Broadening the Scope of Reading Comprehension using Scenario-based Assessments: Preliminary Findings and Challenges

Sabatini, J., O'Reilly, T., Halderman, L. & Bruce, K. (2014). Broadening the Scope of Reading Comprehension using Scenario-based Assessments: Preliminary Findings and Challenges. *International Journal Topics in Cognitive Psychology*, *114*, 693-723.

Peer Review Process: *International Journal Topics in Cognitive Psychology* is a double-blind peer-reviewed topical journal.

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

Abstract

Existing reading comprehension assessments have been criticized by researchers, educators, and policy makers, especially regarding their coverage, utility, and authenticity. The purpose of the current study was to evaluate a new assessment of reading comprehension that was designed to broaden the construct of reading. In light of these issues, we developed a scenario-based assessment (SBA) of reading comprehension that was inspired by the cognitive literature in reading, learning, and instruction. The SBA was designed to measure students' ability to integrate and evaluate a set of thematically related sources for the purpose of achieving an overarching goal. The SBA also measured students' ability to form an integrated and global mental model of the text, student background knowledge, and social interactions in a digital environment (e.g., perspective taking; classifying and evaluating simulated peer claims). A sample of 426 sixth grade students completed the SBA form and a subsample of 161 students completed a reading component skills battery. Results indicated adequate psychometric properties of the SBA, evidence generally in support of the alignment of the SBA to the assessment design, and strong correlations between the SBA and traditional reading comprehension assessments. While students were able to engage with a variety of complex tasks, tasks that measured students' ability to form a coherent mental model (e.g., write a summary) and digital literacy tasks that required students to integrate perspectives from multiple texts were particularly difficult.

Contemporary reading assessments designed for school and classroom use have not been keeping pace with the changing nature of reading theory (van den Broek, 2012). Most published, standardized reading comprehension tests in the United States continue to be comprised primarily of isolated passages and independent questions about those passages (Rupp Ferne & Choi, 2006; van den Broek, 2012). While multiple research studies have probed different techniques for gathering evidence on the formation and quality of examinee mental representations or models of text (e.g., Trabasso, 2005; Zwaan, 2008; Zwaan, Magliano, & Graesser, 1995; Zwaan & Radvansky, 1998), most traditional tests continue to rely almost exclusively on items explicitly designed to measure specific skills. Similarly, many assessments continue to be published and administered using paper-based forms, which limits how well they integrate digital text genres and responses typical of many of those genres. Finally, the broadest goal or purpose students are given for reading in a standardized test is to answer as many questions correctly as possible (Rupp et al., 2006) - a goal that does not map onto authentic, real-world reading activities.

In fact, conventional passage-question reading assessments have been widely criticized by researchers, educators, and policy makers. Some have been critical of the alignment of tests with contemporary theoretical constructs and empirical findings pertaining to reading processes (Hannon & Daneman, 2001; Magliano, Millis, Ozuru, & McNamara, 2007). Others have noted the lack of tasks requiring cross-text integration (Lawless, Goldman, Gomez, Manning, & Braasch, 2012; Rouet & Britt, 2011; Strømsø, Bråten, & Britt, 2010). Yet other researchers have argued that traditional assessments employ tasks and texts that do not represent the range of purposeful literacy activities of 21st Century students (Partnership for 21st Century Skills, 2004, 2008; Pellegrino, Chudowsky, & Glaser, 2001; Rupp, Ferne, & Choi, 2006).

In this paper, we make a case for broadening the reading comprehension construct, and how it is assessed, in ways that we believe are better aligned with contemporary theories of reading literacy. We then present a prototype scenario-based assessment of reading comprehension that represents an attempt to instantiate some of the features of this broader construct.

Background and Literature

Advances in Theory and Research of Reading and their Influence on Reading Assessment. Traditionally, reading comprehension assessments are comprised of items sampled from a list of behavioral skills such as distinguishing main idea from detail, fact from opinion, drawing conclusions, and so forth. These skill targets are typically derived from a list of curriculum skill standards (e.g., NGA & CCSSO, 2010). The skills may also be derived from empirical reviews of skilled readers (e.g., National Reading Panel, 2000) or more broadly aligned with key reasoning skills such as analysis, synthesis, or evaluation (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956) that are often the focus of classroom instruction. In contrast, technical descriptions of published tests rarely include any reference to theories of reading, cognition, or students' ability to form mental models of text (van den Broek, 2012). Instead, assessment technical manuals list the skills and text types targeted by the assessment. Content validity evidence generally consists of reviews by content experts or teachers, confirming that the specific items align with the skill targeted. At best, one might find reference to empirical studies that show evidence that the targeted skills are associated with proficient reading. But on balance, there is typically no overarching theoretical framework that guides the test design.

We believe the research in cognitive science can contribute to improving the quality of reading comprehension assessments, as we review next. Although the research literature on reader cognition is extensive, we focus on two broad issues that we argue are appropriate foundations for broadening the construct of reading comprehension assessment: the formation of mental models of text sources and the processing of electronic or digital text sources. Within each broad issue, we briefly discuss other pertinent strands of research as they relate to the reading constructs and assessment designs (e.g., background knowledge, task-oriented reading).

Forming mental models of text sources. Most cognitive models and theories of comprehension converge on assuming that the reader constructs mental representations or models while reading (Gentner & Stevens, 1983; Johnson-Laird, 1983; McNamara & Magliano, 2009; van Dijk & Kintsch, 1983). Thirty years ago, Norman (1983) usefully distinguished between the conceptual versus user's mental model. A *conceptual model* is an accurate, consistent, and complete representation of the target domain (such as the content of text sources), constructed typically by teachers, assessment designers, researchers, or other experts. The reader's *mental model* is the representation that evolves from interactions with the source content. Norman observes that user mental models are often incomplete, severely limited, unstable, inconsistent, and parsimonious. Theories of comprehension have evolved to explain and study both the processes of mental model formation and the resulting mental models themselves. Prominent cognitive theories that investigate reader mental model formation include constructionist theory (Graesser, Singer, & Trabasso,1994), the construction-integration model (Kintsch 1998), the structure building framework (Gernsbacher 1997), the event indexing model (Zwaan, Langston, & Graesser 1995; Zwaan & Radvansky 1998), the resonance model (Myers & O'Brien, 1998), the landscape model (van den Broek, Young, Tzeng, & Linderholm 1999), and the minimalist hypothesis (McKoon & Ratcliff, 1992).

In the assessment we present in this paper, we are interested in the reader's understanding of informational texts in relation to a general purpose for reading and with respect to their background knowledge on the topic. For this reason, Kintsch's (1998, 2012) construction-integration (CI) theory is perhaps most relevant. CI theory differentiates between two key representation levels relevant to assessment design: the textbase and the situation model. The textbase includes the propositional structure of the text that preserves its core meaning and structure. The situation model integrates the text meaning that is relevant to the reader's goals with the reader's background knowledge.

We view the concept of reader-formed mental models as having multiple implications for assessment design and score interpretation. First, we do not view individual items merely as evidence of the reader's possession of discrete behavioral skills, but rather as providing evidence of the accuracy, coherence, and completeness of the mental models that the reader forms. Second, as assessment designers, we would want to provide the examinee with a situation or scenario that motivates a broader goal for reading than simply answering discrete questions (Rupp et al., 2006), as well as providing the examinee an opportunity to build adequate, relevant mental models. Third, in doing so, we need to formulate and consider the *conceptual model* that we are expecting the reader to approximate with their mental model. For example, are the task instructions sufficiently clear and specific to direct, yet constrain their responses to relevant aspects of the texts? Can we assume that the examinee has adequate prior knowledge to generate appropriate inferences in forming a rich situation model? In sum, designing reading comprehension assessments that capture evidence of reader mental models of texts, versus evidence of application of discrete skills, signals a foundational shift in the assumptions underlying the design process. In the following sections, we describe several additional theory-based issues that we view as helpful in broadening the reading comprehension construct.

Task-oriented reading. As noted, a situation model is a representation that takes into account reader goals. Recent research into text relevant and task-oriented reading is exploring how reader mental models are impacted by reader goals (McCrudden, Magliano, & Schraw, 2011; Narvaez, van den Broek, & Ruiz, 1999; Rouet, 2006; van den Broek, Lorch, Linderholm, & Gustafson, 2001). Reader goals can be as simple as locating a specific text detail, while other goals require deeper processing (e.g., synthesize multiple perspectives on a topic). Evidence is accumulating that task-orientation impacts comprehension processes and outcomes (McCrudden et al., 2011). Later, we describe how one can use scenario-based assessments to afford opportunities for task relevant reading, thus broadening the construct assessed.

Background knowledge. As noted, a well-formed situation model requires incorporating text meaning with relevant background knowledge (Kintsch, 1998, 2012). Background knowledge provides a structure for integrating text information (Mandler, 1984) and helping students make the necessary inferences to bridge gaps in less cohesive texts (Author, 2007). While various experimenter measures have been developed (e.g., Hannon & Daneman, 2001), to our knowledge, no standardized assessments of reading comprehension take into account the role of background knowledge and how it can affect the formation of mental models, and consequently the interpretation of reading comprehension scores (Shapiro, 2004). Traditional assessment designs attempt to minimize the influence of background knowledge by sampling topics widely. However, they do not directly measure background knowledge or its influence on performance, making it impossible to know whether this strategy is successful. Furthermore, this strategy makes it difficult to design a comprehension measure where the goal is to create a mental model on a set of related texts and tasks. By integrating a measure of background knowledge into the assessment, one can investigate directly how student proficiency might interact with their prior knowledge.

Reading literacy in digital literacy genres. Recent advances in technology have expanded the way people read, communicate, and interact (Coiro, 2009; Leu et al., 2007; Rouet, 2006). The introduction of the internet, the use of multimedia, hyperlinks, and nonlinear text has increased access to a wealth of information. However, to appropriately gather and understand this information, individuals need to be strategic in formulating appropriate search terms, deciding on what search results and hyperlinks are relevant, evaluating the quality of the sources, and must also integrate and synthesize sources into a coherent whole (Afflerbach, & Cho, 2010; Coiro, 2009; Coiro, & Dobler, 2007; Leu et al., 2007; Leu, Kinzer, Coiro, Castek, & Henry, 2013; Graesser et al., 2007; Metzger, 2007). There is some evidence to suggest that there are at least some unique skills involved in online reading. For instance, Coiro (2011) found that scores on an online reading task predicted performance on another web-based task over and above the effects of the student's reading ability as measured by a standardized print-based comprehension assessment.

While the unique contribution of online versus offline reading is beyond the scope of the current paper, the importance of incorporating online reading sources into the broader construct of reading literacy seems beyond question. In fact, three of the most influential international assessments of reading comprehension are currently integrating elements of technology, digital literacy, and online reading into their assessment designs. In the early grades, the Progress in Reading Literacy Study (PIRLS) (Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009) is currently building a new assessment called ePIRLS which is intended to measure online reading (International Association for the Evaluation of Educational Achievement, 2013). Although the assessment is under development at the time of writing, ePIRLS incorporates a virtual online reading environment to assess students' ability to navigate, evaluate, integrate, and synthesize online sources in a non-linear environment.

Similar international efforts have addressed the issue of online reading and digital literacy with older students. Both the Programme for International Student Assessment (PISA) for 15 year olds (Organisation for Economic Co-operation and Development, 2009a) and the Programme for the International Assessment of Adult Competencies (PIAAC) for adults (Organisation for Economic Co-operation and Development, 2009b) include aspects of technology and digital environments in their frameworks. In PISA 2009 (OECD, 2009a), the authors cite the addition of electronic texts as one of the most important modifications in the framework and assessment design. The PISA and PIAAC frameworks provide detailed descriptions of the similarities and differences in reading in print versus electronic media. Each provides a classification scheme for a wide range of electronic texts that include hypermedia, interactive message boards, and visual accompaniments to text. Clearly at the international level, assessment designers are envisioning the next generation of reading assessments with technology and digital literacy as a part of the design.

Despite these advances in large-scale assessments, the majority of off-the-shelf, reading assessments in the U.S. are either paper-based, or they do not include aspects of online reading and digital literacy in the construct. Two aspects of digital literacy environments are targeted in the SBA discussed in the article: multiple source integration and evaluation and evaluating interactive communications.

Multiple source integration and evaluation. Skilled reading includes proficiency in evaluating and synthesizing information across multiple text sources and this requirement is driven by the increasing prevalence of digital literacy activities (Britt & Sommer, 2004; Coiro, 2009, 2011; Goldman, 2004; Leu et al., 2013; Rouet & Britt, 2011). Some have explored the integration of information across texts that may provide conflicting or complementary information (e.g., Bråten, Gil, Strømsø, 2006; Strømsø et al., 2010; Wiley et al., 2009). Of interest here are those who have focused on synthesizing and evaluating the quality of sources (Graesser et al., 2007; Metzger, 2007). For example, Lawless et al. (2012) propose and investigate a model of sourcing that includes evaluation of relevance, author, venue (e.g., publisher), currency, and type (e.g., primary versus secondary source). Drawing from aspects of this work, we include elements of source evaluation in the SBA assessment described later.

Evaluating interactive communications. Web environments not only provide access to multiple sources of information, but also increased communication. The advent of social media has created unprecedented sharing of points of view and perspectives that comprise another variant of multiple text integration. Dynamic communication environments include email or text message dialogues, surveys, blogs, user reviews, chat forums, etc. One challenge of reading in digital environments stems from interpreting the multitude of relatively unfiltered voices that comprise the source content. Interpreting such texts requires mental representation of other minds and social interactions among agents (Dijkstra, Zwaan, Graesser, & Magliano, 1995; Graesser et al., 1994). It also demands questioning the relevance of statements in relation to topic context, as well as evaluating the reliability and truth value of those statements, issues explored elsewhere in the reading research, (Beck, McKeown, Hamilton, & Kucan, 1997; Cain & Oakhill, 2012; Graesser, Ozuru, & Sullins, 2009; Palincsar & Brown, 1984), but now taking on a new dimension in digital contexts.

A theory-driven, scenario-based assessment design. Broadening the construct of reading comprehension to include the issues described above is not easily accommodated in the traditional passage and questions assessment design. Instead, we have built upon existing research on scenario-based assessments (see Author, 2009; Bennett, 2011; Bennett & Gitomer, 2009). SBAs are designed to measure various levels of reading comprehension in a range of reading situations. These situations include a goal for reading and associated tasks, appropriate for students in the targeted grade band. Test takers are provided with a specific purpose for reading (e.g., studying for a history test; preparing for a class presentation) and a set of materials focused on a common topic (e.g., websites, newspapers articles, authoritative texts). Test takers progress through the materials in a structured way that enables them to: provide evidence of complex mental models of text content; learn and organize what they read; and apply and synthesize what they have learned to solve a problem that satisfies their original purpose for reading. In what follows, we discuss the specific design features that were incorporated into an SBA on the topic of organic farming that align with our views on how to broaden the construct of reading comprehension.

To address mental model formation, we chose three task designs typically not used in traditional comprehension assessments. We asked students to write summaries according to guidelines/rubrics that required students to report only main ideas (excluding details), use their own words, and exclude their own opinions and outside information. Summarization and summary writing are strategies shown to enhance comprehension (Bean & Steenwyk, 1984; Hill, 1991; Taylor, 1982) and metacognition (Thiede & Anderson, 2003). Developing a summary parallels the formation of a coherent mental model, because it requires organizing the information of a text according to the key content and text structure.

We also developed graphic organizer (GO) tasks. GOs help readers visualize and understand the organizational structure of a text, aiding in building coherent models of text content (Armbruster, Anderson, & Meyer, 1991; Bean, Singer, & Frazee, 1986; Griffin, Malone, & Kammenui, 1995). Each GO is aligned with the text structure. In the SBA prototype, the more complicated a text, the more complex the GO structure (e.g. a 3 x 4 cell structure vs. a 2 x 2 structure). Each of the GOs is partially completed to help students better understand what they have to do and the nature of the response that is expected.

Finally, we have a set of items that target questioning (Beck et al., 1997; Graesser & Lehman, 2006). In each item, the student is given a list of questions and must select which one can be answered by the text content. The questions range from shallow to deep based on Graesser and Person’s (1994) typology, and students are required to select the correct question, making this item type less demanding than producing their own questions. The question items were designed to tap aspects of the mental model by sampling from key ideas in the texts.

To address task-oriented reading, we created a scenario and used simulated peer interactions throughout the assessment to focus reader's attention on task goals. The examinee reads an opening screen that says:

> "Your class has decided to create a website about organic farming to help members of the community become more familiar with the subject. The website will provide information to answer the following questions: What are the natural methods used in organic farming? How are these methods different from the methods used on non-organic, or conventional, farms? What are the pros and cons of organic farming? You will work with three classmates on the project."

We chose the context of working with a website, given the focus of digital literacy in both research and real use settings (Leu et al., 2013). Throughout the session, screens present conversations in which the simulated teacher and peers interact. The teacher serves as a familiar authoritative agent, introducing subsections of the test and explaining tasks and guidelines, while the peers offer a touch of informality and sympathetic commiseration. More importantly, these peer interactions reinforce instructions, model responses, and focus attention on specific task demands. In some tasks, the test-taker responds to questions and problems posed by the simulated peers.

To address relevant background knowledge, the students begin the test session by making a glossary for the website with words related to farming. We used a technique developed by Deane (2012) and colleagues, in which students are asked to decide whether a term is related or unrelated to the topic of farming. Examinees can choose "I don't know" and the instructions make clear that the section will not count towards their total reading score. In prior studies, the results showed that this task is a quick, but valid indicator of student's prior knowledge of the topic (Author, in press). This section also includes a set of synonym-style vocabulary items relevant to the topic and readings. Taken together, performance on these items can differentiate students who have more prior knowledge about farming versus students that know less about the topic.

Although the scenario is set in the context of digital literacy, the aforementioned tasks do not focus on digital literacy. Two sections of the SBA explicitly do. Presented with the results of a simulated web search, students must evaluate the relevancy, currency, author expertise, and credibility of the sources (Lawless et al., 2012). Later in the assessment, examinees are given the role of moderators of a simulated community discussion forum on the topic of organic farming. Their first task is to classify a series of forum comments as accurate, inaccurate (based on information they learned in the assessment), opinions, or off-topic. Further questions ask students to relate their prior readings to the comments, requiring them to apply the information they learned across multiple sources, within a simulated, digital environment.

Rationale for Current Study and Research Questions

The benefits of properly designed and implemented SBAs and associated innovative task designs include increased construct relevant sources of variance, decreased construct irrelevant variance, alignment to cognitive models and theories of learning, and, importantly, broader construct coverage. The SBA approach, while potentially fruitful, also presents various accompanying challenges. In this study, we investigate the following research questions:

1. Are complex scenario-based assessments feasible in real classroom setting? Can students successfully complete a scenario-based assessment with no prior experience or training?
2. Does the test demonstrate traditional psychometric standards of reliability and item adequacy?
3. Is there concurrent validity evidence in comparison to traditional reading comprehension measures?
4. Is time spent on the sections of the test consistent with the design?

5. Is the pattern of task performance consistent with the design?

Method

Sample

A six school sample of 426 sixth grade students completed the SBA form in the Spring of 2012 with a subsample of 283 retaking the form as seventh graders in Fall of 2012. Another 11 students (9 in Spring, 2 in Fall) started, but did not complete the form. A subsample of 161 students also completed a traditional reading comprehension test and components subtests. Students were participating in a larger research study in a northeastern urban school district in the United States, and through that study informed consent was obtained for all participating students. Based on all available demographics[i], the sample was 47% female; 55% white, 11% African-American, 28% Hispanic/Latino/a, and 6% other. About 55% received free or reduced lunch (an indicator of low income) and about 17% were identified by the school as having limited English proficiency.

Instruments

SBA Organic farming form[1].

Sources. Texts and source stimuli included in this assessment cover descriptive information on techniques used in organic farming such as crop rotations, organic fertilizers, etc.; web search results; pro versus con passages; a web forum discussion of opinions on the topic; and cartoons, charts, and graphic organizers. Passage readability ranged from fourth to nearly ninth grade levels based on the Flesch-Kincaid (F-K) readability formula. We recognize that some of the texts used in this study are challenging, but this is in accord with the U.S. Common Core State Standards (NGA & CCSSO, 2010) which state that students should read both texts that are in and out of grade level.

[1] This assessment was derived from a form initially developed by the ETS *CBAL*™ project, using similar scenario-based principles.

Tasks. Students completed sections of the test that include multiple-choice (MC), constructed-response (CR), and graphic organizer (GO) items. We classify these based on categories of formation of Background Knowledge, Mental Models, Digital Literacy, and Other reading comprehension items.

Background knowledge

- Create a website glossary - Designed to measure background knowledge in two sections. The first section focuses on topical vocabulary (24 MC items) where students decide whether words are related or unrelated to the topic of farming. In the second section, other words related to the topic of farming are presented in a sentence context and students select its synonym (5 MC items). For all items, students can select "I don't know" and the instructions explain that this section would NOT count towards their total reading score.

Mental models

- Learning about Organic Farming - A passage about organic farming is presented and tasks are designed to measure modeling of text content (2 CR summary items worth up to 3 point each), organization (4 GO items), sequencing (3 items), and questioning (3 items).
- The Pros and Cons of Organic Farming - A passage on the pros and cons of organic farming is presented and a GO (4 items) is designed to assess the student’s ability to identify its advantages and disadvantages.

Digital literacy

- Find more information about organic farming – Students are presented with brief descriptions of search results and questions are designed to assess the student’s web source evaluation skills, i.e., is the information relevant and reliable (3 MC items).
- The Community Forum - Comments about organic farming from simulated community members are presented and items are designed to measure student’s ability to classify

comments as correct, incorrect, opinion or off-topic (6 MC items) and to integrate comments with prior learned information (4 items).

Other item types

- Did students learn the vocabulary after reading the texts - Students see the same sentence context vocabulary items seen previously in the background knowledge section, but now after reading passages containing the target words (5 MC items).
- Interpret cartoon in relation to organic farming information (1 MC item).
- Interpret bar chart information (1 MC item).
- Interpret pro-con statement (1 item).

Reading comprehension and component skills. A battery of 45-60-minute, web-based tests that target reading comprehension and component skills was administered. Coefficient alpha (α) reliability in sixth grade students from a previous study Author (2013b) is provided for each subtest. All tests are unspeeded (except Efficiency of Basic Comprehension), however, time limits are set based on prior research showing that 95% of students finish within that time duration. The subtest scales run from 300-400 with an average standard deviation of about 25 per subtest. The subtest scales are benchmarked to a state reading test, such that a scale score of 370 is the benchmark for proficient and 330 is the benchmark for below basic proficiency. Intercorrelations of subtests average in the range of r=.61 to .71. The subtests are:

Word recognition/decoding. Students identify whether a stimulus is a word, decodable non-word, or pseudohomophone (50 items, 6 min., α =0.91).

Vocabulary. Students select either a synonym or a word associated with a target word (38 items, 6 min., α =0.86).

Morphology. Students choose which of three morphologically related words fit in a given sentence (32 items, 7 min., α =0.90).

Sentence processing. Students select the appropriate word that fits into sentences of increasing length and syntactic complexity (26 items, 7 min., α =0.81).

Efficiency of reading for basic comprehension. Students have three minutes to read each of three passages. In each passage, the student must select the appropriate word that fits a sentence context, in a forced-choice task with three word options (36 items, 9 min., α =0.90). The time limit per passage makes this a speeded test, hence, it is labeled as a test of comprehension efficiency. The correct choices are intended to be obvious to even less skilled readers, hence, we label the test 'basic comprehension', to distinguish it from reading comprehension tests that include more complex, higher order questions.

Reading comprehension. Traditional reading comprehension MC questions asked of the same three passages used in the previous subtest (22 items, 20 min., α =0.76).

Other variables in study. In addition to the aforementioned scores, several additional variables are available for analysis in this study. Total time on the following sections of the test was collected and analyzed to better understand how students move through the test and prioritize time. Values greater than 2.5 sd from the section mean were omitted in conducting analyses.

Background knowledge time (BK time). Time spent on the background knowledge section of the test (items=29).

Constructed-response time (CR time). Time spent reading the explanatory passage about organic farming and writing summaries of two subsections of the passage (items=2).

Multiple-choice time (MC time). Time spent completing MC items including GO and other novel item types described above (items=35).

Non-item time (NI time). Time spent reading screens with no response required. These include descriptions of the scenario (one screen), directions (three screens), teacher and peer dialogues (11 screens with text in thought bubbles), and one of the passages (the Pros and Cons of Organic Farming presented first without accompanying questions).

Summary writing scores. Examinees wrote two brief summaries of sections of the organic farming passage. Guidelines directed students to include main ideas, not details, no personal opinions, and to put into their own words. Students could score up to 3 points for a well-formed summary; 2 points was considered an adequate score (i.e., contained at least three of five main ideas elements and met other criteria); while scores of 1 or 0 were given to summaries with significant flaws.

Results

The results are organized to address the research questions posed: 1) the feasibility of administration, 2) psychometric qualities, 3) concurrent validity, 4) timing of the sections, and 5) patterns of student responses for this SBA approach to measuring reading comprehension.

1. Are complex scenario-based assessments feasible in real classroom setting? Can students successfully complete a scenario-based assessment with no prior experience or training? Is there evidence for the feasibility of the SBA design?

Score distribution. The SBA raw score mean (sd) was 18.7 (8.3) out of a possible score of 41, with a range of 5-38. The SBA BK raw score mean was 18.1 (5.0) out of 29 with a range of 3-28. The distribution of scores was normal, with skewness and kurtosis values less than 1.0. Visual inspection of histograms showed no evidence of floor or ceiling effects.

Test length in time. The average time to complete the test was 41 minutes. Eighty-seven percent of students completed the SBA test within 50 minutes; about 95% within 60 minutes; and the remainder took up to 81 minutes. We examined all six schools in the sample, and it appears that most students were given as much time as they needed to complete the test. Of 11 students who did not complete the test, the average items completed was 16 (range 6-26) and the average time spent answering items was 35 minutes (range 16-48). Comparing examinees that took the Fall retest (n=283), the repeaters' session time averaged 12.3 minutes less (42.7 vs. 30.4 for Spring to Fall respectively).

2. Does the test demonstrate traditional psychometric standards of reliability and item adequacy?

Reliability. Internal consistency (alpha) reliability for the 37-item SBA form was α(426) = 0.89. The split half reliability was r(426)=.76, with each half of the test showing adequate alpha reliability (α = 0.80 and α = 0.82 respectively). A subsample of 283 students were administered the same form again at the beginning of the next school year in Fall of 2013 as seventh grade students. Thus, they had a summer break between the retest and consequently little to no new instruction. Test-retest reliability was r(283)=.87 and there was no significant difference in mean scores (Spring M=51.0%, Fall M=50.4%, p=.75). Internal consistency reliability for the 29-items BK section was α = 0.77, split half r(426)=.60.

Item properties. The quality of the items on a test is evaluated by examining item difficulty and discrimination. Item difficulty on a typical test should show a range of difficulty, with most items falling between .20 (difficult) to .80 (easy) on a scale of proportion correct between 0 and 1. The 37 items on the test showed an average difficulty of M=.47 (SD=.16), with a range between M=. 21-.76. Item-total correlations are used to evaluate how much information each item is contributing to the scale. Generally, most items should show values above .20 or preferably .30 to be considered as contributing information to the test. Negative values are undesirable as they indicate that the item is negatively correlated with better performance on the test. All items except two showed correlations above .20, with four items falling between .20 and .30 on the scale.

Average item difficulty for the BK section was *M*=.59 (*SD*=.24) with a range between *M*=.22-.97. Two items showed negative item-total correlations and were omitted from scores and subsequent analyses. With these omitted, reliability increased to .78. Five other items had values below .20 and six others below .30. Note that 25 of the items (including the two with negative values) were 2-option choices with the option of choosing "I don't know". The large guessing parameter likely weakens the information contributed by these individual items.

3. Is there concurrent validity evidence in comparison to traditional reading comprehension measures?

Table 1 shows the means, standard deviations, and correlations between reading comprehension and component subtests and the SBA form. The range of mean reading battery scale scores (356 to 371) positions this sample in the average to slightly above average range of ability overall, but with wide sample variability. The reading comprehension subtest is most comparable to conventional standardized, classroom reading tests. As seen in Table 1, the SBA form shows a strong correlation of r=.77 with reading comprehension. Furthermore, the other subtests show correlations of comparable strength to the SBA and reading comprehension.

Insert Table 1 about here

Table 2 shows the means (SD) of the SBA and reading subtests by quartile ability groups on the SBA. Using the reading comprehension scale benchmark values, examinees in the highest ability group

(Quartile4) would score proficient on the state tests, examinees in the lowest group (Quartile1) would be below basic. The middle quartiles show scores in the basic ability range. Using this metric, the SBA form shows a strong concurrent validity with more conventional reading comprehension tests, as well as component reading subtests.

Insert Table 2 about here

4. Is time spent on the sections of the test consistent with the design?

As the correlation matrix in Table 3 shows, spending more time on a section was positively correlated with higher scores, with the exception of the BK section which was weakly, negatively correlated with overall SBA performance. A multiple regression model predicting total score was run using the time variables as predictors, entered in the order: multiple-choice, constructed response, background knowledge, and non-item time (see Table 4). Each time variable added significant variance over and above the prior model (R^2= .341, .360, .488, .516 respectively for the four sections). The beta weight for BK was negative, suggesting that more time spent on BK negatively related to total score, even though total score on the BK was positively related to performance (r=.51). Spending more time on the BK section was also a small, but significant predictor of BK score (F(1, 638)=31.3, p<001; R^2 =.047), and a significant predictor of SBA total score after controlling for the BK score (F(1, 638)=31.3, p<001; R^2=.316, change in R^2=.012).

Table 5 shows the relative time spent on the parts of the SBA by the different ability groups. As a group, students spent the most time on the screens that required answering the questions that counted towards their total scores (MC and CR items). They spent an average of 2.9 minutes responding to the background knowledge questions. They also spent 5.7 minutes reading non-item screens which are comprised of the scenario description, peer/teacher dialogues, as well the Pro-Con passages themselves, which students are asked to read before any questions are given. With respect to ability groups, higher ability groups consistently spent more time on each part of the test with the exception of the BK section.

Insert Table 3, 4 & 5 about here

5. Is the pattern of task performance consistent with the design?

Following the review in the introduction, we classify items as related to reader *mental models* (4 tasks, 16 items), *digital literacy* (3 tasks, 13 items), and *other* (8 items) reading items. Table 6 shows means (SD), reliability, and correlations for these three item categories. The tasks associated with forming a mental model of the text were on average more difficult than the tasks associated with evaluating web sources, which in turn were more difficult than other comprehension questions. Both the mental models and digital literacy sets had adequate subskill reliability (α=.78 & .76 respectively).

Table 7 shows mean percent correct for subsets of items within each of these categories. Column two lists the source stimulus for item clusters. *None* signifies that all information was included in the item question and stem. Next, we examine in more detail the pattern of task results for each category in relation to the design.

Insert Table 6 & 7 about here

Mental model formation. Four task sets consisting of 16 items (α = 0.78) were designed to gather evidence of mental model formation. Three tasks used the organic farming explanation passage as stimuli. While the organic farming passage was not very long (341 words), the Flesch-Kincaid readability index was 6.9. Writing summaries was the most difficult task (22%), followed by the graphic organizer (26%), sequencing (36%), and questioning (55%) task sets. Perhaps surprisingly, the performance on the Pro vs. Con passage GO was higher (45%). The readability index for this 416 word Pro-Con passage was rated as more difficult than the organic farming passage, with a Flesch-Kincaid 8.6 grade level.

Background knowledge. The correlation of background knowledge to GISA score was r(426) =.51. Thus, about 25% of the variance in GISA scores was positively associated with examinees' familiarity with farming-related vocabulary.

Digital literacy tasks. Three task sets consisting of 13 items (α=0.76) were designed to simulate a web environment.

Evaluating credibility of multiple sources. Examinees performed moderately on evaluating the credibility of simulated search results (50%). We examined the pattern of responses across the four item subtypes in the evaluation task (not shown in Table 7) and observed that students found it more difficult to select the most relevant source to match their query (37%) than to identify whether a specific source was biased, written by an expert, or timely (average of 54%).

Evaluating perspectives. Examinees found classifying forum comments as on vs. off topic, opinions, or factually inaccurate to be a relatively easier task set (62%), than formulating responses to those comments by integrating information they had learned earlier in the assessment (37%).

Other items. Performance on three other MC items was moderate (48%). Students related information they read to a cartoon, evaluated information in a bar chart, and answered an inference question about the Pro vs. Con Passage. Performance on five traditional synonym items given as part of the SBA test was moderate (61%). These items had appeared prior to students reading passages in which the terms were used, and performance increased slightly in comparing pre (51%) vs. (61%) post performance (paired sample t-test (425)= 7.9, p<.001).

Task-oriented reading. The evidence in support of the scenario and teacher/peer elements of the test is that spending more time on these 'non-item' sections was a significant predictor of GISA scores even after controlling for time spent answering questions (see Table 4). Examinees spent an average of 16.7% of their total time on the non-item screens, with the highest quartile spending slightly more time than the average (18.3%).

Discussion

The purpose of this study was to design and evaluate a new reading comprehension assessment that was inspired by the theoretical and empirical research in reading and cognitive science and also broadened the construct of reading. The assessment design was premised on a body of research that suggests reading in 21st Century environments is a purpose driven activity (van den Broek et al, 2001) that involves integrating and synthesizing multiple texts (Strømsø et al, 2010), evaluating the relevance of sources (Rouet, & Britt, 2011), requires the use of relevant background knowledge (Shapiro, 2004) to form coherent mental model(s) (Kintsch, 1998), and requires the flexibility to read and perform tasks in digital environments (Leu et al., 2013). Collectively, these findings imply a different type of reading comprehension assessment is needed; one that goes beyond measuring a set of decontextualized, discrete skills, towards an assessment that broadens and contextualizes the construct of reading and takes advantage of advances in technology.

While these aspects of reading have solid foundations in the research literature, building a reliable and valid assessment that incorporates all these constructs into a single assessment is challenging. For instance, new constructs and tasks may be unfamiliar to students and without prior professional support for their teachers, students may not understand what is required of them as they attempt new item types. Compounding the problems associated with unfamiliarity, students might not be able to complete such a complex assessment in a single classroom period. In other words, the novelty and complexity may render a broader and more theoretically based assessment of reading infeasible.

Our approach to this problem was to adopt a scenario based assessment design (Author, 2009; 2013a, b; 2014; Bennett 2011; Bennett & Gitomer, 2009) that contextualized and structured the assessment in a simulated, familiar, social environment. The assessment incorporated a simulated teacher and students to help provide directions, highlight task goals, and to elicit construct relevant processing. In the current assessment, we somewhat limited the constructs coverage to focus on three key targets: background knowledge, mental model formation, and digital literacy skills. By constraining the construct and using a scenario-based assessment design, we hoped to produce an assessment that was theoretical and modern, while maintaining its feasibility for use in a typical classroom.

In light of these aims, the data seem to suggest the assessment was feasible (question 1), technically adequate and reliable (question 2), and displayed concurrent validity with another measure of reading comprehension (question 3). In terms of feasibility (question 1), most students were able to finish the assessment during a 50-60 minute classroom period. Although the scores were normally distributed, the overall mean of the assessment suggested that it was somewhat difficult for students. Part of this difficulty may have been a result of the new item types and constructs included in the assessment. We might anticipate that when students are given adequate preparation and practice responding to the novel task structures, that any construct irrelevant variance associated with adapting to the novelty of the SBAs will be reduced. Nonetheless, even under the novel administration conditions in this pilot, the assessment was reliable (question 2) and the student performance was correlated with more familiar, traditional reading comprehension performance.

Despite these strengths, one possible criticism of using a scenario-based approach is that the purpose-driven and thematic nature of the assessment might not generalize beyond the single assessment. The issue of reduced generalizability was one of the criticisms of the performance assessment movements in the 1990's which used more authentic and purpose-driven tasks (see Ryan, 2006). In reaction to this claim, the current SBA displayed a moderate to high correlation with a more traditional reading comprehension assessment. This result not only provides some evidence of the concurrent validity of the test (question 3), but also provides some evidence that the SBA design is not inherently content dependent. The overlapping variance between the SBA and the traditional measure of comprehension indicates that both assessments are measuring common aspects of a larger reading construct. While this result does not provide definitive evidence of the generalizability of SBAs, it does lend at least some support to the claim that what the SBA measures is not entirely restricted to the topics, texts, and items of the particular thematic form.

Another criticism of richer task environments such as performance assessments is that they were effortful to build and that this effort does not provide very much added value (Ryan, 2006). While a strict interpretation of added value goes beyond the scope of this paper, there is some preliminary evidence that richer aspects of the SBA design are related to student performance. The results of the current study indicate that the process data (timing information) revealed that the time a student spent on scenario related screens (non-item time) was associated with higher reading comprehension scores. The time spent on non-item screens predicted unique variance in comprehension scores even when the time on multiple choice and constructed response items was factored in to the hierarchical regression. In other words, contextual elements such as the teacher and peer interactions seemed to provide some value for students as time on these screens was associated with higher scores. While the precise reason for this association goes beyond the design of the current study, it might result from a number of factors ranging from reducing students' test anxiety by providing a friendlier context, to providing more information that would help students understand task expectations. Future research should explore these issues more systematically.

Beyond the issues related to the scenario, we were also interested in capturing two key aspects of reading comprehension, mental model formation and digital literacy. In particular, we wanted to determine whether the pattern of performance on these tasks was consistent with the design (question 5) and the literature. To this aim, mental model formation was measured by a combination of summary, graphic organizer, and questioning items. Of the mental model items, performance on the questioning items was higher than the performance on the other two item types. The questioning items are an indirect measure of the question generation technique, simplified so that students could choose, rather than generate, appropriate questions (e.g., Beck et al, 1997; Graesser & Lehman, 2006; Graesser & Person, 1994; Graesser et al., 2009) - a design change which may have made them a little less challenging than the typical instantiation of this technique. In contrast, the summary items were the most difficult mental model item type. Although some of this difficulty might be attributed to the response format (constructed response), in general, summary writing tends to be a difficult skill to master (Hill, 1991), as it requires students to extract key information, omit less important information, and organize the key points into a coherent whole. This process is demanding in part because it requires the integration of a number of key skills.

Indeed, forming a situation model (or mental model) requires students to integrate their background knowledge with the text (Kintsch, 1998). Although we did not measure the *integration* of background knowledge with the text, we did provide a measure of background knowledge that was related to the topic of the text. Consistent with the prior literature, we did observe a moderate correlation between background knowledge and comprehension on the SBA (Author, 2007; Shapiro, 2004). That is, students who knew more about the core topic of the assessment (farming) did score higher than students who knew less about the topic. The background knowledge might have helped students to draw knowledge based inferences (McNamara, de Vega & O'Reilly, 2007) or help them integrate what they read into an existing schema (Mandler, 1984).

In 21st century learning environments, building a mental model can be even more complex as students have to navigate the complexities of extracting meaning in digital environments. In the current study, digital literacy was operationalized and constrained to tasks assessing the evaluation of sources, classifying and evaluating different perspectives, as well as tasks that measured how students integrate and evaluate perspectives across multiple texts. Tasks that required students to classify and evaluate perspectives were the easiest in this set. Although this set included theoretically more difficult metacognitive type tasks such as identifying incorrect and off-topic information (Hacker, Dunlosky, & Graesser, 2009), there were other items in this task set that are common in the standards of typical U.S. classrooms such as distinguishing fact from opinion (NGA & CCSSO, 2010). This overlap with the curriculum might be one of the reasons why this task set was one of the easiest in the entire assessment.

In contrast, the digital literacy task that required students to evaluate and integrate different perspectives across texts proved very difficult. This task not only required students to engage in perspective taking (Zhang et al., 2013), but also to navigate and integrate information across multiple texts (Strømsø et al., 2010). Although young students can integrate information across texts when supported (Goldman, 2004), understanding multiple texts is a difficult skill to master. The difficulty in this task might be augmented by the working memory load required to remember what person was associated with what perspective, and where the information was located across multiple passages. Holding all this information in memory at one time might be particularly demanding for the students in this study.

Limitations and Challenges

This was one pilot of one form of a prototype SBA approach. We were missing a quarter of the sample demographic information about students, so we cannot make broader claims to the representativeness of the sample or any biases in test items in subgroups. We also lacked instrumentation to evaluate student motivation or other usability challenges in test administration. We are currently conducting more usability studies that explore such issues, as well as gathering teacher feedback on the quality and appropriateness of the tasks. We were also unable to help students or teachers prepare for this novel testing situation. In the future, we hope to prepare materials so that the experience is not quite so novel. We are in the process of generating forms that vary topic and task type to further explore the generalizability of the design characteristics of these tests.

There remain numerous, complex technical issues that will also need to be addressed in moving towards operational use of scenario-based assessments. These include potential violations of item independence; efficient scoring of constructed-response items; stringent evaluation of generalizability across forms; dimensionality; reporting of new kinds of information beyond traditional, unidimensional scale scores (e.g., reporting on background knowledge); etc. In this study, we presented preliminary evidence of the promise and feasibility that we see as encouragement for moving forward.

Finally, we note that the overall time spent on items was positively correlated with performance. It seems that because of the challenge level, those examinees in the bottom quartile may have chosen not to persist, which is not ideal -- a weakness in this design we must address in the future. However, we also note that the highest quartile students also spent the most time, despite the well-known finding that better readers are more fluent and efficient readers, which could have resulted in them spending less or at least the same time overall as other groups. This is consistent with the challenge level required to reason about content and integrate knowledge and information across texts and tasks. Deeper processing of content may neutralize the automaticity and fluency of advantage they typically enjoy.

Conclusion

The scenario-based approach, while potentially fruitful, also presents various accompanying challenges. The reason that the traditional design of reading comprehension tests may not have changed in decades is because it is efficient and effective in producing reliable and interpretable scores. It is also very familiar to test takers and consequently, students know what to expect and how to respond to traditional item types. The SBA introduces several novel elements to the test session including a section probing examinee background knowledge, an introduction of an overall scenario that sets a larger purpose for the texts and tasks, questions that probe mental models, sections that introduce and assess multiple texts that are thematically related to the larger purpose, and simulated teacher and peer interactions that are situated in a modern 21st century digital environment. Thus, building a case for the importance and validity of broadening the construct of reading and introducing changes to the typical design is necessary. However, even if the changes were deemed valid, their use would be undermined if administration and scoring proved infeasible or impractical for typical classroom application. The evidence in this study can be viewed as a concept proof that SBAs can be designed to address the simultaneous challenge of enhanced theoretical foundations and practical use.

With this particular SBA form, middle grade students were given a broad, task-oriented scenario for reading a set of thematically related sources, then asked to reason, integrate, synthesize and evaluate information from traditional and digital texts. The coordination of these skills is quite complex, and building an assessment to measure them required innovating item and tasks that moved beyond the traditional approaches used in passage-question reading comprehension tests. While not every innovation showed evidence of success, preliminary data indicated students were able to perform complex, integrated skills, and it is possible to measure these skills as evidenced by the psychometric properties of the test for students at this grade. We view this as encouragement to researchers and assessment designers alike that scenario-based, theory-based measures that broaden the construct coverage of tests can and should be designed and studied, so that we can better understand the reading behavior of learners, and use that understanding to inform learning and instruction.

References

Afflerbach, P.A., & Cho, B.Y. (2010). Determining and describing reading strategies: Internet and traditional forms of reading. In H.S. Waters & W. Schneider (Eds.), M*etacognition, strategy use, and instruction* (pp. 201–255). New York, NY: Guilford.

Armbruster, B. B., Anderson, T. H., & Meyer, J. L. (1991). Improving content-area reading using instructional graphics. *Reading Research Quarterly, 26*, 394-416.

Author (2007).

Author (2009).

Author (2013a).

Author (2013b).

Author (2014).

Bean, T. W., Singer, H., & Frazee, C. (1986) The effect of metacognitive instruction in outlining and graphic organizer construction on students' comprehension in a tenth-grade world history class. *Journal of Reading Behavior, 18*, 153-169.

Bean, T. W., & Steenwyk, F. L. (1984). The effect of three forms of summarization instruction on sixth-graders' summary writing and comprehension. *Journal of Reading Behavior 16*, 297-306.

Beck, I. L., McKeown, M. G., Hamilton, R., & Kucan, L. (1997). *Questioning the author: An approach for enhancing student engagement with text.* Newark, DE: International Reading Association.

Bennett, R.E. (2011). *CBAL: Results from piloting innovative K–12 assessments* (ETS Research Report No. RR-11-23). Princeton, NJ: ETS.

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.

Bloom, B., Englehart, M. Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York, NY: Longman.

Bråten, I., Gil, L., & Strømsø, H. (2011). The role of different task instructions and reader characteristics when learning from multiple expository texts. In M. T. McCrudden, J. Magliano, & G. Schraw (Eds), *Text relevance and learning from text* (pp. 53-74). Greenwich, CT: Information Age Publishing.

Britt, M. A., & Sommer, J. (2004). Facilitating textual integration with macro-structure focusing tasks. *Reading Psychology*, *25*, 313-339.

Cain, K., & Oakhill, J. (2012). Reading comprehension development from seven to fourteen years: Implications for assessment. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.). *Measuring up: Advances in how we assess reading ability* (pp. 59-76). Lanham, MD: Rowman & Littlefield Education.

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.

Coiro, J. (2011). Predicting reading comprehension on the Internet: Contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of Literacy Research, 43,* 352-392.

Coiro, J., & Dobler, E. (2007). Exploring the online comprehension strategies used by sixth-grade skilled readers to search for and locate information on the Internet. *Reading Research Quarterly, 42*, 214–257.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta -analytic review. *Review of Educational Research, 66,* 227-268.

Deane, P. (2012). NLP methods for supporting vocabulary analysis. In J. Sabatini, T., O'Reilly, & E. Albro (Eds.), *Reaching an understanding: Innovations in how we view reading assessment* (pp. 117-144). Lanham, MD: Rowman & Littlefield Education.

Dijkstra, K., Zwaan, R. A., Graesser, A. C., & Magliano, J. P. (1995). Character and reader emotions in literary narrative. *Poetics, 23,* 139-157.

Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models.* Hillsdale, NJ: Erlbaum.

Gernsbacher, M. A. (1997). Two decades of structure building. *Discourse Processes, 23,* 265-304.

Goldman, S. R. (2004). Cognitive aspects of constructing meaning through and across multiple texts. In N. Shuart-Ferris & D. M. Bloome (Eds.), *Uses of intertextuality in classroom and educational research* (pp. 317–351). Greenwich, CT: Information Age Publishing.

Graesser, A. & Lehman, B. (2006). Questions drive comprehension of text and multimedia. In M. T. McCrudden, J. Magliano, & G. Schraw, (Eds), *Text relevance and learning from text* (pp. 53-74). Greenwich, CT: Information Age Publishing.

Graesser, A. C., Ozuru, Y., & Sullins, J. (2009). What is a good question? In M. G. McKeown & L. Kucan (Eds.), *Threads of coherence in research on the development of reading ability* (pp. 112-141). New York, NY: Guilford.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*, 104-137.

Graesser, A., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101,* 371-395.

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.

Griffin, C. C., Malone, L. D. & Kammenui, E. J. (1995). Effects of graphic organizer instruction on fifth grade students. *Journal of Educational Research*, *89*, 98-107.

Guri-Rozenblit, S. (1988). Impact of diagrams on recalling sequential elements in expository texts. *Reading Psychology: an International Journal*, *9*, 121-139.

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

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology*, *93*, 103-128.

Hill, M. (1991). Writing summaries promotes thinking and learning across the curriculum -- but why are they so difficult to write? *Journal of Reading*, *34*, 536-639.

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

Katz, I. & Macklin, A. S. (2007). Information and communication technology (ITC) literacy: Integration and assessment in higher education. *Journal of Systemics, Cybernetics and Informatics 3*, 50-55.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kintsch, W. (2012). Psychological models of reading comprehension and their implications for assessment. In J.P. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 21-38). Lanham, MD: Rowman & Littlefield Education.

Lawless, K. A., Goldman, S. R., Gomez, K., Manning, F., & Braasch, J. (2012). Assessing multiple source comprehension through evidence-centered design. In J. Sabatini, T., O'Reilly, & E. Albro (Eds.), *Reaching an understanding: Innovations in how we view reading assessment* (pp. 3-17). Lanham, MD: Rowman & Littlefield Education.

Leu, D. J., Kinzer, C. K., Coiro, J., Castek, J., & Henry, L. A. (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.

Leu, D.J., Zawilinski, L., Castek, J., Banerjee, M., Housand, B., & Liu, Y. (2007). What is new about the new literacies of online reading comprehension? In L. Rush, J. Eakle, & A. Berger (Eds.), *Secondary school literacy: What research reveals for classroom practices* (pp. 37–68). Urbana, IL: National Council of Teachers of English.

Magliano, J., Millis, K., Ozuru, Y., & McNamara, D. (2007). A multidimensional framework to evaluate reading assessment tools. In D. McNamara (Ed.), *Reading comprehension strategies: Theory, interventions, and technologies* (pp. 107-136). Mahwah, NJ: Erlbaum.

Mandler, J. M. (1984). *Stories, scripts, and scenes: Aspects of schema theory*. Hillsdale, NJ: Erlbaum.

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

McKoon, G. & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99,* 440-466.

McNamara, D. S., de Vega, M., & O'Reilly, T. (2007). Comprehension skill, inference making, and the role of knowledge. In F. Schmalhofer & C. A. Perfetti (Eds.), Higher level language processes in the brain: Inference and comprehension processes (pp. 233–251). Mahwah, NJ: Erlbaum.

McNamara, D.S. M., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of Learning And Motivation, 51*, 297-384.

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.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from http://timssandpirls.bc.edu/pirls2011/downloads/PIRLS2011_Framework.pdf

Myers, J. L. & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes, 26*, 131-157

Narvaez, D., van den Broek, P., & Ruiz, A. (1999). Reading purpose, type of text and their influence on think-aloud and comprehension measures. *Journal of Educational Psychology, 91*, 488-496.

National Governors Association Center for Best Practices & Council of Chief State School Officers (NGA & CCSSO) (2010). *Common Core State Standards for English Language Arts*. Washington, DC: Author.

McNamara, D. S., de Vega, M., & O'Reilly, T. (2007). Comprehension skill, inference making, and the role of knowledge. In F. Schmalhofer & C. A. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes* (pp. 233–251). Mahwah, NJ: Erlbaum.

Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 7-14). Hillsdale, NJ: Erlbaum.

Organisation for Economic Co-operation and Development. (2009a). *PISA 2009 assessment framework—Key competencies in reading, mathematics and science.* Paris, France: Author.

Retrieved from http://www.oecd.org/document/44/0,3746,en_2649_35845621_44455276_1_1_1_1,00.html

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

Palincsar, A., & Brown, A. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1,* 117-175.

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 map.* Washington, DC: Author. Retrieved from http://www.p21.org/storage/documents/21st_century_skills_english_map.pdf

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

Robinson, D. H., & Kiewra, K. A. (1995). Visual argument: graphic organizers are superior to outlines in improving learning from text. *Journal of Educational Psychology, 87*, 455-467.

Rouet, J.-F. (2006). *The skills of document use: From text comprehension to web-based learning*. Mahwah, NJ: Erlbaum.

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

Rupp, 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.

Ryan, T. J. (2006). Performance assessment: critics, criticism, and controversy. *International Journal of Testing, 6,* 97-104.

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.

Strømsø, H.I., Bråten, I., & Britt, M.A. (2010). Reading multiple texts about climate change: The relationship between memory for sources and text comprehension. *Learning and Instruction, 20,* 192-204.

Taylor, B. M. (1982). A summarizing strategy to improve middle grade students' reading and writing skills. *The Reading Teacher, 36*, 202-205.

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Educational Psychology 28*, 129-160.

Trabasso, T. (2005). The role of causal reading in understanding narratives. In T. Trabasso, J. P. Sabatini, D. C. Massaro & R. C. Calfee (Eds.), *From Orthography to Pedagogy: Essays in Honor of Richard L. Venezky* (pp. 81-106). Mahwah, NJ: Lawrence Erlbaum.

van den Broek, P. (2012). Connecting cognitive theory and assessment: Measuring individual differences in reading comprehension. *School Psychology Review, 41*, 315-325.

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

van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The landscape model of reading. In. H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 71-98). Mahwah, NJ: Erlbaum.

van Dijk, T.A. & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wiley, J., Goldman, S., Graesser, A., Sanchez. C., Ash, I., & Hemmerich, J. (2009) Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal*, *46*, 1060-1106.

Zhang, X., Fung, H. H., Stanley, J. T., Isaacowitz, D. M., & Ho, M. (2013). Perspective taking in older age revisited: A motivational perspective. Developmental Psychology, 49(10), 1848-1858.

Zwaan, R. A. (2008). Time in language, situation models, and mental simulations. *Language Learning*, *58*13-26.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*, 292-297.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185

Table 1

Means (SD) and Correlations of Reading Subtests to SBA and Reading Comprehension Tests (N = 161)

	Mean (SD)	SBA	Reading Comp
SBA	54.11 (24.05)	1.00	
Reading Comprehension	356 (31.37)	0.77	1.00
Efficiency of Comprehension	366 (31.83)	0.76	0.78
Sentence Processing	356 (31.81)	0.70	0.73
Morphology	366 (30.97)	0.76	0.76
Vocabulary	371 (27.50)	0.71	0.66
Word Recognition/Decoding	364 (28.29)	0.70	0.69

Note: Reading scores of 370 and higher are considered proficient. Scores of 330 and lower are considered below basic proficiency.

Table 2

SBA and Reading Assessment Means (SD) for Quartile Ability Groups Based on SBA Scores (N = 161)

	Quartile 1	Quartile 2	Quartile 3	Quartile 4	Average
SBA % correct	24 (6)	41 (5)	59 (6)	84 (9)	54 (24)
Reading Comp	325 (19)	337 (23)	366 (20)	386 (16)	356 (31)
Efficiency	331 (24)	353 (28)	379 (19)	391 (11)	365 (32)
Sentence	325 (23)	342 (27)	366 (22)	381 (19)	356 (32)
Morphology	333 (26)	353 (24)	376 (20)	391 (12)	365 (31)
Vocabulary	343 (27)	362 (25)	381 (14)	393 (8)	371 (28)
Word Rec/Decoding	339 (27)	348 (25)	372 (17)	387 (13)	364 (28)

Table 3

Means (SD) and Correlations of Time Variables (Minutes) to SBA Total Score (N = 426)

	Mean (SD)	1	2	3	4	5	6
1. SBA % Correct	53.33 (23.09)	1.00					
2. Multiple Choice Time	15.62 (6.48)	0.59	1.00				
3. Constructed Response Time	10.46 (5.43)	0.45	0.57	1.00			
4. Non-Item Time	6.24 (3.15)	0.39	0.40	0.39	1.00		
5. Background Knowledge Time	3.02 (1.01)	-0.16	0.24	0.35	0.21	1.00	
6. Total Session Time	35.34 (12.67)	0.58	0.88	0.84	0.64	0.41	1.00

Table 4

Hierarchical, Multiple Regression with Time Predicting SBA Total Score (n=337)

	Adj R^2	R^2 Change	F Change	Sig. F Change
Multiple Choice Time	.341	.343	174.6	.000
& Constructed Response Time	.360	.021	11.2	.001
& Background Knowledge Time	.488	.129	84.5	.000
& Non-item Time	.516	.029	20.2	.000

Note: Time values above 2.5 SD from mean were omitted, resulting in n=337 listwise.

Table 5.

Means (SD) for Quartile Ability Groups Based on SBA Test Scores (Percent Correct)

	Quartile 1	Quartile 2	Quartile 3	Quartile 4	Average
SBA % Correct	24.90 (5.82)	41.69 (4.77)	59.18 (5.45)	83.58 (8.94)	53.33 (23.09)
Multiple Choice Time	10.18 (5.66)	13.61 (5.38)	18.11 (4.89)	19.92 (4.95)	15.62 (6.48)
CR time	7.08 (5.10)	9.68 (5.14)	11.32 (4.61)	13.40 (4.84)	10.46 (5.43)
Non-Item time	4.64 (2.86)	5.77 (3.37)	6.16 (2.78)	8.20 (2.54)	6.24 (3.15)
BK Time	3.20 (1.21)	3.16 (1.04)	2.93 (0.93)	2.85 (0.79)	3.02 (1.01)

Note: CR = Constructed Response, BK = Background Knowledge

Table 6

Means (SD) of Percent Correct, Reliability, and Correlations for Three Item Types

	Items	Mean (SD)	Alpha	Mental Model	Digital Literacy
Mental Model	16	0.49 (.26)	0.78	1	
Digital Literacy	13	0.52 (.24)	0.76	0.72	1
Miscellaneous	8	0.56 (.25)	0.64	0.70	0.69

Table 7

Percent Correct on Clusters of Items

	Source Text	Items	Percent Correct
Mental Models		**16**	**49%**
Summary Writing	Organic Farming	2	22%
GO - 3 x 4 levels	Organic Farming	4	26%
Sequencing	Organic Farming	3	36%
Questioning	Organic Farming	3	55%
GO 2 x 2 levels	Pros and Cons	4	45%
Background		**29**	**59%**
Vocabulary	None	5	51%
Topic list	None	24	61%
Digital Literacy		**13**	**52%**
Eval.Credibility of Sources	Search Result Descriptions	3	50%
Eval. Perspectives-Classify	Forum Discussion	6	62%
Eval. Perspectives-Integrate	Forum Discussion	4	37%
Miscellaneous		**6**	**56%**
MC-items	*	3	48%
Vocabulary	None	5	61%

Notes: * 1-item each to stimulus Pro and Con passage, cartoon, and bar chart.

Endnotes

[i] Unfortunately, about 25% of the demographic information was missing or could not be matched in this dataset. We compared the full school demographics to the subsample who participated in the study and while there were some minor differences relative proportion at the school level, overall the demographics reported provide a general sense of the sample.