Engineering a 21st Century Reading Comprehension Assessment System Utilizing Scenario-based Assessment Techniques

John Sabatini

Tenaha O'Reilly

Jonathan Weeks

Zuowei Wang

Educational Testing Service

Correspondence should be addressed to John Sabatini, Institute for Intelligent Systems, 365 Innovation Drive, Suite 303, Memphis, TN 38152-3115, USA. Email: jpsbtini@memphis.edu.

Office phone: (901) 678-5102.

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. We are grateful to the Institute of Education Sciences and Educational Testing Service for sponsoring and supporting this research. We would like to also like to thank Michael Kane and Robert Mislevy for their intellectual insights and thoughtful comments; Andre Rupp, Paul Deane and Heather Buzick for their review comments; Kelsey Dreier and Kim Fryer for their editorial assistance. We also wish to express our gratitude to the all of our RfU and Cognitively Based Assessment as, of, & for, Learning (CBAL™) colleagues, who have been and are contributing to this ongoing enterprise.

Peer Review Process: International Journal of Testing (IJT) is a double-blind peer-reviewed topical journal.

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

Sabatini, J., O'Reilly, T., Weeks, J., & Wang, Z. (2019). Engineering a 21st Century reading comprehension assessment system utilizing scenario-based assessment techniques. International Journal of Testing. https://doi.org/10.1080/15305058.2018.1551224

Abstract

The construct of reading comprehension has changed significantly in the 21st century, however, some test designs have not evolved sufficiently to capture these changes. Specifically, the nature of literacy sources and skills required has changed (wrought primarily by widespread use of digital technologies). Modern theories of comprehension and discourse processes have been developed to accommodate these changes, and the learning sciences have followed suit. These influences have significant implications for how we think about the development of comprehension proficiency across grades. In this paper, we describe a theoretically driven, developmentally sensitive assessment system based on a scenario-based assessment paradigm, and present evidence for its feasibility and psychometric soundness.

Keywords: Reading comprehension assessment; reading for understanding; scenario-based assessment

Engineering a 21st Century Reading Comprehension Assessment System Utilizing Scenario-based Assessment Techniques

Digital forms of literacy are reshaping the genres and nature of literacy practices, and consequently the construct of reading comprehension in the 21st century (Leu, Kinzer, Coiro, Castek, & Henry, 2017). Yet some reading comprehension assessment designs are largely the same as decades ago. That is, examinees generally read a set of unrelated passages and respond to questions associated with each. However, a wide range of initiatives and new standards, such as the Common Core State Standards (National Governor's Association Center for Best Practices & Council of Chief State School Officers [NGA & CCSSO], 2010), Partnership for 21st century skills (Partnership for 21st Century Skills, 2008), and the Gordon commission on the future of assessment (Gordon Commission, 2013), have opened the door to assessment reform. Building upon this spirit of reform, in this article we present a relatively new test design approach called scenario-based assessment (SBA), as well as the architecture for building a developmentally-sensitive system of assessments that spans primary through secondary reading comprehension development.

SBA, broadly, represents a cluster of techniques for organizing and sequencing a set of thematically-related sources and items. This feature is particularly relevant in a digital environment, where individuals read for a specific purpose as they access and evaluate related texts on a single device. Many reading comprehension assessments have migrated to computer-based platforms (cf. PISA; Organisation for Economic Co-operation and Development, 2017), yet relatively little has changed with respect to the tried-and-true designs established in the traditional reading comprehension passage-question paradigm. On the surface, SBA designs may appear to be radical departures from this more traditional approach; hence, it is important to

evaluate the validity of this type of assessment design and assure assessment experts that SBAs can be implemented in a manner that satisfies expected measurement standards, such as feasibility of implementation and scoring, and psychometric rigor.

The SBA design presented here is intended to accommodate changes in the world of literacy (wrought primarily by the widespread use of digital technologies), incorporate modern theories of comprehension and discourse processes (Braasch, Braten & McCrudden, 2018; Magliano, McCrudden, Rouet, & Sabatini, 2018), incorporate insights from the learning sciences (Ercikan & Pellegrino, 2017), and be sensitive to how comprehension proficiency changes over developmental periods of time. These elements represent changes in the domain of literacy (what we read has changed), the cognitive processes applied when reading (how we read has changed), learning (how we learn to read has changed), and development (how we develop proficiency across the years of our schooling has changed).

In the following sections, we review three aims we sought to address in the design of reading comprehension assessments, and discuss how SBA techniques can be thought of as possibilities for optimizing assessment designs to meet these aims. Specifically, we describe why we see SBA as a design methodology for: a) addressing changes in the nature of the construct or reading for understanding, b) incorporating advances in reading and learning sciences, and c) enhancing features that support instructional relevance. We follow with a discussion of how we arrayed and adapted SBA elements to construct a developmentally sensitive system of vertically scaled SBA forms that span grades 3-12.

Traditional Tests and the Performance Assessment Movement

Prior to the release of the Common Core Sate Standards in the U.S. (NGA & CCSSO, 2010), traditional reading comprehension tests had been widely criticized for failing to be

transparent about how the cognitive and learning science literature was incorporated into assessment designs (Chudowsky, Glaser, & Pellegrino, 2001; Mislevy, 2008; Mislevy & Haertel, 2006). These criticisms included: the lack of explicit connection between theoretical models of reading and the assessment design, the use of artificial and narrow passages, an over reliance on multiple-choice format, the omission of digital texts and multimedia, weak links to instruction and the lack of diagnostic information, too narrow a focus on the product of comprehension rather than the process of how it unfolds over time, and the failure to control for individual differences such as student motivation and background knowledge (August, Francis, Hsu, & Snow, 2006; Coiro, 2012; Magliano, Millis, Ozuru, & McNamara, 2007; National Research Council, 2000; Perfetti & Adlof, 2012; Rupp, Ferne, & Choi, 2006; Sabatini, Albro, & O'Reilly, 2012).

Early attempts at opening up the design space in the U.S., such as performance/portfolio assessments, were met with significant critical commentary concerning construct coverage, objectivity, consistency of scoring, cost-effectiveness, and time-efficiency (Gearhart & Herman, 1998; Koretz, Stecher, Klein, & McCaffrey, 1994). Thus, while other countries may have made advances in this regard, the feasibility and utility of performance assessments were called into question and design innovations were stymied in the U.S., where many of the original criticisms of this test paradigm remain unaddressed.

In recent years, advances in technology and in measurement techniques have challenged conventional notions about what is feasible and useful in large-scale assessments. These include the migration of much of the construct domain to digital forms and the availability and sophistication of technology-based delivery and scoring platforms. These advances enabled the construction of scenario-based reading comprehension assessments (Bennett, 2015; O'Reilly &

Sabatini, 2013; Sabatini et al., 2012). Before we provide details of our approach to SBA design, we begin with a review of the changes to the construct that provides the rationale for rethinking how we assess reading comprehension.

A 21st Century Construct of Reading Comprehension

The construct of reading comprehension, as measured in some traditional tests, has a strong focus on understanding the content of single source texts (one at a time) from printed materials such as books. In the 21st century, the landscape has shifted to an entire universe of Internet documents and other communications, published in all forms of media, from printed documents to texts on tablets, smart phones, and computer screens (Leu et al., 2017). The sheer volume of digital sources has raised the priority of strategic, goal-directed reading skills. Modern readers often need to construct a mental model that integrates information or resolves discrepancies across multiple sources. They may need to evaluate the importance, relevance, accuracy, or truthfulness of each source, and allocate their attention according to complex purposes (Magliano et al., 2018; Rouet & Britt, 2011). These types of multiple source comprehension skills differ somewhat relative to the skills required in traditional print reading (Leu et al., 2017) While making sense of stand-alone texts is still a primary component of reading proficiency, it understates the complexity of the construct, especially as one considers the literacy skills needed for college and career readiness. Thus, to be proficient in reading, individuals must be able to access multiple sources of text and related materials, often in digital formats, and integrate and evaluate what they read (Sabatini, O'Reilly, Wang, & Dreier, 2018). Further, the act of reading has become increasingly social, as individuals interact in social media contexts which require perspective taking skills (LaRusso et al., 2016).

Another important facet to consider in the assessment design is disciplinary reading (Lee & Spratley, 2010). That is, the manner in which an individual models and reasons about content can vary across disciplines (e.g., Shanahan & Shanahan, 2008). For example, in the context of reading science-related content, one reasons through representations, models, and principles to synthesize relationships and draw conclusions from empirical data. On the other hand, when reading history-related content, one evaluates facts and interpretations, the quality of sources (e.g., primary vs. secondary), corroborates evidence, and evaluates the context in which information was collected. Some general skills and strategies are likely transferable (e.g., locate information). However, one might also expect to identify specific skills and reasoning that are differentially called upon when learning in a discipline (Goldman et al., 2016; O'Reilly, Weeks, Sabatini, Halderman, & Steinberg, 2014).

These considerations raise questions about the new construct features that might deserve attention in redesigned assessments – an issue addressed in the U.S. Common Core State Standards (NGA & CCSSO, 2010). The panel of educators and researchers who developed the standards highlighted features that aligned with a more modern conception of reading such as the central role of content and disciplinary literacy in reading comprehension. In fact, they created standards that are specific for reading in history/social studies and in science/technical subjects. The authors also emphasized the need for cross-disciplinary literacy and the need to comprehend, evaluate, and synthesize multiple texts. In addition, technology and the use of multimedia are encouraged in the standards, as well as scientific inquiry as reflected in the focus on research skills. Perspective taking and an awareness of different cultures is also emphasized. See the Common Core State Standards Initiative (2018) for more information on the general design considerations and the specific standards themselves.

The standards subsequently led to the development of innovative assessments in the U.S. as a part of the Smarter Balanced Assessment Consortium (SBAC, 2018) and the Partnership for Assessment of Readiness for College and Careers (PARCC, 2018). For instance, SBAC developed innovative tasks that draw on many of the skills mentioned above. For example, students might exercise their research and inquiry skills by identifying credible sources that would be useful for gathering information that is relevant to their goal or identify which source can support a particular claim (SBAC, 2018). In some cases, the items are a part of a performance task that contains multiple tasks that are interrelated. Similarly, the PARCC assessments include innovative item types that require a student to write about a theme that is similar across multiple texts or to order a set of statements that effectively summarize a passage (PARCC, 2018). In short, the assessments developed in both consortia represent a major advancement from traditional reading comprehension measures.

Building upon this and other work, we identified several theory and research-driven targets based in the literature (O'Reilly & Sabatini, 2013; Sabatini, O'Reilly & Deane, 2013). Briefly, we see a need to incorporate purpose-driven or goal-directed comprehension (McCrudden & Schraw, 2007; Van den Broek, Lorch, Linderholm, & Gustafson, 2001), multiple-text comprehension (Braasch, Braten & McCrudden, 2018; Britt & Rouet, 2012), disciplinary and content area reading (Lee & Spratley, 2010; Shanahan & Shanahan, 2008), digital literacy, online reading or reading in technological environments (Coiro, 2009, 2012; Leu et al., 2017), and social interaction including collaboration, communication, and perspective taking (LaRusso et al., 2016; NGA & CCSSO, 2010; Partnership for 21st Century Skills, 2008). While some of these elements may have been addressed in traditional tests, a growing body of

theoretical and empirical literature is helping to define and elaborate their significance in modern reading contexts.

Incorporating Learning Science Research into Test Designs

The rationale for incorporating learning science research into test designs is supported in a range of recent reports like the Common Core State Standards (NGA & CCSSO, 2010), frameworks for international assessments of reading such as PISA (Organisation for Economic Co-operation and Development, 2009b), PIAAC (Organisation for Economic Co-operation and Development, 2009a), PIRLS (Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009), ePIRLS (International Association for the Evaluation of Educational Achievement, 2013); and various publications on assessment reform (e.g. Chudowsky et al., 2001). The movement also includes other progressive frameworks and standards such as the Partnership for 21st Century Skills (2008), panels and commissions on assessment reform (Gordon Commission, 2013), assessment reform initiatives at major testing companies (Bennett, 2011, 2015; O'Reilly & Sheehan, 2009). Collectively, these efforts call for a new generation of assessments that reflect a broader conceptualization of the construct that goes beyond what traditional assessments have been designed to measure.

Adapting to a 21st century construct of reading is a good starting point for developing a defensible SBA, but lessons drawn from the learning science literature compel us to seek also to provide useful information for instruction (Gordon Commission, 2013). For example, there are anecdotal stories about teachers who see the need to interrupt their ongoing curriculum and instruction to prepare for summative assessments. If so, attention should be given to developing tests that are aligned with, and supportive of, strong instructional practices, such that teachers

would not need to interrupt their regularly scheduled curricula, but rather implement it with confidence.

Enhancing instructional relevance through the use of scenario-based assessment has been researched in the Cognitively Based Assessment of, for, and as Learning (CBAL) initiative (Bennett, 2011). CBAL is a research initiative that has focused on assessment in K-12 settings in English Language Arts (ELA), mathematics, and science. The CBAL ELA competency and key practice models (and associated learning progressions) are based on syntheses of the literature of reading, writing, thinking, and their connections (e.g., O'Reilly, Deane, & Sabatini, 2015). A key goal of CBAL has also been to integrate the research from learning sciences to make assessments meaningful for instruction. Multiple prototype ELA summative and formative assessments have been developed and evaluated; thus, building interpretive and validity arguments for their value and utility (e.g., Bennett, 2011).[1] With respect to the research reported here, our impetus was a grant (see acknowledgment section for details of this grant) awarded by the Institute for Education Sciences in 2010, whose aims required that we build theory-driven, developmentally sensitive assessments that span kindergarten through 12th grade. This mandate required not only innovative techniques at the individual-test level, but also assessments that could work together to map the trajectory of student's progress from beginning readers through college and career-ready proficiency.

Designing for Tracking Comprehension Development Across Years

[1] The CBAL initiative and Institute for Education Sciences grant had overlapping aims. It is beyond the scope of this article to review relevant CBAL publications. The interested reader is referred to the ETS website (https://www.ets.org/research/topics/) for a more complete bibliography.

A developmental model of reading must take into account how children grow to handle the complexity of text and tasks to achieve complex purposes. This complexity increases exponentially from kindergarten through 12th grade, in concert with academic and social reading expectations. Students are asked to read greater quantities of increasingly linguistically sophisticated texts across diverse disciplinary domains, and to perform a wider variety of thinking and learning tasks (Goldman et al., 2016; NGA & CCSSO, 2010). Individuals are only able to do so because many reading subprocesses have become habitual, routine, and can be orchestrated in such a way as to permit challenging new texts and tasks to be undertaken at each new developmental period.

Unfortunately, the research community has not reached consensus on comprehensive models of reading comprehension development that span beginning reading to proficiency (Perfetti & Stafura, 2014). In the United States, the closest that one comes to a model of reading development is the Common Core K-12 reading standards (NGA & CCSSO, 2010). The Common Core authors applied their expertise to synthesize existing curriculum standards and benchmark reasoning skills, then spread these across the age/grade span.

For our system design to be developmentally sensitive, we reasoned backwards from the endpoint of proficiency - college and career readiness. In this regard, the Common Core provides a useful heuristic. That is, the purpose of K-12 education is to prepare students for the higher-level reading and thinking that is necessary to succeed in college or in the workforce and society in general. This endpoint in itself is still ill-defined, and most likely spans a range of proficiencies. However, it is a target amenable to empirical investigation that hopefully will continue to be clarified over time.

We targeted three strands of developmental expectations as guiding principles in our SBA designs. First, social and conceptual reasoning skills will develop across age (e.g., Metzger, 2007). For example, the maturity of students will determine when students might be expected to understand or reason about mature content or themes, and these expectations themselves are conditioned on societal norms that vary across countries. Second, the linguistic complexity[2] and variety (genres) of texts that students are expected to read changes across years. Third, the sophistication of the tasks that students are expected to perform also changes (Ozuru, Rowe, O'Reilly, & McNamara, 2008; Rouet, 2006). More advanced readers might be expected to do more complex tasks with source texts. This may involve writing an argument, corroborating information across texts, detecting and correcting errors in a source, or integrating all of these processes iteratively towards achieving some broader purpose. In the series of SBA forms that we developed to span the grade levels, we incrementally increased the social and conceptual reasoning demands, as well as the linguistic complexity of texts, and the sophistication of tasks and responses required of students, aligned with learning science results that emphasize sets of skills at different developmental levels (e.g., Sabatini et al., 2013).

Scenario-based Assessment as Instantiated in the GISA system

In order to accommodate and integrate these elements, we chose to move beyond the constraints of the traditional passage – question format of some reading comprehension tests. Below, we discuss how a scenario-based design provides an architecture that increases the

[2] Linguistic complexity is used here to refer to the linguistic demands such as text cohesion, syntactic complexity, and the level of vocabulary sophistication that may impact a reader's ability to from a coherent model of the text (see McNamara, Graesser, & Louwerse, 2012). Linguistic complexity may contribute to item difficulty independent of the task and content demands.

degrees of freedom necessary to accomplish this integration at the individual assessment level. SBAs may not be the only solution to this problem, but we have found them to be robust exemplars for enhancing designs.

To better understand the architecture of the system, it is important to lay out a set of constraints that framed the development of the GISA. Most of these constraints were imposed by the design team for the purpose of providing feasibility of implementation and scoring, scalability, and maintaining psychometric rigor. First, we chose to make the entire system web-administered. This had multiple benefits including: remote recruitment and implementation, ease of administration, data collection, scoring, the implementation of complex, randomized designs within and across schools, and a natural environment for using digital sources. Second, we limited the test length to around 45-50 minutes, a typical classroom period in the U.S. This limited duration made it easier for us to recruit schools and collect student data. In some of our studies, we had students complete a pair of test forms in two sessions.

Third, we limited the use of constructed response (CR) items to questions we believed could be scored using automated processes. We primarily focused on paraphrase, summary, and some short-answer explanations. These item types are important cognitive reading strategies (hence, worth teaching to by instructors), as well as rich sources of comprehension evidence. This approach served several aims simultaneously: less student time on individual CR items (which are often also effort intensive) and amenable to automated scoring. Prior research shows that asking students to write summaries (always of key texts in the scenario) increased the engagement of students in closely reading the source text, which we argue increases validity of the score as a measure of reading comprehension (O'Reilly, Feng, Sabatini, Wang, & Gorin, in press; Wang, Sabatini, O'Reilly, & Feng, 2017). Space precludes discussing other important

external and internal constraints, but these three highlight how the test designers planned not only in how to make this feasible for use in real school environments, but also how to facilitate collecting quality data for validation during the project.

The design of the GISA system was guided by a three-part, reading for understanding assessment framework. The first two parts provide the foundational research principles and interpretive argument for a theory-based construct of reading for understanding (Sabatini & O'Reilly, 2013; Sabatini et al., 2013). The third part of the framework (O'Reilly & Sabatini, 2013) is most relevant here, as it described the notion of performance moderators and some of the key features of scenario-based assessment, including:

- Providing a purpose for reading: establishing a standard of coherence.
- Promoting coherence among a collection of materials: the assessment narrative.
- Gaining more information about test takers: triangulating strengths and weaknesses.
- Promoting collaboration: distributed and collective understanding.
- Simulating valid literacy contexts of use/practice: assess what we want students to be able to do.
- Promoting interest, motivation, and engagement.

The instantiation of these features is briefly summarized below.

In some traditional reading assessments, test takers are presented with a collection of unrelated passages on a range of general topics. Students answer a set of discrete items on each passage and then move on to another unrelated passage. In this traditional design, students are effectively expected (or allowed) to forget what they read previously when answering questions on later passages. In other words, there is no overarching purpose for reading other than to answer discrete, multiple choice questions (Rupp et al., 2006). In contrast to this approach, with

GISA, students are given an overarching purpose for reading a collection of thematically related sources for the purposes of solving problems, making decisions, or completing a higher level task (e.g., making a presentation; editing a wiki). The reading purpose introduces a set of goals, learning aims, or criteria that students use to evaluate sources, or decide what information is relevant.

The collection of sources is always diverse and may include a selection from a book, e-mails, blogs, websites, policy documents, primary historical documents, and so forth. Students are asked a series of questions about the sources ranging from traditional comprehension items (e.g., identify key information, draw basic inferences) to more complex tasks such as the synthesis and integration of multiple texts, perspective taking, evaluating web search results, completing graphic organizers, using a rubric to score given responses, or applying what they read to a new situation or context.

Tasks and activities in a scenario are sequenced to reveal the parts of a more complex task that students can or cannot do. For instance, if a student has trouble writing a summary—thus limiting the evidence of his or her skills—other tasks are provided to determine whether the student can evaluate a given summary, recognize a good summary, complete a graphic organizer, or identify key ideas. Such a collection of graded tasks helps provide an evidence trail that can be used to infer the complexity of tasks a particular student can or cannot handle. In this way, complex tasks are not viewed as an "all or nothing activity," but rather as a way to help triangulate partial student knowledge in the larger context of skill development. Simulated "peer" students are also included into the assessment design to provide guidance—hints—and to serve as a way to identify student misconceptions or errors in understanding. For instance, a

simulated peer may provide an incorrect explanation of a process described in a text and the test taker's task is to identify and correct the error.

As noted, the designs were informed by what we refer to as performance moderators, variables that are not directly considered a part of the reading construct, but may impact the reading process. For example, techniques were incorporated into the test designs to provide more information about relevant student background knowledge on the topic of assessment. For instance, if a measure of background knowledge indicates the student knew a lot about the topic, then the comprehension score could be qualified as possibly reflecting more about the student's knowledge level than the individual's reading ability. To mediate some of the differences in prior knowledge, a commonly used technique was to build up relevant background knowledge incrementally over the course of the assessment, so that students can engage in deeper tasks towards the end of the assessment (e.g., make a decision, apply what was learned to a new situation). Other performance moderators are included in the test design such as engagement/motivation, metacognition and self-regulation, as well as reading strategies, to model and encourage good practice.

While we have not fully worked out how these performance moderators would precisely function in a high stakes testing environment, we believe they are worth investigating so that future research can evaluate their potential added value. For instance, we are currently exploring designs that may allow teachers to track whether student learning is taking place over the course of the assessment. In such designs, background knowledge items are presented before and after the student reads content that provides answers to the questions. This design may help uncover whether students have initial misconceptions about a concept and whether they can overcome these misconceptions after reading. This knowledge revision process has been given recent

attention in the literature, as is relevant to background knowledge, metacognition, and self-regulated learning (Kendeou & O'Brien, 2014).

In short, SBA designs model and reflect the way an individual might interact and use literacy source materials when learning from text or making decisions in or outside of school settings; in contrast to the discrete passage paradigm of traditional reading comprehension tests. SBA presents real problems and issues for students to solve and it involves the use of higher level reading and reasoning skills that are demanded in several current content and assessment initiatives. Despite these more demanding goals, the assessment also presents students an opportunity to evaluate and develop their skills, as complex tasks are broken down into more manageable subtasks, while empirically supported instructional practices are incorporated into the design. In this way, the assessment design is intended to support learning and instruction.

Designing the System for Developmental Sensitivity Across Grades

In developing the assessment framework for the system, we identified a set of new construct facets that we viewed as theoretically important for expanding the reading comprehension construct to satisfy the demands of 21st century literacy. As noted, however, one challenge faced by the design team was the absence of a comprehensive, K12 reading comprehension developmental model. This left the team with decisions about how to adapt these facets of the construct across different developmental levels. Next we describe how we went about these adaptations for several of these key constructs facets.

Overarching purpose/scenario – All SBAs provide students with a purpose for reading sources. The purpose helps establish what is, and is not important to attend to, as well as a final product or outcome for reading. For very young students, the purpose may be to read a story together with a teacher about animal oddities (Do chickens take dust baths?). Middle grades

students may be asked to work with peers to present information to the public on a website (e.g., organic farming). Secondary students might be asked to work together as a study group to learn about a complex topic (e.g., mitochondrial DNA or mtDNA, which is distinct genetic information inside mitochondria in human cells, is inherited from one's biological mother, and is used to trace ancient lineages) to prepare for an essay or a test. In this way, the purpose/scenario was adapted to typical academic literacy practices across different developmental age/grade bands.

Topical background knowledge is measured to determine its impact on the reading score and to measure learning. Items range from sorting topical vocabulary/knowledge related to content into categories or identifying specific facts or concepts that will come up in content. It is not inappropriate for this knowledge to prime relevance processing during the scenario, and answers are sometimes shared with students, if knowledge would not interfere with subsequent test performance, but developmental interactions with item properties are monitored (e.g., Sabatini, Halderman, O'Reilly, & Weeks, 2016). Some background items may be asked again at the end of assessment, to see if students have learned the content after reading. As stated earlier, this can also be used to examine student misconceptions or knowledge revision processes (Kendeou & O'Brien, 2014).

Modeling text content (and strategy use) – A wide range of strategies that have been shown to be effective learning tools for students (McNamara, 2007) are used to gather evidence of students understanding and modeling of text content. These are arrayed across the development range to moderate cognitive complexity. For example, young learners may fill in partially completed graphic organizers, elementary students may be asked to paraphrase key

content, while more mature learners may be required to draft written summaries or explanations of content.

Disciplinary literacy– at every developmental level we employ literary, science, history, and general topics at the core of the scenario. However, within and across scenarios, the mix of genres vary (e.g., a literary focus may also present historically relevant texts; science topics often include policy texts). At higher development levels, items that focus more on the distinct kinds of disciplinary processes are increased (e.g., using primary sources in history; evaluating data in science).

Though space precludes a thorough treatment, multiple source processing, use of digital text format, conceptual reasoning, and social modeling and reasoning are similarly interspersed at every level of development, taking into account the developmental literature on what types of maturity and complexity of content and skills students at different ages are likely to be able to attempt (Sabatini et al., 2013).

A note on text complexity across ages – While we agree that texts should not be so challenging as to frustrate or overwhelm typical students, we chose to present a wide range of text challenge levels within each assessment (see also NGA & CCSSO, 2010). Why? First, at every level there are advanced/precocious students, and we want them to face texts that challenge their text comprehension ability. Second, it is important to understand how students approach challenging text strategically, in light of purpose and supports we provide. Often the purpose or the peers provide guidance in what strategies to use in understanding or identifying the relevant information in the more challenging texts (which we tend to present in briefer chunks). We expect that as students grow, they will often face challenging texts and must

develop strategies to deal with the challenges. Third, we structure and sequence the scenario such that background knowledge for understanding complex texts is built up earlier in foundational texts as students are asked to provide summaries, build organizers, and so forth. Sometimes the peers provide support in understanding complex texts by reviewing earlier information, or in directing attention to relevant information.

The approach is also related to the level of engagement and motivation. We expect students to have a level of persistence and effort, and such challenging texts require students to demonstrate their willingness to put forward the effort required of a proficient reader. We evaluate effort using several process data sources including timing information, as well as response choices.

As for simple texts, there is no lower limit. For example, one of our secondary school texts is based on a Langston Hughes story that is written at a 4th grade level. The themes and tasks are complex, but the linguistic complexity of the prose is simple. This freedom also allows us to use longer texts without overwhelming students' processing capacity in relatively short sessions (less than 50 minutes on average).

Evidence and Results

In this section, we review briefly some of the evidence that supports the GISA system. We have collected a large amount of data—around 100,000 administrations across dozens of studies. Nearly every form was pilot tested with 100 or more students, often across one to three adjacent grade levels. This was done to more carefully evaluate the distribution of scores at each grade level, as well as the difficulty of the forms, prior to conducting our large-scale data collection. As part of the pilot study analyses, we examined classical item statistics, checked for

floor/ceiling effects, and examined score reliabilities and timing information (completion mostly took times less than 50 minutes on average). The results were used to revise or remove items and tasks that did not function as expected. For all piloted forms, we found that the majority of students were able to complete the tasks within the allotted time; they were also able to adapt to the somewhat novel SBA designs without any separate tutorial or training by teachers or administrators. We credit our assessment development team for the success we had with this design out of the gate, so to speak, and to the previous work implementing SBAs by the CBAL group (Bennett, 2011).

We created over 20 operational GISA forms spanning the K-12 ability range. Nineteen of these forms were administered in a field study of around 12,000 students across grades 3-12. Building on the model of pairwise administrations for the pilot studies, we implemented a scaling design where three forms were administered at each grade level. Two randomly equivalent groups of students received the forms in a counterbalanced design with one form serving as a common form for both groups. This common form was also administered to a single group at the subsequent grade level. This design allowed for the evaluation of items both within an equivalent groups design (at each grade level), as well as via a nonequivalent-groups, common item (form) design across grade levels (Kolen & Brennan, 2014). The across-grade design established the linkages for creating a vertical scale that can be used to compare scores across grades.

IRT Scaling

In order to compare scores across test forms (within and between grades) it is important that they be reported on a common scale. Item response theory (IRT; Lord & Novick, 1968) is commonly used for this purpose. In contrast to classical methods which essentially aggregate

scored responses, IRT is a probabilistic approach that relies on the pattern of item responses and item characteristics to obtain estimates of examinee ability. The item parameters, across forms, were calibrated concurrently via a multi-group extension of the 2PL/GPCM (Bock & Zimowski, 1997) using marginal maximum likelihood estimation. The software program MDLTM (von Davier, 2006) was used to estimate the item parameters (and subsequently, examinee abilities). For this estimation approach, examinee groups were defined grade-level and form pairing. To place all of the forms on a common scale, item parameters for the common forms (i.e., the forms administered in two grades) were constrained to be equal across groups. The model was identified by constraining the mean item difficulty and slope to be zero.

After the initial scaling was complete, the results were evaluated to ensure estimation convergence and to identify potentially problematic items. Two items were flagged for review and were ultimately excluded from the scaling. The final set of items was recalibrated using the multigroup concurrent calibration with model constraints described above. The end result of this calibration was the creation of a unidimensional vertical scale spanning grades 3 through 12. Expected *a posteriori* (EAP) examinee abilities were estimated based on the final item parameter estimates.

Dimensionality Analysis

Prior to conducting the final IRT scaling, we examined the dimensional structure of the tests, across grades. When developing a vertical scale it is commonly assumed that the underlying construct is unidimensional, both within and across groups. However, given the novelty of the GISA SBA approach, unidimensionality is not assured. As an extension of the unidimensional IRT scaling approach described above, the data were examined using exploratory and confirmatory factor analytic methods, within a multidimensional IRT framework (Reckase,

2009). This approach was patterned after a previous examination of GISA (O'Reilly et al., 2014). We fit three models: a unidimensional model, a two-factor exploratory model, and a two-factor confirmatory, simple-structure model where items associated with science passages loaded on one factor and items associated with history/language arts passages loaded on the second factor. The correlations between factors under the exploratory and confirmatory models are around r = .83 and r =.69 respectively. An examination of the item slopes (loadings) suggest the possibility of a science-related factor; however, the unidimensional model fits better than either of the multidimensional models on the basis of the BIC (see Table 1). The AIC is slightly higher for the unidimensional model relative to the multidimensional models. Note that lower AIC and BIC values indicate better model fit. In short, the construct measured by the GISA across grades appears to be essentially unidimensional. It is on this basis that the unidimensional vertical scale was created.

Differential Item Functioning

As part of our analyses, we also checked the forms for gender and race/ethnicity differential item functioning (DIF). The criteria for assessing the presence of DIF was based on Dorans and Kulick (2006) and has three levels based on values of the Maentel-Haenszel chi-square statistic. There were very few items that exhibited significant DIF. Only 11 items, out of over 700, were excluded for DIF; seven were excluded for gender DIF and four were excluded for race/ethnicity DIF. After all the exclusions, the scores in final forms showed reliabilities in the range of .80 to .88.

External Validity Evidence

Given the novelty of the SBA format, the changes we claim in the construct, and the topical, thematic focus of each form, we were cognizant that critics would be concerned with

whether the resulting scores still captured significant variance associated with traditional definitions of reading comprehension; or rather was each GISA form a construct unto itself. And, perhaps they might be concerned that the forms were too advanced for general or struggling readers, as we make claims about evaluating students' deep comprehension processes. The psychometric evidence from the item analysis and IRT scaling is one source of evidence that supports that claim that we are measuring essentially a single construct, not 19 separate, form-dependent constructs. Further, the relative normality of our score distributions (with no floor or ceiling effects) in samples aimed at the grade span for each GISA form is evidence that the difficulty of the forms is pitched at an appropriate level, especially in the low stakes administrations where one might expect that motivation and effort are more variable across samples. Still, this does not fully satisfy the question of whether the variance captured in the assessment scores overlap significantly with the construct of reading comprehension[3], as measured in more traditional reading comprehension measures.

To address this issue, we conducted a follow-up study, where we administered multiple GISA forms across grades 4-8 in the spring of the school year. These students also took both the Gates-MacGinitie Reading Tests (GMRT; MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2000) and GISA generally within a 2-3 week time span, and the RISE battery Reading Comprehension (RC) subtest (see Sabatini et al., 2015 for construct and subtest properties) earlier, in the fall of the school year. The GMRT is frequently used in the research community as a standard reading comprehension test. The RISE RC also is a traditional passage-question style comprehension test. It has been shown to correlate with state tests at about .6, which is consistent

[3] Reading comprehension is not directly observable, but the quality of a student's mental model can inferred from the evidence gathered through a sample items and tasks.

with the expected correlation among reading comprehension tests across publishers (e.g., Keenan, Betjemann, & Olson, 2008). The correlation between GMRT and GISA scores was r = .80; the correlation between RISE and GISA scores was r = .65; the correlation between GMRT and RISE RC scores was r = .77. These correlations fall within the expected range from previous literature (e.g., Keenan et al., 2008). The high, positive correlation between the GISA and Gates (and RISE RC) provides concurrent validity evidence for the GISA, that is, the scores are capturing variance associated with the traditional measures of reading comprehension, though there remain differences in the variance explained. Future research will be needed to explore the sources of those differences.

We argue there is added value in using the GISA over a traditional passage-question format, because the design, texts, and tasks resemble the types of reading that occur in more modern, 21st century reading environments, and are aligned with learning science evidence that supports quality instruction. One might expect that the rank order of students based on comprehension ability is not changed by giving different types of tests that call upon their comprehension ability, hence, the GISA and traditional measures should show evidence of concurrent validity coefficients. However, the features embedded in the GISA SBAs vary from one form to another, as do the different mixes of cognitive resources necessary to perform successfully on tasks. It is also our intent that by embedding the new features described, as well as reading strategies, metacognitive, and self-regulatory models and affordances, we would model and encourage good habits of mind for teachers and students alike. We hope that this, in turn, supports learning and instructional practices that yield growth in comprehension that would positively impact future student GISA performance. However, we admit that the evidence of

hypothesized instructional impacts awaits future implementation studies of the GISA forms in school settings.

Conclusions and Future Directions

This paper focused on how to create a test architecture that accommodates simultaneous goals of: applying 21st century reading theories and models, incorporating learning sciences research, enhancing instructional relevance, and ensuring developmental sensitivity, while maintaining feasibility and adequate psychometric properties. We were able to use SBA techniques to address multiple construct facets of our assessment framework for enhancing reading comprehension measurement. These construct facets include purpose-driven reading, understanding digital sources and genres, multiple source evaluation and integration, incorporation of reading strategies, the employment of more disciplinary literacy content and processes, and enhancing the use of social reasoning and modeling, including perspective taking and evaluating different points of view. We also measured students' background knowledge relevant to the topic of the assessment to both contextualize comprehension score performance, and potentially track students' conceptual change and knowledge revision over the course of the assessment.

We chose to move beyond the constraints of the traditional passage – question format of some reading comprehension tests to advance the design of reading assessments to meet the technological needs students face today and to better align assessments with current learning approaches. While it might be possible to incorporate these features in the traditional passage-question format of reading comprehension tests, we believe the SBA design architecture represents a promising approach to incorporating them in feasible, yet rigorous forms. Indeed, PARCC and SBAC assessments have shown progress in including several of these new features.

To address the goal of developmental sensitivity across the assessment system, we adopted strategies to adapt the new construct facets, as appropriate, at different age/grade levels. One element of our approach was to build on our early successes with item and task types, using them as exemplary models, then performing our own *perspective taking* exercise – thinking about the literacy and instructional environments we would find at different grade levels. We used various sources to inform this reasoning process, including theoretical and empirical studies, but also curricular and instructional reviews. In future work, we will explore how well the strands of these constructs are holding up across grades, though that will require more complex analyses or new data collections and studies.

Taken together, the results of this study suggest that the GISA does have defensible psychometric properties, and further, that the scores across grades can be compared along a vertical scale. In other words, from an internal validity standpoint, scenario-based assessment seems to be a feasible way to measure reading ability. In order for a test to be useable, it must have solid internal properties. However, given the large number of features that were novel in the design (e.g., expanded construct, item types, simulated peers), plus the lack of professional support to prepare students, we also found it encouraging that the test scores were correlated to scores on traditional tests – one indicator of external validity.

That being said, the evidence provided here speaks mostly to the internal validity of the GISA. The next step is to consider the external validity of the assessment more robustly, such as correlations to state tests, as well as the practical utility of using this type of measure as a supplement to, or in place of, more traditional reading comprehension measures. We also hope to learn from future research that explores other variants of SBA, both in reading and in other domains.

References

August, D., Francis, D. J., Hsu, H. Y. A., & Snow, C. E. (2006). Assessing reading comprehension in bilinguals. *The Elementary School Journal, 107,* 221-238.

Bennett, R. E. (2011). *CBAL: Results from piloting innovative K–12 assessments* (RR-11-23). Princeton, NJ: Educational Testing Service.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39,* 370-407.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (p. 433-448). New York: Springer-Verlag.

Braasch, J. L. G., Braten, I. & McCrudden, M. T. (Eds.). (2018). *The handbook of multiple source use*. New York: Taylor & Francis/Routledge.

Britt, M., & Rouet, J. F. (2012). Learning with multiple documents: Component skills and their acquisition. In J. R. Kirby & M. J. Lawson (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 276-314). Cambridge, England: Cambridge University Press.

Chudowsky, N., Glaser, R., & Pellegrino, J. W. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Coiro, J. (2009). Rethinking online reading assessment. *Educational Leadership, 66,* 59-63.

Coiro, J. (2012). Understanding dispositions toward reading on the Internet. *Journal of Adolescent & Adult Literacy, 55,* 645-648.

Common Core State Standards Initiative. (2018). *English language arts standards*. Retrived from http://www.corestandards.org/ELA-Literacy/introduction/key-design-consideration/

Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the Mini-Mental State Examination. An application of the Mantel-Haenszel and standardization procedures. *Medical Care, 44*(11), 107-14.

Ercikan, K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. New York, NY: Taylor & Francis.

Gearhart, M., & Herman, J. L. (1998). Portfolio assessment: Whose work is it? Issues in the use of classroom assignments for accountability. *Educational Assessment, 5*, 41-55.

Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., Greenleaf, C., . . . Project READI. (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist, 51*, 219-246. doi:10.1080/00461520.2016.1168741

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

International Association for the Evaluation of Educational Achievement. (2013a). *ePirls online reading 2016.* Retrieved from http://www.iea.nl/fileadmin/user_upload/Studies/PIRLS_2016/ePIRLS_2016_Brochure.pdf

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*, 281-300. doi:10.1080/10888430802132279

Kendeou, P. & O'Brien, E. J. (2014). The knowledge revision component (KReC) framework: Processes and mechanisms. In D. Rapp, & J. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*. Cambridge, MA: MIT Press.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices.* New York: Springer-Verlag.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vemont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*, 5-16.

LaRusso, M., Kim, H. Y., Selman, R., Uccelli, P., Dawson, T., Jones, S., . . . Snow, C. (2016). Contributions of academic language, perspective taking, and complex reasoning to deep reading comprehension. *Journal of Research on Educational Effectiveness, 9*, 201-222.

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

Leu, D. J., Kinzer, C. K., Coiro, J., Castek, J., & Henry, L. A. (2017). New literacies: A dual-level theory of the changing nature of literacy, instruction, and assessment. *Journal of Education, 197*, 1-18.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Welsley Publishing.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2000). Gates-MacGinitie Reading Tests® Fourth Edition. Retrieved from http://www.hmhco.com/hmh-assessments/reading/gmrt

Magliano, J. P., McCrudden, M. T., Rouet, J.-F., & Sabatini, J. (2018). The modern reader: Should changes to how we read affect research and theory? In M. F. Schober, M. A. Britt, & D. N. Rapp (Eds.), *Handbook of discourse processes 2nd ed.* (pp. 342-361). New York: Routledge.

Magliano, J. P., Millis, K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp.107-136). New York, NY: Psychology Press.

McCrudden, M. T., & Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review, 19,* 113-139.

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

McNamara, D. S., Graesser, A., & Louwerse, M. (2012). *Sources of text difficulty: Across genres and grades*. Lanham, MD: R&L Education.

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

Mislevy, R. J. (2008). How cognitive science challenges the educational measurement tradition. *Measurement: Interdisciplinary Research and Perspectives, 6,* 124.

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

Mullis, I. V., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Chestnut Hill, MA: Boston College.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common core state standards initiative: About the standards*. Retrieved from http://www.corestandards.org/about-the-standards/

National Research Council (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC: National Academies Press.

O'Reilly, T., Feng, G., Sabatini, J., Wang, Z., & Gorin, J. (in press). How do people read the passages during a reading comprehension test? The effect of reading purpose on text processing behavior. *Educational Measurement*.

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. (2013). *Reading for understanding: How performance moderators and scenarios impact assessment design* (RR-13-31). Princeton, NJ: Educational Testing Service.

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

O'Reilly, T., Deane, P., & Sabatini, J. (2015). *Building and sharing knowledge key practice: What do you know, what don't you know, what did you learn?* (RR-15-24). Princeton, NJ: Educational Testing Service.

Organisation for Economic Co-operation and Development. (2009a). *PIAAC literacy: A conceptual framework.* Retrieved from http://www.oecd-ilibrary.org/content/workingpaper/220348414075

Organisation for Economic Co-operation and Development. (2009b). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science.* Retrieved from http://www.oecd.org/pisa/pisaproducts/44455820.pdf

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 technical report.* Retrieved from http://www.oecd.org/pisa/data/2015-technical-report/

Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behavior Research Methods, 40,* 1001-1015.

Partnership for 21st Century Skills (2008). *21st century skills and english map.* Retrieved from http://www.p21.org/storage/documents/21st_century_skills_english_map.pdf

Partnership for Assessment of Readiness for College and Careers. (2018). *Grade 11 english lanaguge arts/literacy practice tests.* Retrieved from https://parcctrng.testnav.com

Perfetti, C. A., & Adlof, S. M. (2012). Reading comprehension: A conceptual framework from word meaning to text meaning. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how to assess reading ability* (pp. 3-20). Lanham, MD: Rowman & Littlefield Education.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18,* 22–37.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Rouet, J. F. (2006). *The skills of document use: From text comprehension to web-based learning*. New York, NY: Psychology Press.

Rouet, J. F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In G. Schraw, M. T. McCrudden, & J. P. Magliano (Eds.), *Text relevance and learning from text* (pp. 19-52). Charlotte, NC: Information Age Publishing.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*, 441-474.

Sabatini, J., & O'Reilly, T. (2013). Rationale for a new generation of reading comprehension assessments. In B. Miller, L. E. Cutting, & P. McCardle (Eds.), *Unraveling the behavioral, neurobiological, and genetic components of reading comprehension* (pp. 100-111). Baltimore, Maryland: Paul H. Brookes Publishing Co., Inc.

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

Sabatini, J., Bruce, K., Steinberg, J., & Weeks, J. (2015). *SARA reading components tests, RISE forms: Technical adequacy and test design* (RR-15-32). Princeton, NJ: Educational Testing Service.

Sabatini, J., O'Reilly, T., Halderman, L., & Weeks, J. (2016). Assessing comprehension in kindergarten through third grade. *Topics in Language Disorders*, *36*(4), 334-355.

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

Sabatini, J., O'Reilly, T., Wang, Z., & Dreier, K. (2018). Scenario-based assessment of multiple source use. In J. L. G. Braasch, I. Braten, & M. T. McCrudden (Eds.), *The handbook of multiple source use* (pp. 447-465). New York: Taylor & Francis/Routledge.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review, 78*, 40-59.

Smarter Balanced Assessment Consortium (2018). Sample Items. Retrieved from: http://sampleitems.smarterbalanced.org/

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

von Davier, M. (2006). Multidimensional latent trait modelling (MDLTM) [Software program]. Princeton, NJ: Educational Testing Service.

Wang, Z., Sabatini, J., O'Reilly, T., & Feng, G. (2017). How individual differences interact with task demands in text processing. *Scientific Studies of Reading, 21*, 165-178.