasks involve close reading of the passage and sorting the details and a main idea, as shown in Figure 5.

[Figure 5 here]

This task is followed by a few foundational skill tasks for teachers to determine whether students' difficulties in identifying a main idea and details were attributed to lexical and syntactic knowledge in certain sentences. Toward the end of Part 1, the tasks assess the students' higher-

order reading skills, where they have to identify reasons and details by comparing and integrating information across multiple sources (See Figure 6).

[Figure 6 here]

ELFA also includes teacher resource materials to accompany the ELFA assessment forms. The main materials are the ELFA Teachers' Guide (Shore, Wolf, & Blood, 2013) and the ELFA Observation and Teacher Probes. These documents describe how teachers can use the scenario-based ELFA assessment tasks for formative purposes as part of their instruction. As ELFA was designed for classroom use, inherent in the design framework is the collection of additional evidence. This evidence is collected not only through the assessment items themselves, but also through teachers' observation and probing questions during Part 1 of ELFA.

Collecting Validity Evidence for ELFA

A number of pilot and field studies were conducted during the development process to explore the item properties, usability and applications of ELFA in classroom settings. First, pilot studies were conducted for all nine forms, focusing primarily on task and item qualities, both quantitatively and qualitatively. Analyses were done on the forms to determine the internal consistency of the items, confirm item difficulty level, and demonstrate discrimination among items and between levels. At the form level, internal consistency reliability estimates were moderate, ranging from .73 to .84. Overall, however, the reliability estimates were found to be at an acceptable level for classroom-based assessments. The correlation coefficients between Part 1 and Part 2 scores ranged from .67 to .78 across the forms. These moderate correlations were not surprising as Part 1 was completed collaboratively with Part 2 being done individually. In addition, Part 1 and Part 2 item types were somewhat different despite the fact that they

measured different aspects of the same construct. Part 1, the formative assessment done collaboratively, had several constructed-response items and Part 2, the independently completed test, consisted of only selected-response items.

Usability studies were also conducted using a collective case-study approach (Shore, Wolf, & Heritage, 2016). As formative assessment, by definition, centers on the teacher's practice and process of collecting evidence of learning to inform instructional next steps, a usability study to investigate teachers' use of ELFA is an essential step in ELFA's validation work. The results of this collective case study indicated that ELFA was seen as adding unique value to classroom tools available for EL students. In particular, the scenario-based design approach to meaningfully sequence the crucial reading skill tasks, as well as scaffolding tasks to ultimately engage ELs in grade-level higher-order reading tasks, was perceived positively by the teachers who participated in the usability studies. Further, the results suggested that teachers enhanced their understanding about formative assessment by way of implementing assessments that were specifically designed for formative purposes like ELFA (Shore, Wolf, & Heritage, 2016).

Essential Considerations in Developing Scenario-Based Reading Assessments

In this chapter, we described how SBA design features could be applied in creating new reading assessments for school-aged children. We focused on illustrating concrete examples of scenario-based tasks using two research and development projects, GISA and ELFA. We described how tasks were designed to measure higher-order reading skills expected of school-aged children. We also briefly discussed ongoing research to evaluate the validity of claims stemming from the construct frameworks and intended uses of the tests.

Students, especially those who may struggle to read subject-area texts, are best served with sensitive, engaging measurement tools that can inform instruction (Francis et. al, 2006; Turner & Purpura, 2016). Whether outcome-based or formative, classroom-based, the value of reading comprehension measurement is enhanced when it can be used to identify learner challenges, take into account students' knowledge, and when it can inform decisions with regard to student learning. Drawing from the prototype development and empirical research we conducted, we now summarize a few key elements to consider in designing scenario-based reading assessment for both native English-speaking and EL students. These may be broken down to describe how SBAs have addressed three key issues: (1) measuring 21st century reading abilities in complex and evolving literacy environments, (2) supporting the learning of essential reading skills while engaging in assessment, and (3) ensuring that results are instructionally meaningful.

As described in this chapter, increased attention on ensuring assessments attend to complex and evolving reading skills has inspired innovation. Both assessments described in this chapter attempt to address the primary goal of reading assessment innovation by the use of SBAs. First, GISA uses specific scenarios, to measure a variety of integrated and complex higher-order reading skills aligned with the 21st century skills, such as multiple text comprehension, disciplinary literacy, digital literacy, and perspective taking. It also captures information on performance moderators, such as background knowledge, to help interpret test scores, and in the case of reading strategies, to help encourage good habits of mind. While higher-order reading skills are assessed, tasks and activities are sequenced and modeled to help gather information on whether students can complete tasks that contribute to understanding of

more complex skills. In these ways, GISA integrates components of reading comprehension in authentic and meaningful ways.

ELFA takes a different approach to SBA, but also aims to measure the multilayered processes of reading comprehension, specifically those involving the reading abilities of EL students. It uses scenarios for tasks that work from foundational to higher-order skills progressively, using collaborative and individual forms, to assess and describe EL reading profiles. Using SBA techniques, both approaches are offered to meet the challenge of measuring multi-faceted reading processes.

Both GISA and ELFA also support the idea of engaging in a learning activity while completing a measurement task. That is, both GISA and ELFA assessment tasks are designed to be learning experiences themselves. GISA forms work through scenarios, engaging learning in strategic reading behaviors, and mirroring activities that support reading, like reflection and peer interaction, through tasks themselves. ELFA is also designed to echo an authentic learning experience in reading, moving learners through the stages reflected in the reading process. Collaborative and individual forms, along with Teacher Probes that guide teachers to extract individualized information on learning, further underlie ELFA's SBA approach as a learning activity as well as a way to gather measurement information. As further illustration, teachers in the ELFA case study reported that using ELFA was like a form of professional development on instructional approaches to formative assessment and reading components, indicating that this type of SBA could be easily integrated into instruction (Shore, Wolf & Heritage, 2016).

Finally, a goal in assessments such as GISA and ELFA is to ensure that the results are instructionally meaningful. In this respect, GISA not only measures higher-order reading skills, but also the subskills that feed into it and performance moderators like background knowledge to

help contextualize the reading score. This combination of information is aimed at providing instructional relevance, ensuring that information about an individual's skill level can be parsed apart and analyzed to ensure that information truly relevant to individual reading challenges, at a granular level. ELFA is framed in the same way, to provide evidence that is meaningful to instruction. In this case, ELFA's collaborative form involves teacher interaction and instructional engagement guided by Teacher Probes, making the form itself a prompt to collect instructionally relevant information. In these ways, both GISA and ELFA intend to get to the essence of reading challenges, ensuring these challenges exposed by reading tasks can inform specific pedagogical decisions.

In the effort to bring purpose and engagement to assessment designs to foster both learning and teaching in 21st century environments, SBAs represent a promising set of techniques that broaden the construct of reading to accommodate different needs. However, continued empirical studies to support the benefits of SBAs for both teachers and learners are necessary.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grants R305F100005, R305G04065, and R305A100724 to Educational Testing Service as part of the Reading for Understanding Research Initiative and the English Learners Formative Assessment research grant programs. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We would like to thank members of ETS's Cognitively Based Assessment *of, for* and *as* Learning team (CBAL) for their collaboration on these two projects.

References

August, D., Francis, D., Hsu, H.-Y. A., & Snow, C. (2006). Assessing reading comprehension in bilinguals. *Instructional Research on English Learners. Special Issue of Elementary School Journal, 107*(issue numbers), 221–239.

Bailey, A. L. (2007). *The language demands of school: Putting academic English to the test.* New Haven, CT: Yale University Press.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life.* New York, NY: Guilford Press.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–62). New York, NY: Springer.

Britt, A., & Rouet, J. (2012). *Learning with multiple documents: Component skills and their acquisition*. In M. J. Lawson & J. R. Kirby (Eds), The quality of learning: Dispositions, instruction, and mental structures (pp. 276–314). Cambridge, UK: Cambridge University Press.

Bunch, G. C., Kibler, A., & Pimentel, S. (2012). *Realizing opportunities for English learners in the Common Core English Language Arts and disciplinary Literacy Standards*. Commissioned paper by the Understanding Language Initiative. Stanford, CA: Stanford University. Retrieved from http://ell.stanford.edu/papers/practice

Coiro, J. (2009). Rethinking reading assessment in a digital age: How is reading comprehension different and where do we turn now? *Educational Leadership, 66,* 59–63.

DeLuca, E. (2010). Unlocking academic vocabulary. *The Science Teacher, 77,* 27–37.

ETS (2015). Reading for Understanding. Retrieved from: http://www.ets.org/research/topics/reading_for_understanding/

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for instruction and academic interventions*. Portsmouth, NH: Research Corporation, Center on Instruction. Retrieved from https://www2.ed.gov/about/inits/ed/lep-partnership/interventions.pdf

Goldman, S. (2012). Adolescent literacy: Learning and understanding content. *Future of Children, 22,* 89–116.

Gordon Commission (2013). *To assess, to teach, to learn: a vision for the future of assessment*. Princeton, NJ: Author. Retrieved from http://www.gordoncommission.org/rsc/pdfs/gordon_commission_technical_report.pdf

Gottardo, A., & Mueller, J. (2009). Are first and second language factors related in predicting L2 reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology, 101,* 330–344.

Graesser, A. C., Wiley, J., Goldman, S., O'Reilly, T., Jeon, M., & McDaniel, B. (2007). SEEK web tutor: Fostering a critical stance while exploring the causes of volcanic eruption. *Metacognition and Learning, 2,* 89–105.

Guthrie, J., & Davis, M. (2003). Motivating struggling readers in middle school through an engagement model of classroom performance. *Reading and Writing Quarterly, 19,* 59–85.

Hacker, D. J, Dunlosky, J., & Graesser, A. C. (2009). *Handbook of metacognition in education*. Mahwah, NJ: Erlbaum.

Institute of Education Sciences. (2010). *Reading for Understanding initiative*. Washington, DC: U.S. Department of Education. Retrieved from http://ies.ed.gov/ncer/projects/program.asp?ProgID=62

Lee, C. D., & Spratley, A. (2010). *Reading in the disciplines: The challenges of adolescent literacy*. New York, NY: Carnegie Corporation.

Lesaux, N. K. & Kieffer, M. J. (2010). Exploring sources of reading comprehension difficulties among language minority learners and their classmates in early adolescence. *American Educational Research Journal, 47,* 596–632.

Leu, D., Kinzer, C., Coiro, J., Castek, J. & Henry L. (2013). New literacies: A dual-level theory of the changing nature of literacy, instruction, and assessment. In D. E. Alvermann, N. J. Unrau, & R. B. Ruddell, (Eds.), *Theoretical models and processes of reading* (6th ed., pp. 1150–1181). Newark, DE: International Reading Association.

Linderholm, T., Virtue, S., Tzeng, Y., & van den Broek, P. (2004). Fluctuations in the availability of information during reading: Capturing cognitive processes using the landscape model. *Discourse Processes, 37,* 165–186.

McCrudden, M. T., Magliano, J. P., & Schraw, G. (Eds). (2011). *Text relevance and learning from text.* Greenwich, CT: Information Age Publishing.

McNamara, D. S. (2007). *Reading comprehension strategies: Theories, interventions, and technologies.* Mahwah, NJ: Erlbaum.

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology, 58,* 2078–2091.

Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25,* 6–20.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects. Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

O'Reilly, T. & Sabatini, J. (2015, July). *Effect of local and global reading skills on argumentation skill.* Paper presented at the Society for Text and Discourse conference, Minneapolis, MN.

O'Reilly, T., & Sabatini, J. (2013). *Reading for Understanding: How performance moderators and scenarios impact assessment design* (ETS RR-13-31). Princeton, NJ: ETS.

O'Reilly, T., Sabatini, J., Bruce, K., Pillarisetti, S., & McCormick, C. (2012). Middle school reading assessment: Measuring what matters under an RTI framework. *Reading Psychology Special Issue: Response to Intervention, 33,* 162–189.

O'Reilly T., Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing reading comprehension assessments for reading interventions: How a theoretically motivated assessment can serve as an outcome measure. *Educational Psychology Review, 26,* 403–424.

O'Reilly, T., Sabatini, J., Halderman, L., Bruce, K., Weeks, J., & Steinberg, J. (2015, March). *Building theoretical and developmentally sensitive reading assessments for students in 3rd through 12th grade: implications for intervention, and potential changes in reading proficiency*. Paper presented at Society for Research on Educational Effectiveness conference, Washington, D. C.

O'Reilly, T., & Sheehan, K. (2009). *Cognitively based assessment of, for, and as learning: A framework for assessing reading competency* (ETS RR-09-26). Princeton, NJ: ETS.

Partnership for 21st Century Skills. (2004). *Learning for the 21st century: A report and mile guide for 21st century skills.* Washington, DC: Author. Retrieved from http://www.p21.org/storage/documents/P21_Report.pdf

Partnership for 21st Century Skills. (2008). *21st Century skills and English map.* Washington, DC: Author. Retrieved from http://www.p21.org/storage/documents/21st_century_skills_english_map.pdf

Proctor, C. P., Carlo, M. S., August, D., & Snow, C. E. (2005). Native Spanish-speaking children reading in English: Toward a model of comprehension. *Journal of Educational Psychology, 97,* 246–256.

Purpura, J. E., & Turner, C. E. (Fall, 2014). *A learning-oriented assessment approach to understanding the complexities of classroom-based language assessment*. New York, NY: Teachers College, Columbia University. Retrieved from http://www.tc.columbia.edu/tccrisls/

Sabatini, J., Albro, E. & O'Reilly, T. (2012). Measuring up: Advances in how we assess reading ability. Lanham, MD: Rowman & Littlefield.

Sabatini, J., Bruce, K., & Steinberg, J. & Weeks, J. (2015). *SARA reading components tests, RISE forms: Test design and technical adequacy 2nd Edition. ETS RR* Princeton, NJ: ETS.

Sabatini, J., & O'Reilly, T. (2013). Rationale for a new generation of reading comprehension assessments. In Miller, B., Cutting, L., & P. McCardle (Eds), *Unraveling reading comprehension: Behavioral, neurobiological, and genetic components* (pp. 100–111). Baltimore, MD: Brookes Publishing.

Sabatini, J., O'Reilly, T., & Albro, E. (2012). Reaching an understanding: Innovations in how we view reading assessment. Lanham, MD: Rowman & Littlefield.

Sabatini, J., O'Reilly, T., & Deane, P. (2013). *Preliminary reading literacy assessment framework: Foundation and rationale for assessment and system design*. (ETS RR-13-30). Princeton, NJ: ETS.

Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective*. Mahwah, NJ: Erlbaum.

Shanahan, C. Shanahan, T., & Misischia, C. (2011). Analysis of expert readers in three disciplines: History, mathematics, and chemistry. *Journal of Literacy Research, 43,* 393–429.

Shapiro, A. M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal, 41,* 159–189.

Sheehan, K.M. (2012). *A cognitively-based text analysis system designed to help test developers ensure that admissions assessments incorporate suitably complex text.* Riyadh, Saudi Arabia: National Center for Assessment in Higher Education.

Sheehan, K. M., Kostin, I., & Napolitano, N. (2012, April). *SourceRater: An automated approach for generating text complexity classifications aligned with the Common Core Standards.* Paper presented at the National Council on Measurement in Education, Vancouver, BC.

Shore, J., Wolf, M. K., & Blood, I. (2013). *English Learner Formative Assessment (ELFA) Teacher's Guide.* Princeton, NJ: ETS. Available at https://www.ets.org/s/research/pdf/elfa_teachers_guide.pdf

Shore, J., Wolf, M. K., & Heritage, M. (2016). Formative Assessment Tools as Teacher Professional Development: A Case Study in the Use of the English Language Formative Assessment System.

Silverman, R., & Hines, S. (2009). The effects of multimedia-enhanced instruction on the vocabulary of English-language learners and non-English language learners in pre-kindergarten through second grade. *Journal of Educational Psychology, 101,* 305–314.

Stahl, S. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56,* 72–110.

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in the classroom. In D. Tsagari & J. Banerjee (Eds.). *Handbook of second language assessment* (pp. 255-274). Berlin, Germany/Boston, MA: DeGruyter Mouton.

Taboada, A. (2009). English language learners, vocabulary, and reading comprehension: What we know and what we need to know. *Yearbook of the College Reading Association, 30,* 307–322.

van den Broek, P., Lorch, R. F., Jr., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29,* 1081–1087.

Watkins, N., & Lindahl, K. (2010). Targeting content area literacy instruction to meet the needs of adolescent English language learners. *Middle School Journal, 4,* 23–33.

Wolf, M.K., Shore, J., & Blood, I. (2014). *English Learner Formative Assessment (ELFA): A*

design framework. Princeton, NJ: ETS. Retrieved from https://www.ets.org/s/research/pdf/elfa_design_framework.pdf

Wong Fillmore, L., & Snow, C. (2000). *What teachers need to know about language*. Washington, DC: Center for Applied Linguistics.

Figures

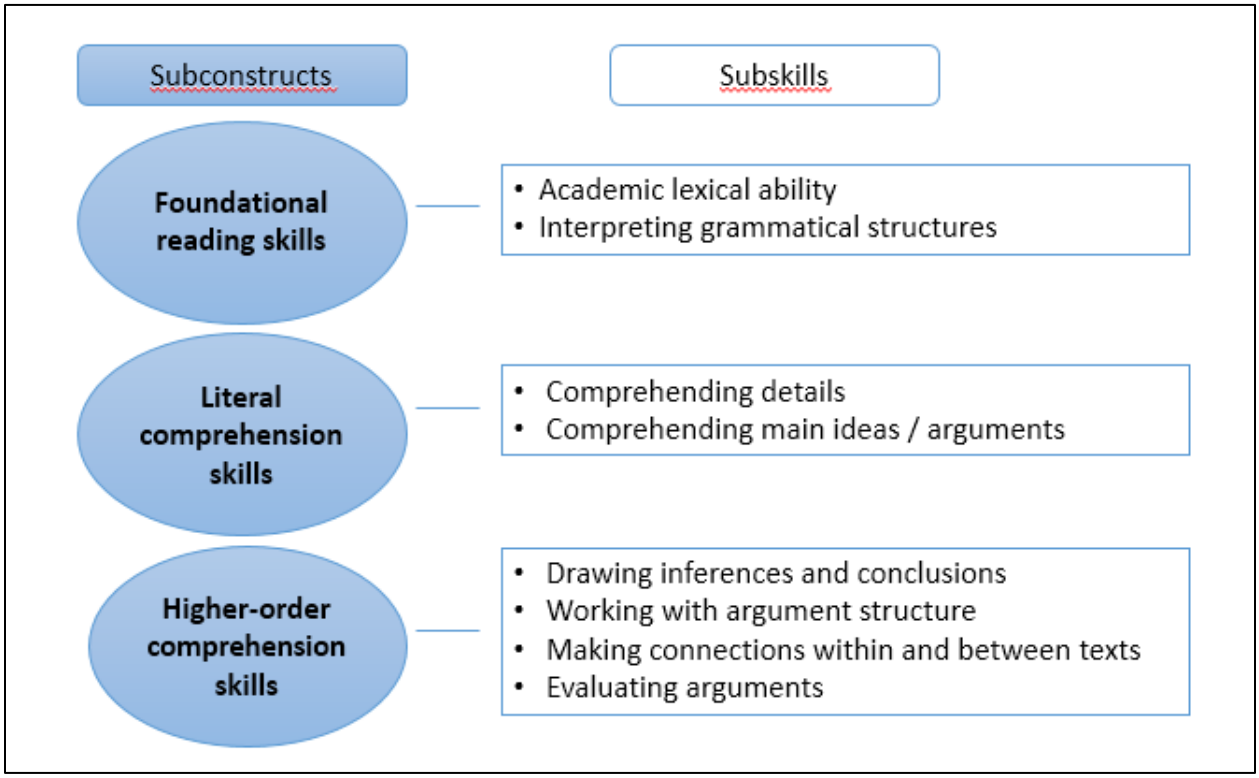

Figure 1. ELFA subconstructs and subskills.

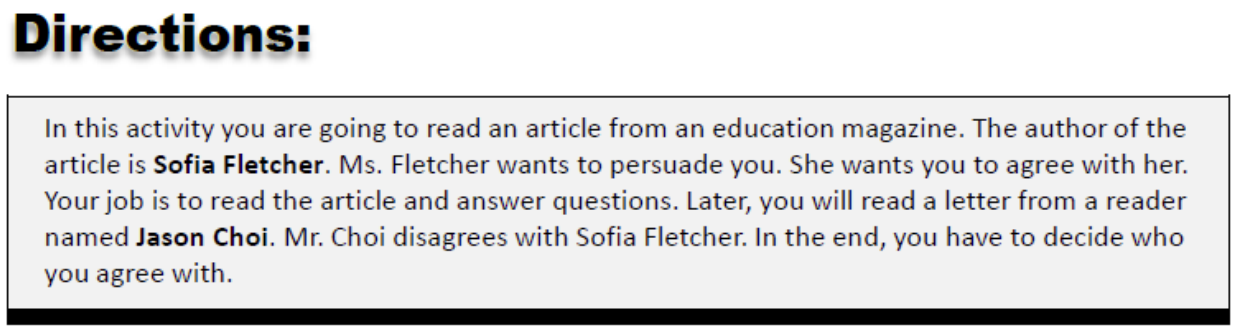

Figure 2. ELFA example item directions.

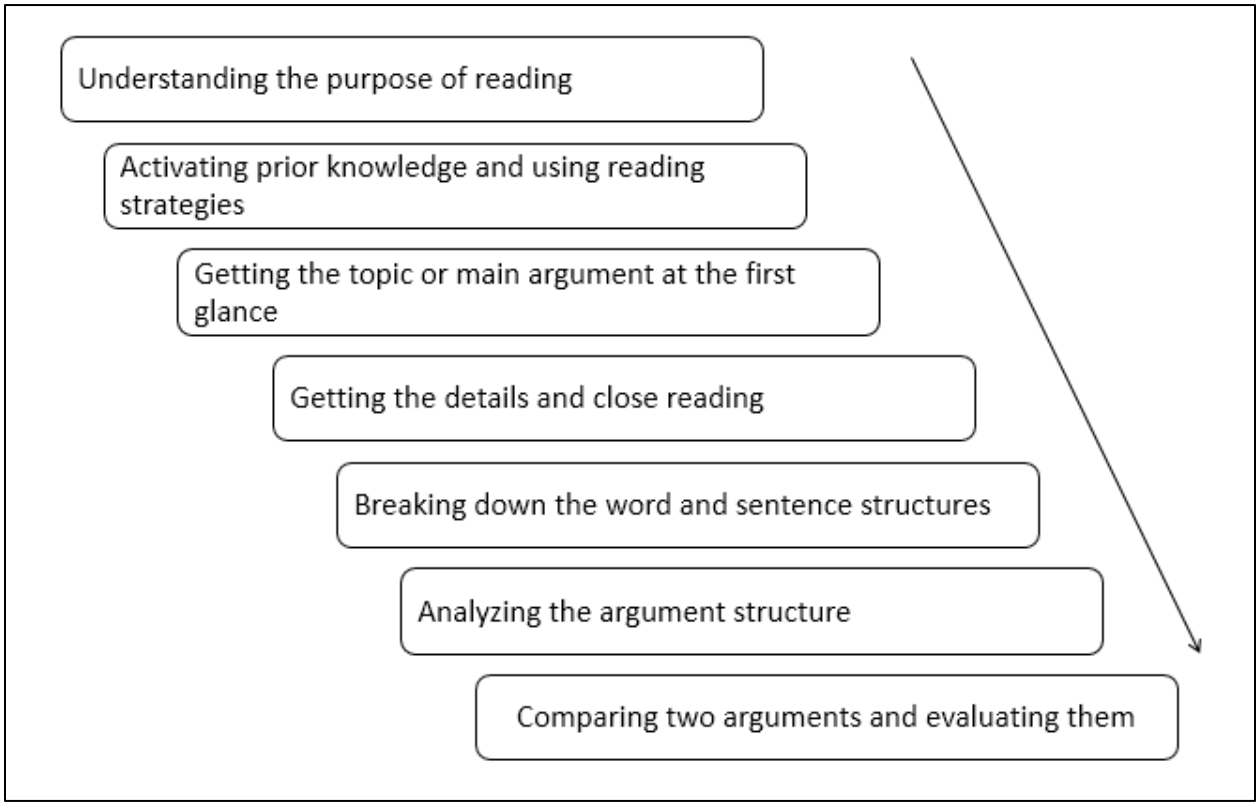

Figure 3. Task sequencing in ELFA.

Before you read...

Look at the article by Sofia Fletcher. Look at the title, the images, and the first sentence of each paragraph.

Where should cell phones be allowed? Where should they *not* be allowed? What is your opinion? Discuss with your partner and make two lists.

Places where cell phones should be allowed:

- _____
- _____
- _____
- _____

Places where cell phones should not be allowed:

- _____
- _____
- _____

Figure 4. Task sample: Warm-up.

Directions: Read each sentence in the ANSWER CHOICES. One of the sentences is the **main idea** of the article. Three of the sentences are **other ideas** in the article. Two sentences are **not in the article**. Write letters in the blanks in the table below to show where each sentence belongs.

ANSWER CHOICES
A. Cell phones with internet access can be expensive.
B. Cell phones should be allowed in schools, but with clear and strict rules.
C. 63 percent of all students send text messages every day.
D. Cell phone trucks are a new service that can be found near some schools in New York City.
E. Students, teachers and parents have different thoughts about cell phones in schools.
F. Cell phones became very popular in the 1990s because of advertising.

Main Idea	Other Ideas	Not In The Article
2 ____	3 ____ 4 ____ 5 ____	6 ____ 7 ____

Figure 5. Task sample: Getting a main idea.

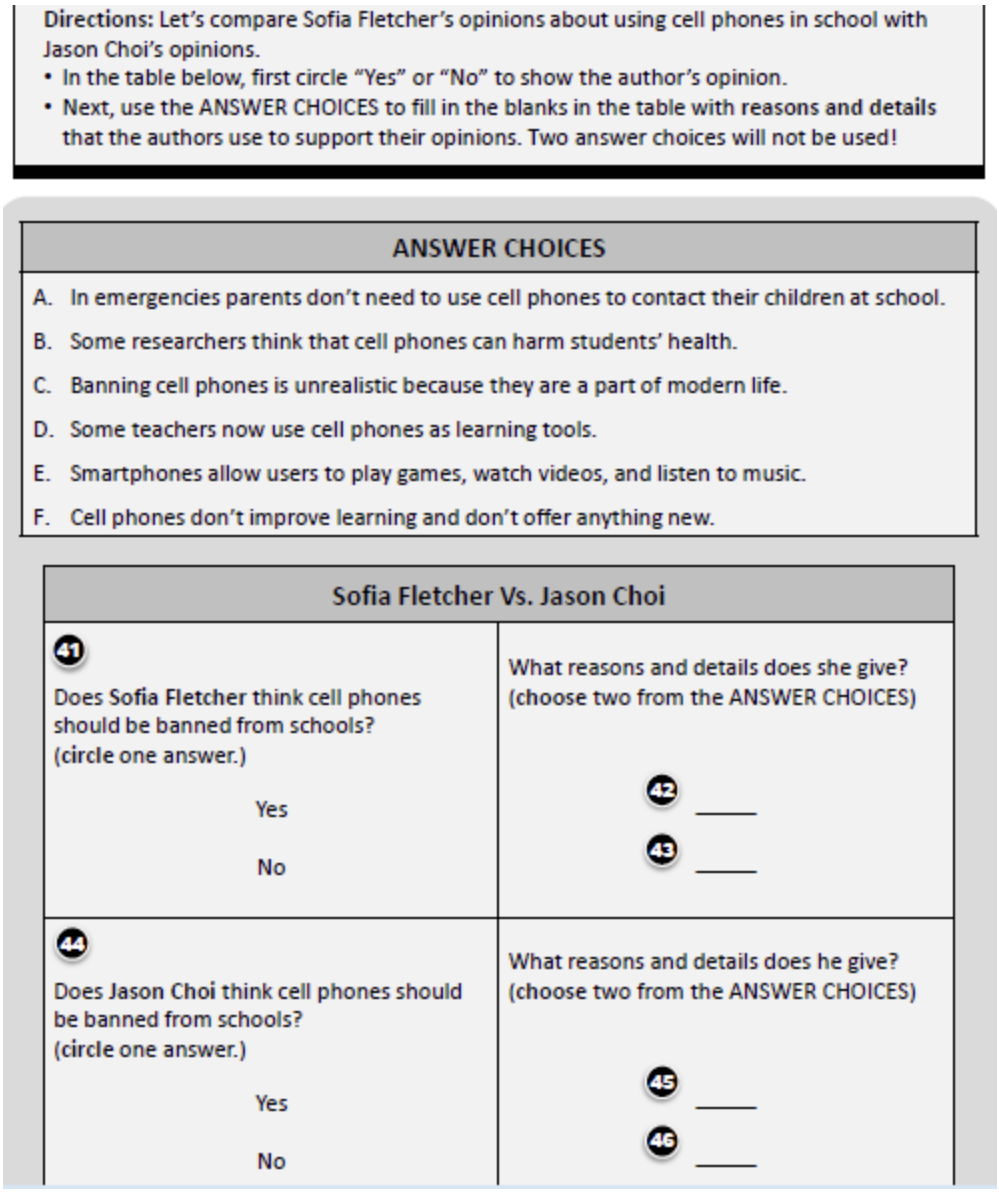

Figure 6. Task sample: Comparing and integrating multiple sources of information.