Listening. Learning. Leading.®

Research Report
ETS RR-13-31

Reading for Understanding: How Performance Moderators and Scenarios Impact Assessment Design

Tenaha O'Reilly

John Sabatini

December 2013

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Reading for Understanding: How Performance Moderators and Scenarios Impact Assessment Design

Tenaha O'Reilly and John Sabatini
Educational Testing Service, Princeton, New Jersey

December 2013

Action Editor: Donald E. Powers

Reviewers: Paul Deane and Jane Shore

Abstract

This paper represents the third installment of the Reading for Understanding (RfU) assessment framework. This paper builds upon the two prior installments (Sabatini & O'Reilly, 2013; Sabatini, O'Reilly, & Deane, 2013) by discussing the role of performance moderators in the test design and how scenario-based assessment can be used as a tool for assessment delivery. Performance moderators are characteristics of students that impact reading performance but are not considered a part of the reading construct. These include (a) background and prior knowledge, (b) metacognitive and self-regulatory strategies and behavior, (c) reading strategies, and (d) student motivation and engagement. In this paper, we argue there is added value in incorporating performance moderators into a reading test design. We characterize added value with respect to the validity of the claims derived from test scores, the interpretation of the test scores, and the relevance to instruction. As a second aim, we present a case for using scenario-based assessments and how they can be used to integrate into the test design both the performance moderators as well as other features that make the assessment more instructionally relevant.

Key words: reading comprehension assessment, scenario-based assessment, background knowledge, motivation, reading strategies, metacognition, self-regulation, GISA, reading for understanding

Acknowledgments

This report was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305F100005 to 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, the U.S. Department of Education, or Educational Testing Service. We are extremely 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 Paul Deane, Jane Shore, and Don Powers for their thoughtful comments and Jennifer Lentini 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, and for, Learning (*CBAL*™) colleagues who have been and are contributing to this ongoing enterprise.

Table of Contents

List of Tables

Page

Background and Context: The Two Prior Frameworks

This paper represents the third installment of the Reading for Understanding (RfU) assessment framework. Part one of the framework covered the rationale for a new generation of reading assessments and a set of six guiding principles for test design (Sabatini & O'Reilly, 2013). These principles represented a distillation of the research in reading and cognitive science that is useful for merging theory and empirical findings with the next generation of reading assessments. While some of the principles discuss empirical and theoretical issues, such as vocabulary, that are already covered on many existing reading tests, other principles cover issues that are not routinely addressed, such as goal-directed reading (or task-oriented reading), multiple source integration, and digital literacy.

The second installment of the reading framework built upon the first by providing a definition of reading for understanding, the key constructs to be assessed, a position on reading development, and an overview of two types of assessments (Sabatini, O'Reilly, & Deane, 2013). Five dimensions of reading literacy were described: writing (or print) system, language (or verbal) system, text and discourse, conceptual modeling/reasoning, and social modeling/reasoning (see Table 1). These dimensions serve as analytic categories for decomposing literacy tasks, such that one can describe or evaluate the relative contribution of skills necessary to perform the task successfully. In addition to the five dimensions, two types of assessments were also described: reading components and a scenario-based assessment called the Global Integrated Scenario-Based Assessment (GISA; also see Table 1).

The Assessments

Reading components assessments are used to evaluate proficiency in the first two dimensions of the model, the writing (or print) system and the language (or verbal) system, and to some extent, the text and discourse dimension. Thus, the components' tests are designed to measure the reading subskills that enable basic processing of text including word recognition, morphology, word reading efficiency, vocabulary, syntax, and lower level reading comprehension. Each of the reading subskills is measured separately in distinct subtests to help isolate component skill strengths and weaknesses (O'Reilly, Sabatini, Bruce, Pillarisetti, & McCormick, 2012; Sabatini, Bruce, & Steinberg, 2013).

Table 1

The Two Types of Assessments in the Reading for Understanding (RfU) Framework as a Function of Dimension

Assessment types	Writing/print (includes typographical information, decoding, word recognition)	Language/ verbal (includes vocabulary, syntax, sentence and local understanding)	Text & discourse (includes genre and text structure and global understanding)	Conceptual (includes conceptual reasoning, evaluation, integration, synthesis, application)	Social (includes social reasoning, intent, motive, author purpose)
Component assessments (assess foundational reading skills and lower level comprehension; skills specific to printed material)	Primary target	Primary target	Secondary target	Not covered	Not covered
GISA (assess higher level comprehension, critical thinking, deep understanding, application, transfer; skills not specific to printed materials)	Indirectly covered	Secondary target	Primary target	Primary target	Primary target

Note. GISA = Global Integrated Scenario-Based Assessment.

In contrast to the components' assessments, the second type of assessments, GISA, is designed to measure higher level reading comprehension in an integrated way. In the RfU model, GISA covers all levels but focuses primarily on the text and discourse, conceptual, and social dimensions. Rather than trying to isolate specific skills, GISA is designed to evaluate the orchestration of the five dimensions in complex reading literacy tasks. This orchestration is achieved by incorporating a scenario-based design that organizes the assessment around a central theme and goal for reading (e.g., work with fellow students to study for an exam or prepare a presentation on a science or history topic, unravel the controversy surrounding who was the

model for Da Vinci's painting of the Mona Lisa), a set of diverse sources (e.g., blogs, Web sites, videos, charts and diagrams, traditional text genre excerpts), and a sequence of subtasks to achieve the final goal (e.g., evaluate sources, identify important or relevant ideas, integrate information across sources, make decisions, edit a wiki). In this manner, GISA is designed to resemble the types of reading activities one might engage in school, work, or leisure. This design feature not only helps bolster the authenticity of the assessment but also is intended to encourage the deeper processing demanded by a host of recent initiatives such as the Race to the Top (U.S. Department of Education, 2009), Common Core State Standards (National Governors Association Center for Best Practices [NGACBP] & Council of Chief State School Officers [CCSSO], 2010) and Partnership for 21st Century Skills (2004, 2008), as well as other seminal efforts for assessment innovation (Bennett, 2011a, 2011b; Bennett & Gitomer, 2009; Gordon Commission, 2013; Pellegrino, Chudowsky, & Glaser, 2001).

New Challenges

Collectively, these two installments of the framework offer an alternative approach to reading assessment. The approach advocated here aims to broaden the construct of reading beyond what is traditionally measured in many off-the-shelf reading assessments. The challenge of broadening the construct is also met with the simultaneous aim of engaging readers in activities that are designed to elicit deeper processing. While this approach has several advantages, it also poses some new challenges for test designers and measurement specialists. Some of these challenges concern design and implementation, while others impact scoring of tasks or interpreting ensuing reading scores. For instance, broadening the design space to more explicitly draw upon student's metacognition (e.g., thinking about one's own thinking) and reading strategies introduces new challenges of how to build tasks to effectively implement these ideals in an assessment context. Similarly, creating themed assessments on a particular topic to enable deeper processing introduces new challenges of how to account for individual differences in student background knowledge and motivation. Both student background knowledge and motivation can impact the interpretation of a reading score. For instance, students who already know more about the theme or topic of the assessment may have an advantage and score higher than students who have low knowledge. Similarly, students who are more interested in the topic or theme of the assessment might be more engaged and score higher than students who are less interested in the topic. In either case, it is difficult to determine whether the reading score

actually reflects true reading ability or individual differences in motivation and or background knowledge. This uncertainty represents a potential threat to test score interpretation.[1]

Goals of This Report

The purpose of this installment of the framework is to describe the role of performance moderators in a new assessment design like GISA. This report is organized into two major sections. This first section discusses the notion of performance moderators. By *performance moderators*, we mean theoretical and practical factors that impact reading performance but are not typically considered direct targets of proficiency. That is, in empirical studies, these factors are associated with reading proficiency, but it has not been established that these are necessary aspects of skilled performance or the product of completing literacy tasks proficiently. The performance moderators we include are (a) background and prior knowledge, (b) metacognitive and self-regulatory strategies and behavior, (c) reading strategies, and (d) student motivation and engagement.[2] For each class of moderators, we discuss the theoretical underpinnings and preliminary ways one might measure them; for two moderators (background knowledge and motivation), we discuss techniques that can be applied to enhance the interpretation of reading scores. Our goal is to make factors such as these more explicit in the assessment design and, thus, enhance interpretation of test scores. At the same time, we do not see strong rationales or evidence to support positing them as required performance outcomes in and of themselves.

Subsequent to this discussion, we describe one technique we use, the scenario, to organize the assessments. Scenarios are useful techniques for structuring and sequencing assessment tasks to account for performance moderators and to gather more information about test takers. While the notion of a scenario might be useful for assessment, it is not well specified in the existing literatures and can be used to describe a wide array of instantiations. For this reason, a key aim of this section will be to describe the various facets of a scenario ranging from minimalist interpretations (e.g., providing single purpose for reading up front) to more complex and intricate instantiations (highly organized sequences of task and interactions). Below we begin our discussion with an overview of performance moderators.

Performance Moderators and Global Integrated Scenario-Based Assessment (GISA)

In this section, we address what we call *performance moderators* as they impact assessment design, score interpretation, and the utility of assessment results. Performance

moderators are theoretical constructs (or individual differences) derived from the theoretical and empirical literature, which can impact or interact with the interpretation or claims about what the test is measuring. In some cases, performance moderators can be viewed as introducing construct-irrelevant variance (e.g., low motivation underestimates true reading ability). In other cases, the concern is more about the absence of any attempt to measure or understand their potential impact on scores despite considerable discussion of their importance in the learning sciences literature (e.g., whether a student deploys appropriate self-regulatory or reading strategies). Below is a short list of some commonly voiced concerns about how performance moderators might detract from test score interpretation and utility.

- The score is not a measure of reading ability, but rather a test of knowledge. An individual with knowledge of the content would not have to read the text to answer the items correctly (see Katz & Lautenschlager, 2001).
- Skilled readers have been found to have strong self-regulation and metacognitive abilities (Hacker, Dunlosky, & Graesser, 2009; Pressley, 2002). However, the assessments do not address these constructs at all.
- Skilled readers have been found to utilize reading strategies (McNamara, 2007). However, the assessments do not address these strategies at all.
- The score does not reflect true reading ability because the students were not motivated or interested, and therefore, the test underestimates their true abilities (see Braun, Kirsch, & Yamamoto, 2011).

In the sections below, we outline the nature of the issues or concerns raised then introduce some assessment design techniques for addressing them.[3]

Claims About Reading Proficiency

To contextualize the approaches below, we suggest the following claims about general reading proficiency. In Sabatini et al. (2013), the authors stated that reading proficiency is composed of the knowledge, skills, strategies, and dispositions that enable readers

- to learn and process the visual and typographical elements and conventions of printed texts;
- to learn and process the verbal elements of the English language including grammatical structures and word meanings;

- to learn and process the discourse structures, forms, and genres of print;
- to model and reason about conceptual content; and
- to model and reason about social content.

Given that reading is a goal-directed, purposeful cognitive activity, we elaborate on the above claims to add that in order to achieve a reading purpose or goal, proficient readers will (see Table 2).

- apply relevant background or prior knowledge, as necessary;
- apply appropriate self-regulatory, metacognitive, or reading strategies to construct their understanding, as necessary; and
- exert sufficient cognitive effort.

These three dispositional aspects of reading behavior are consistent with what Guthrie and colleagues term *engagement* (see Guthrie, McGough, Bennett, & Rice, 1996; Guthrie & Wigfield, 2000). Note the requirement is not that readers possess appropriate background knowledge or strategies, but rather that if they do, they deploy them as demanded by the reading context. The third aspect states an expectation of intrinsic motivation to deploy one's entire repertoire of reading skills in reading situations that demand the use of those skills. Below we discuss how these performance moderators—a) background and prior knowledge, b) metacognitive and self-regulatory strategies and behavior, c) reading strategies, and d) student motivation and engagement—impact and can shape the design of reading comprehension assessments. Table 2 provides an overview of the measurement goals, potential score interpretation value, and instructional benefits they can afford to reading assessments.

Background, or Prior, Knowledge

Definition. In the current manuscript, the terms *background knowledge* or *prior knowledge* are used interchangeably[4] to refer to the associated topical vocabulary, concepts, relations among concepts, and associated knowledge-based inferences that are not explicitly mentioned in text, but are necessary (or useful) for the reader to form a coherent understanding of text. At a more specific level, background knowledge includes common cultural references, knowledge associated with specific experiences or age cohorts, and various nuances associated with a particular language (e.g., clichés or idioms).

Table 2

The Reading for Understanding (RfU) Performance Moderators as a Function of Score Interpretation Value and Instructional Benefit

Performance moderator	Measurement goal	Score interpretation value	Instructional benefit
Background knowledge	To provide an indicator of students' background knowledge on the topic of the texts and sources in the assessment	Can be used as an indicator of students' ability to learn new information Can be used as an indicator of students' ability update existing knowledge and apply it to complex tasks	Identify students who may have insufficient knowledge to understand text Provide additional resources to increase knowledge before reading
Metacognitive and self-regulatory strategies	To provide an indicator of students' ability to monitor their understanding and their ability take action to repair gaps, errors, and misconceptions	Can be used as an indicator of the accuracy of students' judgments of learning Can be used as an indicator of students' ability to recover from errors and use available resources to solve problems	Identify students with weakness in self monitoring and self regulation behaviors Provide training in improving judgments and repairing strategies
Reading strategies	To provide an indicator of students' strategic use of text	Can be used as an indicator of students' ability to use local strategies such as paraphrase Can be used as an indicator of students' ability to use global strategies such as summarization	Indentify students who are not strategically processing text Provide training in the use of local, global, and other strategies
Motivation and engagement	To provide an indicator of students' willingness to expend sufficient effort to understand text	Can be used as an indicator of students' interest on the topics and texts of the assessment Can be used as an indicator of students' engagement with specific tasks	Indentify students who are not motivated Provide motivation training that involves the use of engaging materials and activities

Background and context. As its title suggests, the Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects (NGACBP & CCSSO, 2010) puts strong emphasis on the importance of content area reading and cross-disciplinary thinking. The Standards recognizes the critical role that reading comprehension plays in learning content, but at the same time, this new emphasis poses real challenges for assessment designers. In particular, blending reading comprehension with topic-specific reading materials can reduce the validity of the interpretation of test takers' scores on a reading comprehension test—high reading scores may provide evidence of skilled reading or, alternatively, the scores may be more reflective of test takers' existing knowledge on the topic.

Traditional assessment designs avoid providing test takers with topic-specific materials because of the risk that doing so unfairly advantages or disadvantages some test takers over others. Recognizing the impact of background knowledge on reading comprehension, test designers try to reduce the impact of this *construct-irrelevant* variance by utilizing two primary design techniques. First, passages are selected to ensure the reading material is general enough to be accessible to a wide range of readers. Topics that require special knowledge are avoided in favor of topics that most people should be familiar with. A second technique to reduce the impact of background knowledge is to include a wide range of passages on the test that cover a number of general topics. By including a large number of topics on the test form, it is expected the effects of background knowledge will be mitigated because it is assumed test takers will have background knowledge on some passages but not others.

While this logic makes intuitive sense, there is no guarantee that it is effective in reducing the impact of background knowledge, especially at the individual level (see Shapiro, 2004). An individual might know all or none of the topics sampled. It is also an unsatisfactory solution in that it does not address the important role of background knowledge in comprehension so much as ignoring it and hoping it is not that significant an influence.[5] In part, because assessments have evolved to be time efficient (i.e., to maximize measurement precision at the minimal test duration), attempts to measure background knowledge in a standardized test are rare, and thus, there is little evidence as to how much influence background knowledge may have on an individual's reading comprehension score. As interest in reading for understanding in the content areas increases, the influence of background knowledge may also increase and further reduce

confidence that comprehension scores reflect reading ability versus content knowledge (see O'Reilly & McNamara, 2007a, 2007b).

Why background knowledge matters. Rather than treating background knowledge solely as a construct-irrelevant nuisance, it also can be seen as an opportunity to improve the interpretation of reading scores and to model good practice. Like key theories of reading (e.g., construction-integration [Kintsch, 1998], landscape [van den Broek, Young, Tzeng, & Linderholm, 1999]), we regard background knowledge as an important influence on students' ability to comprehend text. Test takers who have more background knowledge comprehend more from text than test takers with low background knowledge (Adams, Bell, & Perfetti, 1995; Alexander, Sperl, Buehl, & Chiu, 2004; Cromley & Azevedo, 2007; Dochy, Segers, & Buehl, 1999; Fincher-Kiefer, Post, Greene, & Voss, 1988; Hambrick & Engle, 2002; McNamara, 1997, 2001; McNamara, de Vega, & O'Reilly, 2007; McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996; Murphy & Alexander, 2002; O'Reilly & McNamara, 2007a, 2007b; Ozuru, Best, Bell, Witherspoon, & McNamara, 2007; Ozuru, Dempsey, & McNamara, 2009; Recht & Leslie, 1988; Schneider, Körkel, & Weinert, 1989; Shapiro, 2004; Spilich, Vesonder, Chiesi, & Voss, 1979; Thompson & Zamboanga, 2004; van den Broek, 2012; Voss & Silfies, 1996; Walker, 1987). While background knowledge can facilitate comprehension, in some cases background knowledge can actually interfere with reading comprehension when knowledge is irrelevant or violated by the text (Kucer, 2011).

Background knowledge can support reading comprehension by enabling test takers to infer unstated relationships between elements in the text (i.e., knowledge-based inferences [McNamara et al., 2007]). Alternatively, background knowledge may serve as a template or schema to integrate new information (Mandler, 1984; Piaget, 1957; Rumelhart, 1980; Schank, 1977). When background knowledge is high, the test taker has already built a mental model of the topic even before reading. In such cases, the text merely provides a way to refresh, reactivate or update what is already known; it is not used to determine whether a test taker can form a mental representation from scratch. Thus, when test takers have high knowledge on the topic of the texts, the assessment might be measuring background knowledge rather than reading ability.

Ways to measure and support background knowledge. While we do not expect to solve the problem of background knowledge, we argue that there is added value in incorporating features into the design that address the issue head on. As in previous assessments, we advocate

including a wide range of passages, genres, and perspectives in each assessment form (as time permits). This is designed both to address individual differences in background knowledge as well as to broaden students' exposure to different content topics, text structures, and genres. Going beyond traditional practice, however, in GISA designs, we also attempt to measure and support background knowledge. In the 21st century literacy environment, information and knowledge about any topic is now readily accessible via anytime, anywhere technological devices (e.g., computers, tablets, handhelds, phones) linked to a networked, virtual universal library (i.e., the Internet/World Wide Web). It seems counter to 21st century reading to assume a literacy environment that totally isolates the reader from access to relevant sources that could be used to fill gaps in one's current knowledge. In fact, for the skilled reader, knowing what one knows and knowing what one does not know are at the core of applying one's background knowledge (versus generating sensible inferences to fill gaps) in building a coherent mental model of a text.

Therefore, one design approach we are experimenting with will assess test takers' topical knowledge before they read texts. The background knowledge measure may sample information related to the topic of the target passages or information directly covered in the passages or both. While this background knowledge quiz will not contribute to the final score, it will provide an estimate of what students know before they read the texts. Thus it can be used to embellish interpretations about whether the reading score is reflective of true reading ability or background knowledge.

In addition to measuring background knowledge directly, we are also advocating building up test takers' knowledge over the course of a test. This growth can be achieved by providing introductory audio or multimedia on the topic of the assessment or by providing additional texts that supplement the content (e.g., provide cultural references for English language learners; provide texts that elaborate the history or context). Texts and other source materials are introduced in sequence, such that earlier texts are general and provide more of an overview of the topic, while subsequent texts dig deeper into the topic. By the end of the assessment, test takers have access to a progressively richer set of content materials they can draw upon to address a "big idea" question that serves as the main purpose of the assessment. This additional content does not totally resolve the issue of background knowledge, however. Knowing a topic

well is likely more advantageous than trying to learn the content with little or no relevant knowledge about the topic. Nonetheless, both are necessary skills of a proficient reader.

We are also interested in understanding the degree to which students take advantage of supplemental resources to further their understanding of key ideas discussed in text—an issue both of background knowledge and self-regulation (taken up in the next section). Supplemental resources may include video, audio assists, additional text, diagrams, and vocabulary definitions that are not designed to contribute to the final test score but can be used to interpret it. The supplemental materials are not strictly necessary to answer items correctly if one has the presumed general background knowledge (or makes good inferences), but they are included as resources for students who may be having difficulty with the content and vocabulary.

Implications for instruction. Including a measure of background knowledge in a reading test can potentially be useful for informing instruction. Prior to teaching a lesson on a particular subject, a background knowledge measure can be administered to students to determine what key words or concepts are unknown. With information at hand, teachers can focus on creating units that build up the vocabulary and context before they require their students to read unfamiliar text. Similarly, a background knowledge screening measure can be administered to students before they read texts and this information can be used to classify students into high and low knowledge groups. Students who are placed in the low knowledge group can be given special activities to build up their knowledge before reading and answering basic comprehension questions, while the students in the high knowledge group can be given more demanding comprehension tasks that require them to apply or transfer what they have read to new situations. Having a more nuanced measure of background knowledge in this way can alter the path of instruction for students' individual needs.

Another illustration of how background knowledge can be beneficial for instruction concerns the use of supplemental resources to build background knowledge as mentioned above. The extent to which a test taker uses available resources can be tracked and provided in a report to the teacher. More specifically, if students score low on the reading assessment, we can check to see if they at least tried (or were motivated enough) to use the supplemental resources to build up their background knowledge. Accordingly, the willingness to exert the effort to access supplemental information becomes another source of added information about the readers. It is an expectation that skilled readers will exert the appropriate level of effort to take advantage of

information provided for their use. Students can then get an item correct by applying a knowledge-based inference or by accessing supplemental resources. Similarly, they can get it wrong by ignoring the supplemental information or by using or interpreting the information inappropriately. The decision to use the resources is not scored as part of the total score, but the information available to an instructor about those decisions can be useful in understanding the reading behavior that underlies the score. If students are not using the supplemental resources, then teachers can discuss the potential value of solving problems by using external resources and encourage such adaptive behaviors in the future. Below we discuss this issue in more detail as it relates to self-regulatory and metacognitive strategies.

Self-Regulatory and Metacognitive Strategies and Behavior

In the current report, we treat metacognitive and self-regulatory strategies and behavior as complementary and synergistic processes that keep understanding and learning on track; that is, they encompass the processes of managing reading for understanding and learning. The term *metacognition* is generally used to refer to all the processes encompassing knowing about one's own cognition, that is, knowing what one knows in any domain (Schraw, 1998). Reading for understanding and learning is a specific application case of one's metacognition (e.g., Cromley & Azevedo, 2007; McNamara, 2007). Here it also refers to the process of monitoring one's understanding in light of potential comprehension difficulties. The term *self-regulation* is used to refer to the set of processes that adjust, alter, or maintain processing in light of intended learning or reading goals and progress toward those goals. Skilled readers notice when comprehension breaks down or whether goal-directed behavior is off track, and they then take action to repair gaps and misconceptions, fix errors, and problem solve by using a host of adaptive strategies and available resources. Metacognition and self-regulation are overlapping constructs that are typically discussed jointly in the reading literature (e.g., King, 2007; McNamara, 2007), and as such, we refer to both as *management processes* in the discussion below.

Managing understanding. The previous section on the role of background knowledge in assessment reminds us of how an individual difference such as background knowledge can alter the interpretation of a reading score. When leveraged appropriately, measuring background knowledge in the context of a reading assessment may improve score interpretation and potentially shed light on instruction. Despite the potential benefit of measuring background knowledge in the context of an assessment, other individual differences and skills must be

considered. For instance, merely possessing background knowledge doesn't necessarily mean that students will use that knowledge when necessary or use it appropriately (Reeves & Weisberg, 1994; Ross, 1989, 2008). A set of associated abilities for goal setting, planning, monitoring, error detection, updating, and repair are also critical for effective knowledge use and reading comprehension. This management of cognition and understanding is often collectively referred to as self-regulation and metacognition in the research literature (see Hacker et al., 2009; McKeown & Beck, 2009; Pressley, 2002; Schraw, 1998).

Viewed as a goal-directed activity, reading for understanding requires that the reader sets comprehension goals and subgoals, allocates and directs attention and effort toward achieving those goals, monitors coherence and progress toward the goals, and adjusts or adapts strategies if and as barriers or problems with comprehension are encountered (Linderholm & van den Broek, 2002; McCrudden, Magliano, & Schraw, 2010; van den Broek, Lorch, Linderholm, & Gustafson, 2001). The importance and complexity of the management of one's reading increases as reader goals become more complex (McCrudden, Magliano, & Schraw, 2011a, b; McCrudden & Schraw, 2007) and as the text complexity (see McNamara, Graesser, & Louwerse, 2012) and number of sources increases. Studying content sources for an exam or preparing to write a research synthesis are more complex goals than understanding the basic gist of a single text. Reading proficiency therefore requires sophisticated self-regulatory and metacognitive strategies (Eilers & Pinkley, 2006; Hacker et al., 2009). Yet the most one could say of traditional comprehension tests is that they implicitly demand some aspects of self-regulation and metacognition but make no explicit effort to measure or even describe how the assessment requires these abilities, and therefore, they provide no additional information for score interpretation. In most cases, the single text, discrete question format is underrepresenting the range of important applications of these skills that occur in nonassessment literacy situations.

Partial understanding and resiliency. We view this as an opportunity to leverage research in the areas of metacognition and self-regulation to make a test that is more sensitive to and aligned with the process and demands of reading in nontesting situations. For example, one key aspect of metacognition and self-regulatory behavior is the detection of errors in understanding (or diversion from a goal) and the disposition and skills to adapt one's reading to repair one's understanding (McNamara & Magliano, 2009a; Oakhill, Hartt, & Samols, 2005; Paris, Wasik, & Turner, 1991; Skarakis-Doyle & Dempsey, 2008). If a student gets an item

incorrect on a traditional comprehension test, it might mean the student has no understanding, or alternatively, it might mean the student had partial understanding but the test was not designed to capture it. If alternate ways of representing the information and feedback were used (for incorrect answers), it is possible the test taker could show evidence of either partial understanding or, more importantly, resiliency (e.g., getting an answer correct after receiving feedback). Implicit in this approach is the recognition that reading comprehension is more than the end product of understanding text, but also the set of core abilities that capture the *process* of reading (see Magliano, Millis, RSAT Development Team, Levinstein, & Boonthum, 2011; Millis & Magliano, 2012), including the ability to use strategies and resources to repair or augment impoverished representations of text.

Theoretical models and empirical findings converge in viewing reading comprehension as an iterative and strategic process that must be managed effectively (McNamara & Magliano, 2009b). A skilled reader is a resilient reader who can tolerate error, respond to feedback, recover from mistakes, and use alternate resources and representations to achieve coherence (Linderholm, Virtue, Tzeng, & van den Broek, 2004). In 21st century learning environments, these core abilities are more important as the sheer amount and quality of information available alters the nature of the construct (Coiro, 2009; Lawless, Goldman, Gomez, Manning, & Braasch, 2012; Metzger, 2007; Rouet, 2006; Rouet & Britt, 2011).

Ways to measure metacognitive and self-regulatory strategies. Designers of assessments of reading comprehension are recognizing the relative importance of metacognition and self-regulation on reading performance. To gain insight on these constructs, some international assessments of reading plans to use self-report measures of the construct (see Organisation for Economic Co-operation and Development (OECD), 2009a, engagement section) or alternatively supply students with scenarios and ask them to choose the best strategy or course of action to take given the circumstances (OECD, 2009b). While both of these approaches are useful for gaining insight on student metacognition and self-regulation, they are not measures embedded in the process of performing reading tasks. For instance, self-report measures are susceptible to bias and enumeration errors that plague such designs (Paulhus, 1991). Embedding explicit requirements for self-regulation and metacognition in assessment task designs would be a more direct technique. The current designs we are developing include indicators of self-regulation and metacognition in a number of different ways. These include

sequencing, the use of feedback, peer response tasks, and demands on resource allocation. These techniques are described more in detail below.

Adjusting to new information. In many traditional testing environments, reading comprehension is static: test takers are asked a series of questions about one text and then are required to move on to answer a series of questions about another unrelated text. However, in more authentic contexts, texts are not isolated entities but rather dynamic sources that relate to some larger goal. Over time, sources may also be modified and updated to reflect new evidence, new discoveries, or different views on a topic. This dynamic nature of text (and knowledge accumulation) places additional demands on readers as they need to update their old understanding with new evidence, or reinterpret text sources with a new perspective. This type of thinking is not only germane to the sciences, but its use is also demanded by the vast amount of information contained on the World Wide Web. Web sites vary in terms of currency, quality, and perspective and the most current and evidence-based sources must be identified, reconciled, and synthesized (Coiro, 2009; Lawless et al., 2012; Metzger, 2007).

One way we can instantiate a more dynamic form of reading is through sequencing the source information provided to gather evidence of test takers' ability to regulate and adapt their understanding. One approach to sequencing involves presenting test takers with different source materials in a predefined order. Sources presented earlier in the sequence could represent original or early views on a particular topic. Later sources then can be designed to reflect current or conflicting views, requiring that learners update their prior understanding in light of new evidence. The test taker's task is not only to understand the early sources in isolation, but also to reinterpret them in light of the new sources. Using this technique, texts are reread and reexamined under different lenses or perspectives (Schraw, Wade, & Kardash, 1993). Such rereading behavior may affect the accuracy of metacognitive judgments (Bråten, Gil, & Strømsø, 2011; Dunlosky & Rawson, 2005; Griffin, Wiley, & Thiede, 2008; Rouet & Britt, 2011).

Adapting to feedback. The sequencing task described above is designed to assess whether readers can adjust to new information. Another technique to gather more evidence on test takers' ability to adapt is feedback (Shute, 2007). In most testing situations, test takers do not know whether they answered the item correctly. Providing feedback in the form of hints or by partially completing the task after an error gives test takers another opportunity to display their resilience. For example, test takers can add to, edit, or delete part of their original answer. It also

provides more information on whether test takers have some knowledge and skill as opposed to none at all (Attali, 2011; Millis & Magliano, 2012). This feedback and scaffolding approach is useful for both promoting resilience and also for creating a more sensitive measure of reading skill.

Simulated peer collaboration. As the Common Core (NGACBP & CCSSO, 2010), the Partnership for 21st Century Skills (2004, 2008), and the modern workforce move toward collaborative learning environments, issues surrounding self-regulation become increasingly important. Not only do students have to be aware of their own learning, but they also have to manage peer interactions. Every day, students collaborate to solve problems and communicate feedback on fellow students' understanding. We believe these interactions are not only authentic, but also represent a great opportunity to assess meta or regulatory behaviors. One way to assess these behaviors is to provide students with a peer (or other) response task (see, PISA collaborative problem solving, OECD, 2009b). These tasks also call upon perspective taking and social modeling skills.

A peer response task set includes a text on a particular topic and an accompanying peer response that summarizes, comments on, interprets, or critiques the text. The simulated peer response is carefully constructed to reveal errors, misconceptions, overgeneralizations, or inappropriate elaborations the peer has about the text. The test takers may be asked to identify the errors in the peer response and then correct them. The focal point, with respect to metacognition and self-regulation, is the ability to know when comprehension breaks down or is faulty and the ability to correct and repair the faulty understanding. Adding the peer element to the assessment not only makes the construct more authentic to 21st century learning environments, but it also allows the test designer to target specific areas of the text that may cause problems for various readers. It also addresses the social dimension of the construct: in static text authors don't argue back, but in digital text peers do. Thus, in digital and workforce environments, students need to be able to discuss and debate text with their peers, but also reconcile perspectives as well as accommodate similarities and differences that accompany such collaboration.

Using resources. The peer response task is one way to build self-regulation (and simulated collaboration) into an assessment environment. However, skilled reading in authentic environments also demands that readers can use additional resources to help them solve key

problems. When skilled readers reach an impasse, they may seek help from outside sources to help them better understand what they read (as noted in previous section on background knowledge). For example, if students do not know the meaning of a word, they may look it up in a dictionary, or if they are unfamiliar with a concept, they may watch a video about it on YouTube. Naturally, skilled readers use the resources available to them to foster their own understanding. To simulate some of these ideas in an assessment context, we provide opportunities for students to access and use supplemental resources.

Implications for instruction. From a diagnostic and instructional perspective, a student's self-regulatory and metacognitive strategies and behavior are useful constructs to measure. If a test taker cannot adapt to task demands, respond to feedback, identify and fix comprehension breaks, and use available resources, then interventions that focus on building these core strategies are apt to be useful. These interventions are likely to be different from interventions that simply focus on a test taker's ability to extract the basic meaning from text. For example, if students are not good at accurately estimating their performance, students can be given repeated tests to help them more accurately calibrate their judgments of learning in light of their actual performance. Similarly, if tasks and items are structured in such a manner as to reveal student understanding with many different approaches, the tests are more likely to provide evidence of partial understanding and be more sensitive. This more nuanced information can potentially make it easier for teachers to triangulate the problem or identify the student's current level of development and adjust instruction accordingly.

In sum, we believe there is added value in measuring metacognitive and self-regulatory strategies and behaviors on a reading test from both an instructional and measurement perspective. Below we discuss specific ways students can regulate their understanding through the use of reading strategies and why including reading strategies in the context of a reading assessment is beneficial.

Reading Strategies

Reading strategies are strategic actions, behaviors, or habits that help readers form coherent models of text. Reading strategies may be used before, during, or after reading to help simplify, organize, or elaborate text in a more meaningful way for the reader. While their execution is often conscious, over time, reading strategies may become routine and resemble automatic performances. Although reading strategies may help improve the process of

understanding text, they are not synonymous with skilled reading. Skilled readers may not choose to use reading strategies if they deem it unnecessary.

Reading strategies and their function. Metacognition and self-regulation help to ensure the process of reading comprehension is on track (Hacker et al., 2009; McKeown & Beck, 2009). At a high level, metacognition and self-regulation represent global reading strategies that govern more specific actions for moving the process of comprehension forward. At a more specific level, there are a number of empirically validated reading strategies that can be helpful in supporting the construction of coherent models of text (see McNamara, 2007). Reading strategies include, but are not limited to visualization/imagery (Oakhill & Patel, 1991), paraphrasing (Fisk & Hurst, 2003), elaborating (Menke & Pressley, 1994), predicting (Afflerbach, 1990), self-explanation (McNamara, 2004), note taking (Faber, Morris, & Lieberman, 2000), summarization (Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005), previewing (Spires, Gallini, & Riggsbee, 1992), and the use of graphic organizers and text structure (Goldman & Rakestraw, 2000; Meyer & Wijekumar, 2007). Reading strategies such as these are often used by skilled readers to simplify, organize, restructure, remember, and embellish text. Not surprisingly, practitioners routinely incorporate reading strategies into English language arts curriculum and reading specialists habitually integrate them into reading interventions.

Despite the empirical support and routine use of reading strategies in the classroom, few assessments of reading comprehension require the use of a broad array of reading strategies in performing tasks on a test. In some cases reading strategies are considered ancillary processes that are not directly related to the construct; to measure them would be to detract from the efficiency of the test. In other cases, reading strategies are deemed as facilitative acts that occur during the process of reading rather than constitute the product of reading for understanding per se. While it is beyond the scope of this paper to argue whether reading strategies represent components of reading skill or whether they are allied processes, we argue that value is added by incorporating strategies in the design.

Improved measurement. At least two key reasons are evident to include reading strategies on a test of reading comprehension: improved measurement of the reading construct and positive wash back effects. Certain reading strategies represent opportunities to directly measure the five dimensions of reading literacy previously described. For instance, the verbal dimension can be measured by tasks designed to assess paraphrasing. In a paraphrasing item,

students are provided with a target sentence from a text and a set of options that correctly or incorrectly paraphrases the target sentence. Incorrect distracters may reorder the syntax in such a way as to change the meaning of the target sentence or may replace key words with incorrect synonyms. If test takers understand the target sentence, they should be able to demonstrate their understanding by selecting the paraphrase that preserves the meaning. These paraphrase items can be delivered in a peer response format in which test takers are asked to determine if a peer correctly put the target sentence in his/her own words. In this manner, paraphrasing not only represents a measureable strategy, but it also serves as a way to measure reading for understanding as defined in this framework.

In a similar vein, the discourse level of the model can be assessed by using summary and graphic organizer items. Both summary and graphic organizer items encourage the test taker to form a more global representation of how the key ideas connect to one another. This more global representation may better reflect discourse level understanding than more typical items that require test takers to comprehend isolated parts of the text. Similarly, a twist on the visualization strategy can be used to measure the discourse level as it pertains to narrative text. Narratives have sequences of events, and these events can be represented pictorially as a sequence. Items that target the discourse level for narrative text can provide test takers with a series of visuals that correctly or incorrectly relate to the events of the narrative. Distracters may depict incorrect events, or incorrect sequences of events, while the key depicts accurate events depicted in the proper sequence. By representing the story in a picture format, the item models the use of images to embellish the text representation. In short, strategies can represent both ways to approach a comprehension task, and in some cases, they may also serve as ways to measure the construct of reading comprehension.

Positive wash back effects. Summative tests are high stakes for those who are impacted by their consequences. Historically this has led to the unfortunate practice of “teaching to the test” (see Crocker, 2003; Volante, 2004). In dealing with this issue, we view reading strategies as one way for tests to have a positive impact on teaching. If reading strategies are included on the test, it may encourage their use in classrooms. While in many schools the pedagogy of reading strategies is routine, in others, they may not be taught as frequently. Inclusion signals importance, and it encourages teaching practices that are likely to be helpful for many students.

In cases where reading strategies are taught frequently, including them as items on the assessment makes the test more familiar and seamless with everyday instruction.

Implications for instruction. While creating positive wash back effects is one way an assessment can have an impact on instruction, there are others. For instance, providing an indicator of students' use of reading strategies helps teachers gain insight into how strategic their students are when processing text. Poor readers who are also identified by the assessment as not being adept with handling a host of strategies can potentially be given training in a wide variety of reading strategies. A more ambitious approach could tailor the strategy instruction to specific areas in which students have particular deficiencies. Of course, this more ambitious approach would be tempered by the quality of the instrument used to make such precise decisions. None the less, we argue that including reading strategies on an assessment is a positive first step.

A final note of caution. A primary challenge we face when incorporating reading strategies into an assessment is striking a balance between the desire to improve the process-oriented instructional value of the assessment and the requirement that we measure student's reading-based understanding (i.e., the product of reading). A simple test of this is whether high ability students would also perform well on tasks designed to measure reading strategies. For example, the demands for successfully responding to a graphic organizer item needs to be intuitively obvious to good readers, whether they were ever explicitly taught or exposed to graphic organizers before entering the test session. Next, we turn to the issue of motivation and engagement and how it impacts test design.

Motivation and Engagement

The terms *motivation* and *engagement* are used here to reflect a reader's willingness to expend the appropriate amount of effort necessary to form a coherent model of text and to complete tasks required by the assessment. Motivation is affected by a number of factors including the reader's interest in the topic, texts and tasks on an assessment, as well the value and stakes the reader ascribes to the assessment and its potential impact on his or her life.

The potential threat to validity. Both metacognition/self-regulation and reading strategies help students manage their comprehension under simple and complex text and task demands. This strategic behavior is often effortful and requires students to focus their attention and persist. If students are not giving their best effort, then one cannot be confident of the claim that a poor score reflects the lack of specific reading skills.[6] Low scores on a reading

comprehension test may be interpreted as reflecting low reading skill, or they may simply mean the test taker was not interested and did not try his or her best (see Braun et al., 2011). In any event, assessments could be designed to confront the empirical relationship between motivation, engagement, interest, and reading comprehension (De Naeghel, Van Keer, Vansteenkiste, & Rosseel, 2012; Guthrie & Davis, 2003; Guthrie et al., 1996; Guthrie & Wigfield, 2000).

Approaches for addressing concerns. There are many ways to address motivation and engagement concerns in a reading comprehension test. One approach that we have advocated is to include motivation in the reading ability construct: define skilled reading as, in part, the disposition to expend appropriate effort as the literacy environment demands. According to this view, if a student is unmotivated and decides to "blow off" the test, then this decision is an indicator of nonproficiency. However, this occurrence does not address the concern voiced earlier—is poor performance attributable to lack of skills or to lack of motivation? While we are somewhat supportive of this stance, we think it places perhaps too much responsibility on the shoulders of the student (and instructors who are charged with helping to build positive dispositions toward reading). This approach also ignores the responsibility of the test designer. Can test designers reasonably expect everyone to read and answer arbitrary questions about dull, dry, disconnected sequences of texts with equivalent engagement and enthusiasm?

Toward the other end of the spectrum, some advocate that tests, instruction, and curricula should be maximally engaging to students by including texts and topics that are highly relevant and interesting for students (see Moley, Bandre, & George, 2011). This position argues that motivation is a prime responsibly of test designers. If the assessment is uninteresting, then it is the fault of the test, not the test taker. This position does not address some construct-relevant issues, specifically, that as learners we are often called upon to read and learn about topics and texts that are uninteresting to us. And it is not only learners confined to school who find themselves coerced into dealing with uninteresting and unsavory text activities. Adult participation in society often demands processing of texts that a majority find uninteresting (e.g., tax forms, rental or credit card agreements). In any case, this approach is itself practically impossible. No known set of texts is universally interesting to every student; designing an assessment to match interesting texts to each test taker would be neither logistically sound nor feasible.

Our approach to the problem of motivation and engagement represents neither extreme view, but a blend of the strengths of both approaches. GISA is designed to support motivation by presenting texts and tasks in the context of a goal-oriented scenario that is more closely aligned with literacy activities that occur outside an assessment environment. As in all simulated or virtual environments, there are simplifications that diverge from authenticity (Petraglia, 1998).[7] Still, GISA goal-oriented scenarios are an attempt to require less suspension of disbelief by learners than traditional comprehension assessments. GISA designs meet the student halfway, not by trying to match to or generate interest in topical content, but rather by engaging students in applying their reading and problem solving skills in a goal directed way, as they might in nonassessment literacy contexts. At the same time, GISA is also designed to provide indications of whether students are in fact putting honest effort into completing the test.

While test design plays an important part in student motivation, teachers also play a key role. Teachers have an intuitive sense of how engaging the materials and activities are for their students, and they can take this information into consideration when designing their instruction. However, with some assessment indicators of student motivation at hand, teachers are in a better position to interpret and adjust instruction than if no measures of motivation were available (see Table 2). Before we describe our approach, we discuss student motivation and engagement in the context of the GISA design.

GISA design features. Student motivation involves the disposition to persist in the face of difficulty and the inclination to expend the effort as required by literacy tasks (which can be quite cognitively demanding). This notion of motivation, embedded in the construct, can help guide design decisions and features that can enhance students' opportunity to show their engagement during testing. One key design feature in this regard is the use of scenarios that are more closely aligned with authentic purposes for reading (Bennett & Gitomer, 2009; Sheehan & O'Reilly, 2012). Other features are used to reduce test anxiety and increase confidence early in the test session. In some scenarios, for example, students can view a video or multimedia presentation at the beginning of the assessment. This technique is designed to help build and activate relevant background knowledge, as well as help to reduce test anxiety. In other cases, complex performances are broken down into more manageable units; this allows lower ability students a better chance of success, and as a result, it potentially builds confidence and self-efficacy and moderates the students' level of effort.

Another way GISA deals with the issues of motivation and engagement is to embed items into the assessment that signal whether students are trying. These items include, but are not limited to, embedding constructed responses in the test and examining the frequency of insincere responses (e.g., "I don't care"), asking test takers to rate their understanding of concepts not included in the assessment, examining test takers' usage of help functions and other resources, and asking students to rate their interest in the text and tasks directly. These indexes of motivation can be instrumental in helping interpret the scores on the reading test. If students score low on a reading test, then the case could be that they lack specific reading skills (as the test scores suggest) or that they did not expend sufficient effort to score better. Note that the latter does not address the issue of whether students possess skills, but it is one step closer and, therefore, potentially of value to instructors who are in the position to follow up.

Implications for instruction. Knowing more about the level of student engagement has potential added value because it suggests different courses of instruction. Students who score low on a reading test because they have low ability but are motivated are likely to benefit from reading strategy training that focuses on both foundational and comprehension skill development (see McNamara, 2007). On the other hand, students who score low because of low motivation might require a different approach to boosting their performance (Guthrie & Davis, 2003). In other words, instructional differentiation can depend upon and be aided by the more nuanced information provided by an assessment. In sum, supporting and measuring effort and motivation can potentially improve the design, interpretation, and use of the reading scores. In the next section, we describe a technique for integrating performance moderators and authentic purposes for reading into test design- scenario-based assessment.

A Case for Scenario-Based Assessment

In the beginning of this paper and elsewhere (Bennett, 2011a, 2011b; Bennett & Gitomer, 2009; O'Reilly & Sheehan, 2009; Sabatini & O'Reilly, 2013; Sabatini et al., 2013; Sheehan, & O'Reilly, 2012), we advocate for a different kind of reading comprehension assessment. This new reading assessment is intended to broaden the construct of reading comprehension and promote deeper processing through the use of purpose-driven reading. In designing such an assessment, we introduce some new challenges that impact how tests are designed and interpreted. These challenges stem from incorporating a host of performance moderators into the test design. In the space above, we describe how the GISA assessment design accounts for these

performance moderators and argue why they are important to consider in testing. In this section, we describe the notion of scenario-based assessment and how it is a useful technique for dealing with purpose-driven assessment and the various performance moderators. During this discussion, we also outline the various aspects that constitute a scenario at different levels of sophistication. Table 3 outlines some of the key features of scenario-based assessment and their potential impact on measurement and instruction.

Background and Context

Many existing summative assessments of reading comprehension are often designed to optimize the information about students' proficiency on a unidimensional scale. Such an approach frequently involves making decisions that favor item discrimination, efficiency, and cost. An assessment that uses the fewest number of items to accurately measure test takers' reading ability in the shortest time possible has been considered ideal. While this approach is economical for administration and scoring purposes, unintended consequences may negatively impact both score interpretation and instructional practice. For instance, in many traditional reading comprehension assessments, passages are presented one at a time in a decontextualized manner (e.g., Ozuru, Rowe, O'Reilly, & McNamara, 2008). There is little or no purpose for reading other than to answer multiple choice questions correctly (Rupp, Ferne, & Choi, 2006). Because the passages are presented at the same time as the questions, students use and are often taught test-taking strategies that encourage locate and retrieve behaviors rather than strategies that require students to form a coherent model or to synthesize and apply information to solve real problems (see Cordon & Day, 1996; Farr, Pritchard, & Smitten, 1990).

In traditional testing environments, the perspective-taking of the student is dominated by a theory of mind of the assessment designer. Put yourself in the head of the test maker, and you score well. While reading in testing situations does in fact represent an authentic and a real purpose for reading (Farr et al., 1990), it grossly underrepresents the full range of reading situations required for college readiness, the workforce, and effective citizenship (McCrudden et al., 2011a,b; O'Reilly & Sheehan, 2009; Sabatini, Albro, & O'Reilly, 2012; Sabatini, O'Reilly, & Albro, 2012). Assessing reading comprehension in a traditional and decontextualized manner does not guarantee the scores will apply to other purposeful reading activities, nor does it encourage processing strategies that will ensure flexibility or transfer to situations outside the testing window.

Table 3

Key Features of Scenario-Based Assessment and Implications for Score Interpretation Value and Instruction

Feature	Design and measurement goals	Score interpretation value	Implications for instruction
Multiple test forms	Increase construct coverage	Increase the generalizability of GISA	Allow for pre- / postintervention designs, growth and progress monitoring
Purpose for reading	Provides readers with a standard of coherence	Increases the external and ecological validity	To simulate valid literacy contexts To promote interest and engagement
Multiple sources	To create an assessment narrative and to promote coherence among sources	Increases generalizability and reliability by increasing the number of passages and questions	Promotes integration and synthesis Can be used to promote cultural diversity and appreciation for different viewpoints
Independent performances	To determine if students can perform a task independently	Better measurement at the upper end of the distribution	Can identify students for possible advanced tasks
Scaffolded performances	To determine partial knowledge and partial skill development	Better measurement at the lower end of the distribution	Used to help triangulate strengths and weaknesses in particular subskills
Peer assessment	To promote collaboration and collective understanding	Expands the construct Can be used to increase discrimination at the upper and lower ends of the distribution	Used to help stimulate discussion and debate Support peer mentoring
Performance moderators	To account for factors that impact reading ability	Improve the interpretation of scores Increase the validity of the assessments	To suggest different modes of instruction

Note. GISA = Global Integrated Scenario-Based Assessment.

Six Features Defining Scenario-Based Assessment

Recognizing these limitations, we propose a different kind of reading assessment that focuses on and supports reading behaviors that are valued by research, effective teaching practice, and workforce readiness. By using a broader assessment design space, we can provide scores that are potentially more reflective of the situations we want and encourage students to read within them. Students should be problem solvers and decision makers (NGACBP & CCSSO, 2010), critical evaluators of sources and evidence (Graesser et al., 2007; Lawless et al., 2012; Metzger, 2007), and collaborators in the co-construction of knowledge (Partnership for 21st Century Skills, 2004, 2008), not masters of test taking. Below we outline six features that collectively define what we mean by scenario-based assessment and how these features may potentially impact testing. In short, scenario-based assessment should be designed to provide a standard of coherence, promote coherence among a collection of materials, gather more information about test takers, promote collaboration, simulate valid literacy contexts of use, and promote interest and engagement.

1. To provide a purpose for reading: establishing a standard of coherence. People read for a variety of reasons, ranging from the relatively simple (e.g., to find a date of an historical event) to the relatively complex (e.g., to write a report recommending the best form of alternative energy for a particular community). Not surprisingly, the level of comprehension demanded by the reader varies greatly depending upon the goals of reading (Carver, 1997; van den Broek et al., 2001; van den Broek, Risden, & Husebye-Hartman, 1995). When looking for a date in a textbook, readers scan for numbers; they do not read the entire text for deep meaning. Conversely, when writing a report on the best form of alternative energy for a particular community, readers need to extract basic meaning, evaluate sources, integrate and synthesize information, make a cost benefit analysis, then make a decision. Clearly, the purpose for reading dramatically changes the level of processing, effort, and attention readers need to regulate (McCrudden et al., 2010, 2011a,b; Sheehan & O'Reilly, 2012). The reading purpose defines what is and what is not important to attend to and how to adjust resources accordingly. This fluctuating metric and decision making process is referred to as the *standard of coherence* in the research literature (Linderholm et al., 2004; van den Broek et al., 1995, 2001). When a low standard of coherence is chosen, gaps in understanding are deemed tolerable to the reader, whereas readers who adopt a high standard of coherence must expend additional effort to deepen

and embellish understanding to ensure that intricate details of a text are integrated into a coherent situation model and to monitor and repair breaks in understanding.

In a traditional testing situation, the standard of coherence is relatively unspecified because there is no targeted purpose for reading other than to answer questions correctly (Rupp et al., 2006). That is, there is no metric for adjusting, guiding, and evaluating how the materials the reader is provided with should be processed. At best, the standard of coherence is dictated locally by each question. While a local purpose for reading is valid, it underrepresents the range of reading situations. Given this discussion, probably the most important function of a scenario is to provide test takers with a purpose for reading. The purpose for reading helps define the standard of coherence readers should adopt as they engage with the reading materials. It becomes the standard for which all texts and materials are judged as relevant and how much and what type of processing is required. To ensure generalizability, we developed multiple GISA forms. A form contains a set of texts and tasks that are organized by a specific purpose for reading (e.g., to modify a wiki; to make a presentation). These different reading purposes provide a variety of contexts in which readers can engage. Varying the reading purposes in this way creates and encourages a broader array of reading contexts than would be expected by a traditional reading assessment.

2. To promote coherence among a collection of materials: the assessment narrative. In traditional reading assessments, test takers are provided with a collection of texts and questions. While the reading assessments include a variety of texts, topics, and genres, often no connection is stated or demands placed on the test taker to integrate them.[8] Students are expected to read a text, answer the accompanying questions, forget it, and move on to the next unrelated passage. In today's digital world, this reading situation is not only artificial, but it also supports compartmentalized thinking. Literacy skills in the 21st century require readers to synthesize, evaluate, and integrate diverse sources (Britt & Rouet, 2012; Graesser et al., 2007; Lawless et al., 2012; Metzger, 2007; Partnership for 21st Century Skills, 2008). This diversity not only includes the variety of text types used in the assessment, but also the range of different perspectives students need to understand, manage, and integrate.

Rather than treating each source in an assessment as a discrete and independent entity, we advocate creating coherence among disparate sources through the use of scenarios. The scenario defines the purpose for reading, the standard of coherence, and the "glue" to connect seemingly

disparate content. While answers to some questions are found in a single text, other questions demand the integration of multiple texts to corroborate claims, verify evidence, evaluate content, and present a balanced and synthesized view of uncertain issues (Lawless et al., 2012). Providing test takers with a purpose for reading sets the stage for this type of thinking, but it does not guarantee it will spontaneously occur, particularly with developing students (Boveri, Millis, Wiemer, Sabatini, & O'Reilly, 2012; Britt & Rouet, 2012). Therefore, it adds value in understanding reader behaviors.

To ensure the assessment captures the type of processing intended, another function of the scenario is to help model synthesis in a structured way. That is, in addition to providing test takers with a general purpose for reading at the beginning of the assessment, other techniques are used to link sources throughout the course of the assessment. Before and after each source is presented to the test taker, new information in the form of a scenario can be given that provides a reason why the test taker is receiving the new source and what he or she is supposed to do with it. This interspersed organization of a scenario serves both as a model for setting subgoals within a larger purpose for reading, as well as providing explanations for the ongoing flow of tasks and activities. In this way, the scenario effectively functions as an assessment narrative that builds, connects, and models construct-relevant processing over the duration of the assessment. In tandem, the purpose of the assessment represents the destination, while the assessment narrative provides the road map to get there.

3. To gain more information about test takers: triangulating strengths and weaknesses. The Race to the Top and Common Core State Standards are designed to promote college and career readiness (NGACBP & CCSSO, 2010; U.S. Department of Education, 2009). In doing so, they effectively raise the bar for achievement. The positive side of this reform is the promise that students will be better prepared to compete in the 21st century economy. The potential concern of this reform is the lack of assessment information that triangulates where students stand in their development. For instance, many traditional summative reading comprehension assessments provide a single score that can be used to measure proficiency, but they provide little or no information on what parts of the larger tasks students can do. One unintended consequence of such a test is evident: it is easier to make inferences about what a student cannot do than what he or she is capable of doing. Complex tasks are either correct or

incorrect, and information on component and allied processing is often overlooked or lost in the integrated performance (see Attali, 2011).

The real conundrum then is how to design a test that encourages higher level processing (as the Common Core demands), while simultaneously informing student development and instruction (Gordon Commission, 2013). Another way to frame the problem is to ask: how can we measure independent and new complex performances in light of the fact that many students are not at a point in development to handle the complexity demanded by these new tasks? While no one perfect solution to this problem exists, the use of well-designed scenario-based assessments can help. Rather than requiring students to carry out complex tasks in one step (i.e., measuring independent performances only), we can use *scaffolding techniques* to help break down the larger more complex task into several smaller and more manageable steps. The scaffolding approach sequences tasks in a particular way so as to reduce load, model strategic thinking, and gather more information on what students can do along the way. Using this procedure, it is possible to identify students who may in fact be able to do parts of the complex task but not the entire task independently. If only the independent complex task were provided, one might falsely conclude the test taker did not have any of the skills that were prerequisite to the desired complex performance (O'Reilly, Sabatini, Bruce, & Halderman, 2012). By providing information on what parts of the task a student can and cannot handle, instructional decisions can be more targeted and focused on the student's individual needs.

While this approach is promising, there is a potential problem: if tasks and activities are always scaffolded, sequenced, and broken down, then the desired outcome of independent performance is not realized. To help address this issue, tasks can be designed to elicit complex and independent performances first, before any scaffolding, sequencing, and task breakdown occurs. By asking for the independent performance first, the test can determine which students have mastered the knowledge, skills, and abilities. After the independent attempt has been executed, follow-up tasks both probe and support the steps needed to achieve the complex performance (O'Reilly, Sabatini, Bruce, Pillarisetti, et al., 2012). This process is analogous to the "show your work" concept in math with the exception that the steps to solve the problem are aided by the design of the test.

4. To promote collaboration: distributed and collective understanding. The Common Core State Standards highlight the importance of communication in the listening and speaking

section of the English language arts standards (NGACBP & CCSSO, 2010). Students need to be able to understand, collaborate, and communicate what they have learned to targeted audiences. Similarly, in the 21st century economy, it is not sufficient to simply understand what one reads in isolation, but also to know how to interact and communicate to others. Many work environments are now team based and people work together to collectively solve problems. These environments require people to distribute responsibility, navigate different perspectives, resolve misunderstandings, and propose alternative solutions as they collaborate and communicate on larger projects. These environments include both face-to-face and digitally mediated communications. In short, modern reading environments demand the social skills required to effectively interact with one's associates in language and text communications (Partnership for 21st Century Skills, 2004, 2008).

Recognizing the social nature of reading, another important function of a scenario is to simulate and provide a context for peer interaction. In GISA, the peer interaction is simulated in the context of the scenario and the assessment narrative. As new sources and events unfold, simulated peers comment on sources, suggest new courses of action, make mistakes, go off topic, provide new evidence, or help adjust a test taker's understanding. The test taker's task is to respond to the simulated peers by using textual evidence to support the response. For instance, a simulated peer might make a comment in a threaded discussion that misrepresents what the author of a text is intending. The test taker is then asked to determine if the information in the peer response is correct and to write a response to correct it if necessary. The peer format of the test is designed to target specific areas of the text that may cause trouble for some students, to support evidence-based reasoning, and to support socially distributed processing. An indirect intention of the peer response format is also to help improve test-taker motivation and engagement by simulating the type of interaction that occurs in effective classroom discussions (Guthrie & Wigfield, 2000). Note that a primary technique we have employed in as assessment narrative are *student peers*, but other peers or agents such as a simulated teacher, expert, or other relevant persons can also be used to simulate social interactions as the scenario demands.

5. To simulate valid literacy contexts of use/practice: assess what we want students to be able to do. As mentioned earlier, reading in the context of an assessment is a valid purpose for reading (Farr et al., 1990). However, it is not the only purpose for reading, nor does it universally generalize to all reading situations in the real world. To widen the band of assessment

activities, scenarios are designed to represent a wide range of rich learning environments. As represented in Table 3, scenarios are organized into different forms that vary in terms of reading purpose, topic, and expected outcome. By varying these dimensions across forms, construct coverage is not only broadened, but the ecological validity is also potentially expanded. In this way, the scenarios are intended to take a slice of the broad array of curricular and extracurricular activities in which we want students to engage. These activities include, but are not limited to, making decisions/recommendations, solving problems, evaluating evidence, providing feedback, creating a Web site, modifying a wiki, posting appropriate and informed responses on a threaded discussion, writing a report, or creating an informational booklet.

These types of reading contexts are designed to promote key reading skills such as critical thinking, evaluation, source integration, synthesis, learning, knowledge integration, strategy use, and self-regulation. By strategically crafting the scenario, these processes are demanded, encouraged, and supported. It remains an empirical question as to whether these more specific purposes for reading will generalize. However, with careful design (e.g., measuring background knowledge) and sampling from a range of topics, scenarios can approximate the situations we want students to read in. This claim can be investigated empirically to evaluate the generalizability of the assessment.

6. To promote interest, motivation, and engagement. As stated earlier in this framework, motivation and student engagement represent a distributed responsibility between the test taker and the test designer. Test takers are expected to expend their best effort when taking an assessment because effort is construct-relevant—it helps to partly define skilled reading. Conversely, test designers have a responsibility to create testing situations that provide affordances for test takers to demonstrate their skills. The scenario is one vehicle that can be used to help promote interest and engagement by supplying test takers with meaningful purposes for reading a collection of diverse materials. Students display more motivation if topics are goal-driven, age appropriate, and relevant to the issues that concern them (see Guthrie & Davis, 2003).

In the context of a scenario, the sequencing, scaffolding, and structuring of the test may also seem more engaging as students are given more opportunities to display their partial knowledge and skills. Similarly, the use of peer interactions may promote engagement as they resemble more familiar classroom and extracurricular environments. Collectively, the purpose for reading, the increased opportunities for success, scaffolding, and the use of peer interaction

are all designed to allow students to demonstrate their best effort. While promoting engagement is not the primary function of using scenario-based assessment, it does represent an indirect benefit we hope is achieved.

Summary and Conclusions

This paper represents the third installment of the RfU assessment framework. The first installment provided the rationale for a new generation of reading assessments and described six principles for assessment design (Sabatini & O'Reilly, 2013). The second installment built upon the first by providing a definition of reading for understanding, the key constructs to be assessed, a position on reading development, and an overview of two types of assessments (Sabatini et al., 2013). The third part of the framework, provided in this installment, introduces a set of performance moderators and describes how they impact assessment design. These moderators—(a) background and prior knowledge, (b) metacognitive and self-regulatory strategies and behavior, (c) reading strategies, and (d) student motivation and engagement—not only impact the interpretation of reading comprehension scores, but also potentially add value for instruction. However with any new complex design, there are always challenges in how to implement innovations in a feasible and accessible way. Accordingly, we argued for a richer, more meaningful assessment design that uses scenarios to present the purpose for reading a collection of thematically related materials. We also outlined the various facets of the scenario and the potential added value to measurement and instruction. Interested readers are welcome to visit our Web site to learn more about GISA and see some released screen shots and item descriptions.[9]

It is perhaps important to reiterate, in closing, that the innovations and approaches described here are both enabled and necessitated by rapid changes in the technological and social environment of literacy that emerged in the late 20th century and continues unabated in society in the 21st century. In the 1980s, only a handful of futurists could have predicted the pervasive, transformational spread of personal information-communication technologies and the organization and instantaneous accessibility of the world's knowledge at every individual's fingertips. Despite richer audio and visual capabilities stemming from these technologies, reading and writing literacy skills continue to be primary engines of human capital and personal growth in this brave new environment. Greater sophistication in how we measure this critical human capability and the factors that moderate performance are warranted, as well as enhancing the utility of such measurement to aide in helping individuals achieve ever higher levels of

proficiency. While we are only on the threshold of experimenting with innovations that apply the science of the cognitive of reading to assessment designs, we see no diminishing of the necessity, potential utility, and capability to do an increasingly better job of measuring and supporting individuals in achieving reading proficiency.

References

Adams, B., Bell, L., & Perfetti, C. (1995). A trading relationship between reading skill and domain knowledge in children's text comprehension. *Discourse Processes, 20,* 307–323.

Afflerbach, P. (1990). The influence of prior knowledge and text genre on readers' prediction strategies. *Journal of Reading Behavior*, *22*, 131–48.

Alexander, P., Sperl, C., Buehl, M., & Chiu, S. (2004). Modeling domain learning: Profiles from the field of special education. *Journal of Educational Psychology, 96*, 545–557.

Attali, Y. (2011). Immediate feedback and opportunity to revise answers: Application of a graded response IRT model. *Applied Psychological Measurement, 35*, 472–479.

Bennett, R. E. (2011a, June). *Theory of action and educational assessment*. Paper presented at the National Conference on Student Assessment, Orlando, FL.

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

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.

Boveri, D., Millis, K., Wiemer, K., Sabatini, J., & O'Reilly, T., (2012, July). *Assessing comprehension: The effects of multiple-documents and scenarios.* Paper presented at the Society for Text and Discourse, Montreal, QC.

Bråten, I., Gil, L., & Strømsø, H. I. (2011). The role of different task instructions and reader characteristics when learning from multiple expository texts (pp. 95–122). In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text.* Greenwich, CT: Information Age.

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An empirical study to examine whether monetary incentives improve 12th grade NAEP reading performance. *Teachers College Record, 113*, 2309–2344.

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

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

Carver, R. P. (1997). Reading for one second, one minute, or one year from the perspective of rauding theory. *Scientific Studies of Reading, 1*, 3–43.

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.

Cordon, L. A., & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology, 88*, 288–295.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice, 22*, 5–11.

Cromley, J., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 99*, 311–325.

De Naeghel, J., Van Keer, H., Vansteenkiste, M., & Rosseel, Y. (2012). The relation between elementary students' recreational and academic reading motivation, reading frequency, engagement, and comprehension: A self-determination theory perspective. *Journal of Educational Psychology*, *104*, 1006–1021.

Dochy, F., Segers, M., & Buehl, M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research, 69*, 145–186.

Dunlosky, J., & K. A. Rawson (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes, 40*, 37–55.

Educational Testing Service (2013). *Reading for understanding*. Retrieved from http://www.ets.org/research/topics/reading_for_understanding/

Eilers, L. H., & Pinkley, C. (2006). Metacognitive strategies help students to comprehend all text. *Reading Improvement, 43*, 13–29.

Faber, J. E., Morris, J. D., & Lieberman, M. G. (2000). The effect of note taking on ninth grade students' comprehension. *Reading Psychology, 21*, 257–270.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27,* 204–226.

Fincher-Kiefer, R., Post, T., Greene, T., & Voss, J. (1988). On the role of prior knowledge and task demands in the processing of text. *Journal of Memory and Language, 27,* 416–428.

Fisk, C., & Hurst, C. B. (2003). Paraphrasing for comprehension. *Reading Teacher, 57,* 182–185.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley. S. (2005). Summary Street: Computer support for comprehension and writing. *Journal of Educational Computing Research, 33,* 53–80.

Goldman, S., & Rakestraw, J. (2000). Structural aspects of constructing meaning from text. In M. Kamil, P. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 311–335). Mahwah, NJ: Erlbaum.

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

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, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory and Cognition, 36,* 93–103.

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

Guthrie, J. T., McGough, K., Bennett, L., & Rice, M. E. (1996). Concept-oriented reading instruction: An integrated curriculum to develop motivations and strategies for reading. In L. Baker, P. Afflerbach, & D. Reinking (Eds.), *Developing engaged readers in school and home communities* (pp. 165–190). Hillsdale, NJ: Erlbaum.

Guthrie, J. T., & Wigfield, A. (2000). Engagement and motivation in reading. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 403–422). Mahwah, NJ: Erlbaum.

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

Hambrick, D. Z., & Engle, R. W. (2002). Effects of domain knowledge, working memory capacity, and age on cognitive performance: An investigation of the knowledge-is-power hypothesis. *Cognitive Psychology, 44*, 339–387.

Institute of Education Sciences. (2009). *Request for applications: Reading for Understanding research initiative* (CFDA Number: 84.305F). Washington, DC: U. S. Department of Education. Retrieved from http://ies.ed.gov/funding/pdf/2010_84305F.pdf

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

Katz, S., &. Lautenschlager, G. (2001). The contribution of passage and no-passage factors to item performance on the SAT reading task. *Educational Assessment, 7*, 165–176.

King, A. (2007). *Beyond literal comprehension: A strategy to promote deep understanding of text*. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 267–290). Mahwah, NJ: Erlbaum.

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

Kucer, S. B. (2011). Going beyond the author: What retellings tell us about comprehending narrative and expository texts. *Literacy, 45*, 62–69.

Kyllonen, P. (2013). *Cognitive intellectual abilities.* Manuscript submitted for publication.

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.

Linderholm, T., & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology, 94*, 778–784.

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

Logan, S., Medford, E., & Hughes, N. (2011). The importance of intrinsic motivation for high and low ability readers' reading comprehension performance. *Learning and Individual Differences, 21,* 124–128.

Magliano, J. P., Millis, K. K., RSAT Development Team, Levinstein, I., & Boonthum, C. (2011). Assessing comprehension during reading with the reading strategy assessment tool (RSAT). *Metacognition and Learning, 6,* 131–154.

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

McCrudden, M. T., Magliano, J. P., & Schraw, G. (2010). Exploring how relevance instructions affect personal reading intentions, reading goals and text processing: A mixed methods study. *Contemporary Educational Psychology, 35,* 229–241.

McCrudden, M. T., Magliano, J. P., & Schraw, G. (2011a). Relevance in text comprehension (pp. 1–18). In M. T., McCrudden, J. P., Magliano, &, G. Schraw (Eds.), *Text relevance and learning from text.* Greenwich, CT: Information Age.

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

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

McKeown, M. G., & Beck, I. L. (2009). The role of metacognition in understanding and supporting reading comprehension. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 7–25). Mahwah, NJ: Erlbaum.

McNamara, D. S. (1997). Comprehension skill: A knowledge-based account. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 508–513). Hillsdale, NJ: Erlbaum.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55,* 51–62.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38,* 1–30.

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

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., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability,* (pp. 89–118). Lanham, MD: Rowman & Littlefield.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22,* 247–288.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14,* 1–43.

McNamara, D. S., & Magliano, J. P. (2009a). Self-explanation and metacognition: The dynamics of reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 60–81). Mahwah, NJ: Erlbaum.

McNamara, D. S., & Magliano, J. P. (2009b). Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 297–384). New York, NY: Elsevier Science.

Menke, D. J., & Pressley, M. (1994). Elaborative interrogation: Using "why" questions to enhance the learning from text. *Journal of Reading, 37,* 642–645.

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.

Meyer, B., & Wijekumar, K. (2007). A Web-based tutoring system for the structure strategy: Theoretical background, design, and findings. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theory, interventions, and technologies* (pp. 347–374). Mahwah, NJ: Erlbaum.

Millis, K., & Magliano, J. (2012). Assessing comprehension processes during reading. In J. Sabatini, T. O'Reilly, & E. Albro (Eds.), *Reaching an understanding: Innovations in how we view reading assessment* (pp. 35–53). Lanham, MD: Rowman & Littlefield Education.

Moley, P. F., Bandre, P. E., & George, J. E. (2011). Moving beyond readability: Considering choice, motivation and learner engagement. *Theory into Practice*, *50*, 247–253.

Murphy, K., & Alexander, P. (2002). What counts? The predictive powers of subject-matter knowledge, strategic processing, and interest in domain-specific performance. *Journal of Experimental Education, 70,* 197–214.

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

Oakhill, J., Hartt, J., & Samols, D. (2005). Levels of comprehension monitoring and working memory in good and poor comprehenders. *Reading and Writing: An Interdisciplinary Journal, 18,* 657–686.

Oakhill, J., & Patel, S. (1991). Can imagery training help children who have comprehension problems? *Journal of Research in Reading, 14,* 106–115.

O'Reilly, T., & McNamara, D. S. (2007a). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43*, 121–152.

O'Reilly, T., & McNamara, D. S. (2007b). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional “high stakes” measures of high school students’ science achievement. *American Educational Research Journal, 44,* 161–196.

O'Reilly, T., Sabatini, J., Bruce, K., & Halderman, L. (2012, July). *Does length matter? The relative contribution of local and global understanding on students’ ability to write summaries*. Paper presented at the Nineteenth Annual Meeting of the Society for the Scientific Study of Reading, Montreal, QC.

O’Reilly, T., Sabatini, J., Bruce, K., Pillarisetti, S., & McCormick, C. (2012). Middle school reading assessment: Measuring what matters under an RTI framework. *Reading Psychology Special Issue: Response to Intervention, 3*3(1–2), 162–189.

O'Reilly, T., & Sheehan, K. M. (2009). *Cognitively Based Assessment of, for and as Learning: A 21st century approach for assessing reading competency* (Research Memorandum No. RM-09-04). Princeton, NJ: Educational Testing Service**.**

Organisation for Economic Co-operation and Development (2009a). *PIAAC literacy: A conceptual framework.* Paris, France: Author. 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.* Paris, France: Author. Retrieved from http://www.oecd.org/document/44/0,3746,en_2649_35845621_44455276_1_1_1_1,00.html

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction, 25*, 399–438.

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19,* 228–242.

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.

Paris, S. G., Wasik, B., Turner, J. C. (1991). The development of strategic readers. In R. Barr, M. Kamil, P. B. Mosenthal (Eds.), *Handbook of reading research* (Vol. 2, pp. 609–640). Hillsdale, NJ: Erlbaum.

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

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

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). New York, NY: Academic Press.

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

Petraglia, J. (1998). The real world on a short leash: The (mis)application of constructivism to the design of educational technology. *Educational Technology Research and Development, 46*, 53–65.

Piaget, J. (1957). *Construction of reality in the child.* London, UK: Routledge & Kegan Paul.

Pressley, M. (2002). Metacognition and self-regulated comprehension. In A. E. Farstrup & S. Samuels (Eds.), *What research has to say about reading instruction* (pp. 291–309). Newark, DE: International Reading Association.

Recht, D., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology, 80*, 16–20.

Reeves, L., & Weisberg, R. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin, 115*, 381–400.

Ross, B. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning Memory, and Cognition, 15*, 456–468.

Ross, B. (2008). Category learning: Learning to access and use relevant knowledge. In M. A. Gluck, J. R. Anderson, & S. M. Kosslyn (Eds.), *Memory and mind: A Festschrift for Gordon H. Bower* (pp. 229–246). Mahwah, NJ: Erlbaum.

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 processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Relevance instructions and goal-focusing in text learning* (pp. 19–52). Greenwich, CT: Information Age.

Rumelhart, D. E. (1980). *Schemata: The building blocks of cognition*. Hillsdale, NJ: Erlbaum.

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.

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

Sabatini, J., Bruce, K., & Steinberg, J. (2013). *SARA reading components tests, RISE form: Test design and technical adequacy* (Research Report No. RR-13-08). Princeton, NJ: Educational Testing Service.

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

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

Sabatini, J., O'Reilly, T., & Deane, P. (2013). *Preliminary reading literacy assessment framework: Foundation and rationale for assessment and system design* (Research Report No. RR-30-30). Princeton, NJ: Educational Testing Service**.**

Schank, R. C. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Erlbaum.

Schneider, W., Körkel, J., & Weinert, F. E. **(**1989). Domain-specific knowledge and memory performance: A comparison of high- and low aptitude children. *Journal of Educational Psychology, 81*, 306–12.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26*(1–2), 113–125.

Schraw, G., Wade, S. E., & Kardash, C. A. (1993). Interactive effects of text-based and task-based importance on learning from text. *Journal of Educational Psychology, 85*, 652–661.

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

Sheehan, K., & O'Reilly, T. (2012). The case for scenario-based assessments of reading competency. In J. Sabatini, T. O'Reilly, & E. Albro (Eds.), *Reaching an understanding: Innovations in how we view reading assessment* (pp. 19–33). Lanham, MD: Rowman & Littlefield Education.

Shute, V. (2007). *Focus on formative feedback* (Research Report No. RR 07-11). Princeton, NJ: Educational Testing Service.

Skarakis-Doyle, E., & Dempsey, L. (2008). The detection and monitoring of comprehension errors by preschool children with and without language impairment. *Journal of Speech, Language, and Hearing Research, 51*, 1227–1243.

Spilich, G., Vesonder, G., Chiesi, H., & Voss, J., (1979). Text processing of domain related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18*, 275–290.

Spires, H. A., Gallini, J., & Riggsbee, J. (1992). Effects of schema-based and text structure-based cues on expository prose comprehension in fourth graders. *Journal of Experimental Education, 60*, 307–320.

Thompson, R., & Zamboanga, B. (2004). Academic aptitude and prior knowledge as predictors of student achievement in introduction to psychology. *Journal of Educational Psychology*, 96, 778–784.

U.S. Department of Education (2009). *Race to the Top Program executive summary*. Washington, DC: U.S. Department of Education. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

van den Broek, P. (2012). Individual and developmental differences in reading comprehension: Assessing cognitive processes and outcomes. In J. P. Sabatini, E. R. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how to assess reading ability* (pp. 39–58). Lanham, MD: Rowman and Littlefield Education.

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

van den Broek, P., Risden, K., & Husebye-Hartman, E. (1995). The role of the reader's standards of coherence in the generation of inference during reading. In R. F. Lorch, Jr., & E. J. O'Brien (Eds.), *Sources of coherence in text comprehension* (pp. 353–373). Mahwah, NJ: Erlbaum.

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

Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy, 35.* Retrieved from http://www.umanitoba.ca/publications/cjeap/articles/volante.html

Voss, J., & Silfies, L. (1996). Learning from history text: The interaction of knowledge and comprehension skill with text structure. *Cognition and Instruction, 14,* 45–68.

Walker, C. (1987). Relative importance of domain knowledge and overall aptitude on acquisition of domain-related information. *Cognition and Instruction, 4,* 25–42.

Notes

[1] We imagine this threat likely exists to some degree in current off-the-shelf reading comprehension tests as well, but it goes unnoticed because it is not measured or estimated. While directly addressing the issue poses numerous challenges, we do not feel that leaving it unaddressed is a solution.

[2] We are not including as performance moderators the measurement of basic human abilities (e.g., working memory, processing speed, spatial ability; see Kyllonen, 2013, for review of contemporary theories and models of human ability measurement). Individual differences in basic human abilities exist, but we will assume that individuals within normal parameters of these basic abilities should be able to develop a high level of reading literacy proficiency. That is, we do not see these kinds of individual differences as threats to validity claims or interpretation of scores, unless one were to posit that they constitute a disabling condition rendering a student unable to perform literacy tasks with proficiency. While basic human abilities are discussed in the reading literature, they are not treated as elements of the learning and instruction environment in the same way as are background knowledge, motivation and engagement, and reading, metacognitive, and self regulatory strategies.

[3] At the time of writing, these techniques are currently being implemented, evaluated, and refined in the RfU research program (see Educational Testing Service, 2013; Institute of Education Sciences, 2009, 2010). We recognize the experimental nature of the innovations we propose; some of these approaches may succeed, while others will fail. The most promising approaches will be codified in later drafts of the framework.

[4] One could make a distinction between the terms, such that background knowledge applies to general knowledge about the world, while prior knowledge might refer to idiosyncratic experiences of an individual (e.g., a walk in the park). In this report, we do not.

[5] Not measuring background knowledge in a traditional reading assessment doesn't make the effect of background knowledge go away. Shapiro (2004) has shown that even when attempts have been made to reduce knowledge demands such as by using texts with general or "unfamiliar" topics, measures of background knowledge still account for significant variance in comprehension scores.

[6] While low intrinsic motivation is associated with poor comprehenders (Logan, Medford, & Hughes, 2011), it is possible that low motivation may mask students' ability to demonstrate their reading potential.

[7] Petraglia (1998) argues that no simulation is a perfect representation of reality and users need to be persuaded they are participating in an authentic environment. However since 1998, one could argue there is now a blur between simulations of reality and reality itself. Online social networking sites and virtual worlds such as Second Life are some good examples.

[8] Some large scale tests do assess multiple text comprehension such as NAEP and AP.

[9] http://www.ets.org/research/topics/reading_for_understanding/