3

The Assessment of Reading for Understanding[1]

Panayiota Kendeou, *University of Minnesota*

CONTENTS

[1] The writing of this report was funded in part by the Guy Bond Chair in Reading from the College of Education and Human Development at the University of Minnesota to P. Kendeou and by grant R305U160002 from the U.S. Department of Education to the National Academy of Education. P. Kendeou would like to thank Elly Orcutt, Reese Butterfuss, and Jasmine Kim at the University of Minnesota for their assistance. The opinions are those of the author and do not represent the policies of the U.S. Department of Education.

EXECUTIVE SUMMARY

The U.S. Department of Education's Institute of Education Sciences (IES) created the Reading for Understanding (RfU) research initiative with the ultimate goal being to improve reading comprehension across pre-kindergarten (pre-K) through grade 12 in U.S. schools. The initiative funded a set of six connected projects (teams) that designed, developed, and tested new interventions and assessments in pre-K through grade 12. This chapter focuses primarily on the three main assessments developed by the assessment consortium, which consisted of the Educational Testing Service (ETS) in collaboration with the Florida Center for Reading Research (FCRR) at Florida State University (FSU). This consortium was tasked specifically with the development of a new summative assessment of reading comprehension across all grades. The five teams that designed and tested intervention programs in different age groups (elementary, middle, and high school) also developed assessments of various reading-related constructs. When relevant, this chapter also includes a discussion of a selected set of these measures because they showed evidence for innovation, technical adequacy, and promise for further development.

To address the RfU core assessment mission, the assessment consortium defined the construct of reading comprehension as *reading literacy,* which was measured by two assessment types: components of reading and global reading literacy. Two assessment systems were developed to assess *components of reading* in K–12: the Reading Inventory and Scholastic Evaluation (RISE) and the FCRR Research Reading Assessment (FRA). One assessment system was developed to assess *global reading literacy* in grades 3–12: the Global Integrated Scenario-Based Assessment (GISA). In addition to these three main assessments, a variety of measures were developed by the other teams (Language and Reading Research Consortium [LARRC], Catalyzing Comprehension through Discussion and Debate [CCDD], Promoting Adolescents' Comprehension of Text [PACT], and Reading, Evidence, and Argumentation in Disciplinary Instruction [READI]) to assess reading-related constructs, such as inference making, social perspective taking, knowledge acquisition, evidence-based argumentation, epistemic beliefs, and academic language, as well as classroom survey tools to assess teaching strategies and student strategies.

Our review and evaluation of these assessments and tools led to the conclusion that the RfU research initiative had a profound impact in the area of reading comprehension assessment. The initiative enabled innovative, large-scale work in diverse populations and contexts. Collectively, the set of assessments developed by ETS and FCRR can be characterized as a *new generation* of reading assessments. These assessments reflect a broader and more authentic conceptualization of reading comprehension, are developmentally sensitive, emphasize instructional sensitivity and value, and reflect the consequences of reading with comprehension. All assessments, those developed by the assessment consortium and the other teams, have a strong theoretical basis and defensible psychometric properties. The overall result is a set of *forward-thinking assessments* that promise to advance both research and practice in reading comprehension for years to come.

An important goal in the future research agenda would be to use these assessments in place of more traditional standardized reading comprehension measures. The use of these assessments in various populations and contexts will, in turn, inform further

development and refinement of reading comprehension theories and models, help evaluate with better precision additional aspects of reading comprehension in younger and older readers, and help understand more deeply the implications of integrating important moderators (such as prior knowledge) into the assessment design. Finally, because these new assessments reflect some of the inherent complexities of the comprehension process that only now have been realized in assessment, they open new possibilities for a future research agenda that can significantly advance theories of reading comprehension.

RECENT HISTORY OF LITERACY INITIATIVES

In 1999, the U.S. Department of Education's Office of Educational Research and Improvement (the predecessor office to the Institute of Education Sciences) charged the RAND Reading Study Group (RRSG) with developing a research agenda to address pressing issues in literacy over the next 10 years. This initiative materialized in a 2002 publication (RRSG, 2002), in which the RRSG made recommendations for a future research agenda that focused on three areas: comprehension instruction, teacher education, and assessment. Pertinent to this report were the recommendations with respect to the assessment of reading comprehension. The RRSG proposed a new approach to assessment, advocating for a strong theoretical basis that is at the same time flexible to adapt and change in the presence of new empirical evidence. The group also advocated for using assessment to directly inform and improve instruction. Specifically, the call was for the design of technically adequate measures of reading comprehension that are sensitive to instructional interventions as well as to specific forms of reading instruction for all readers. The research agenda put forth by the RRSG informed the research focus and priorities set by the RfU research initiative 10 years later.

The RAND Reading Study Group: Needs in Reading Comprehension Assessment

The findings of the RRSG report were consistent with persistent criticisms of widely used reading comprehension assessments. These assessments have long been criticized for inadequately representing the complexity of reading comprehension and its development, lacking instructional utility (Klingner, 2004; Pearson & Hamm, 2005; Snyder, Caccamise, & Wise, 2005), and not meeting technical adequacy criteria (Mislevy, 2006, 2008). These assessments depend primarily on immediate recall and basic literal and inferential multiple-choice questions. Most important, none of these assessments are based on a current theory of reading comprehension (RRSG, 2002).

According to the RRSG, new assessments of reading comprehension needed to (a) reflect the dynamic, developmental nature of comprehension; (b) represent adequately the interactions among the dimensions of reader, activity, text, and context; and (c) satisfy criteria set forth by psychometric theory. Furthermore, these new assessments needed to also reflect the consequences of reading with comprehension, such as acquiring and applying knowledge. Most important, developing new assessments was of the highest priority as good assessments are a prerequisite to making progress with all other aspects of the research agenda on reading comprehension.

The minimum criteria for the development of new assessments put forth by the RRSG were the following:

1. Capacity to reflect authentic outcomes;
2. Consistency with actual comprehension processes;
3. Developmental sensitivity;
4. Capacity to identify poor comprehenders;
5. Capacity to identify subtypes of poor comprehenders;
6. Instructional sensitivity;
7. Openness to intraindividual differences;
8. Usefulness for instructional decision making;
9. Adaptability to individual, social, linguistic, and cultural variations; and
10. A basis in measurement theory and psychometrics.

It is important to note that the RRSG acknowledged that no single assessment could meet all of these criteria. Rather, the research agenda called for an *assessment system or systems* that would address different purposes, audiences, and populations.

The Reading for Understanding Research Initiative

In 2010, IES funded the RfU research initiative (IES, 2010) to provide rigorous research to guide the development of better interventions and assessments across pre-K through grade 12. The Institute funded a set of connected projects that would design and test new interventions and assessments to improve *reading for understanding* across all readers in U.S. schools (Douglas & Albro, 2014). The RfU not only renewed professional interest in reading comprehension across the entire pre-K through grade 12 range, but also presented a unique opportunity to develop a community of researchers who undertook innovative work in the area of reading comprehension, with the potential to advance both research and practice.

Core Assessment Mission

To address the need for the development of a new reading comprehension assessment system, the RfU funded one assessment consortium, consisting of the ETS in collaboration with the FCRR at FSU. This consortium was tasked specifically with the development of a new summative assessment of reading comprehension in pre-K through grade 12. In this context, the assessment consortium expanded the definition of the construct of reading comprehension. The construct was identified as that of *reading literacy* and was measured by two assessment types: components of reading and global reading literacy (O'Reilly, Sabatini, Bruce, Pillarisetti, & McCormick, 2012; Sabatini & Bruce, 2009; Sabatini, Bruce, & Steinberg, 2013; Sabatini, O'Reilly, & Deane, 2013). The *components of reading* were assessed with RISE (Sabatini, Bruce, Steinberg, & Weeks, 2015; Sabatini, Weeks, et al., 2019) and with the FRA (Foorman, Petscher, & Schatschneider, 2015a, 2015b). *Global reading literacy* was assessed with the GISA (Sabatini, O'Reilly, Weeks, & Steinberg, 2016; Sabatini, O'Reilly, Weeks, & Wang, 2019).

Additional Assessment Development

It is important to note that the RfU also resulted in a set of additional measures and survey tools that were developed by the other teams in the context of their intervention work; that is, the teams needed to develop additional, often more specific, measures of reading comprehension or related language, knowledge, or cognitive processes in order to fully evaluate the impact of their interventions. For the purposes of this report, a selected set of these assessments were reviewed because they showed evidence for promise and technical adequacy for further development. Specifically, the LARRC developed an Inference Task (LARRC & Muijselaar, 2018) to assess local and global *inference processes*. The CCDD team developed two measures, the Assessment of Social Perspective-taking Performance (ASPP; Kim, LaRusso, Hsin, Selma, & Snow, 2018) to assess *social perspective taking* and the Core Academic Language Skills Instrument (CALS-I; Phillips Galloway & Uccelli, 2019; Uccelli et al., 2015a, 2015b) to assess *academic language*. The PACT team developed a Causal Inference Task to assess *inference making* (BRIDGE-IT; Barth, Barnes, Francis, Vaughn, & York, 2015) and a Background Knowledge measure (ASK; Vaughn et al., 2013) to assess *knowledge acquisition*. The READI team developed the Evidence-Based Argument (EBA) assessment (Goldman et al., 2016, 2019) to evaluate *evidence-based argumentation* and the literature epistemic cognition measure (Yukhymenko-Lescroart et al., 2016) to evaluate domain-specific *epistemic beliefs* in content areas. With respect to survey tools, the PACT team developed the Contextualized Reading Strategy Survey (CReSS; Denton, Wolters, et al., 2015) to evaluate *students' strategy use*, and the READI team developed a teacher survey scale to evaluate *attitude*, *self-efficacy*, and *argument/multiple source practices* as well as a classroom observation scale to evaluate *teaching practices* and *student activities* (Goldman et al., 2019). All assessments and surveys reviewed are listed in Figure 3-1.

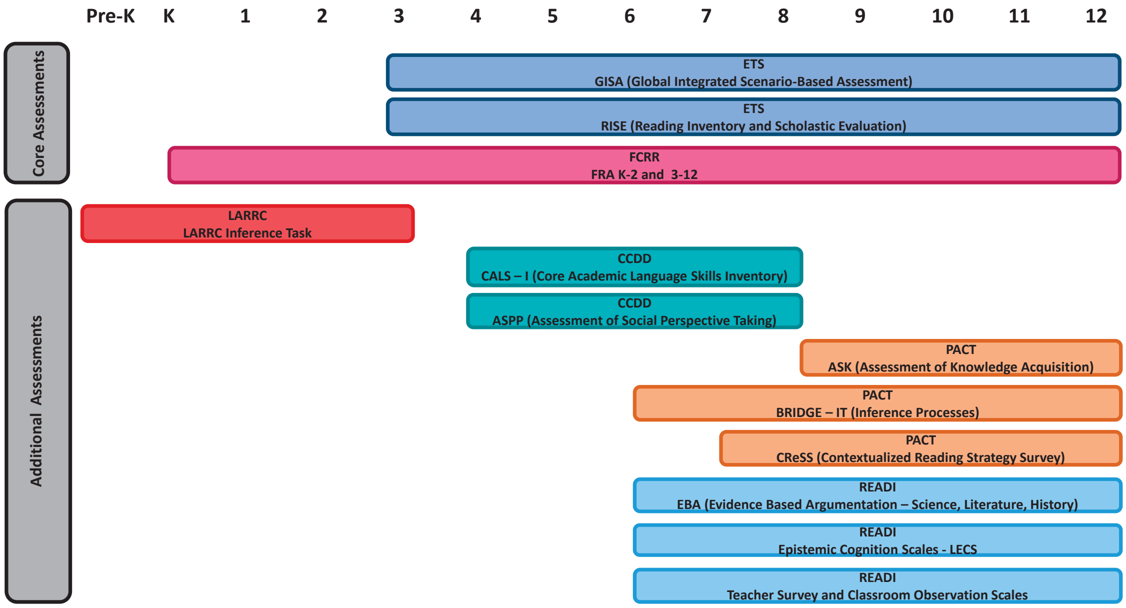

FIGURE 3-1 Assessments and classroom surveys reviewed.

CONTRIBUTIONS OF THE RFU RESEARCH INITIATIVE AND A FUTURE RESEARCH AGENDA

To evaluate the contributions of the RfU research initiative on assessment, we followed an integrative approach that focused on the minimum criteria put forth by the RAND Research Study Group (2002), current trends in reading comprehension research, and an in-depth review of each assessment. The review of each assessment focused on the conceptual framework guiding development, content and sample items, administration and scoring guidelines, and evidence for technical quality focusing specifically on validity, reliability/precision, fairness in testing, and intended use of scores (AERA, APA, & NCME, 2014).

It is important to keep in mind the distinction between the assessments that emerged from the **core assessment mission** versus those developed to allow researchers to measure key facets of their interventions. For the assessments involved in the core mission, it was absolutely imperative to adhere to the highest psychometric standards; this meant that the ETS-FCRR team needed to engage in extensive and iterative large-scale, validity studies of the assessments. The other five teams were not funded to engage in extensive psychometric analyses, but they did engage in standard procedures for establishing the reliability, validity, and utility of their measures for the populations of students with whom they carried out their interventions. Nonetheless, we applied the same standards to all of the assessments introduced in this chapter and elaborated in the more detailed accounts in Appendix 3-1. Our hope in doing so was that readers of this report might understand the comprehensiveness of assessment tools that the RfU has made available to the worldwide community of researchers and educators.

Through this integrative, evaluative process, nine themes emerged that helped summarize the contributions of the RfU assessment research. What follows is the discussion of each of those theme contributions. The discussion of each theme concludes, where appropriate, with suggestions for more research that may be needed.

Authenticity: Complicating the Reading Comprehension Construct

Reading comprehension is among the most complex of human activities. It involves processing words, connecting words using rules of syntax to understand sentences (Perfetti & Stafura, 2014), integrating meaning across sentences, drawing on relevant knowledge, generating inferences, identifying the structure of the text, and taking into consideration the authors' goals and motives (Graesser, 2015). The end product is a mental representation, what has been termed the "situation model" (Kintsch & van Dijk, 1978), which reflects the overall meaning of the text. For all of these processes to be successful, many interacting factors are playing a role, such as reader characteristics, text properties, context, and the demands of the reading task (Kintsch, 1998; RRSG, 2002).

The assessment consortium embraced the complexity of reading comprehension and expanded the construct definition. The construct was identified as that of *reading literacy*, defined as

> the deployment of a constellation of cognitive, language, and social reasoning skills, knowledge, strategies, and dispositions, directed towards achieving specific reading purposes. (Sabatini, O'Reilly, & Deane, 2013, p. 7)

The decision to define and assess a broad construct such as reading literacy was innovative and contemporary. The decision was driven by recent policy efforts, including the Common Core State Standards for K–12 education in the United States (NGA & CCSSO, 2010), new social studies (NCSS, 2013) and science standards (NRC, 2012), the Partnership for 21st Century Skills (2008), frameworks for international assessments of reading such as the Programme for International Student Assessment (PISA; OECD, 2009a), the Programme for the International Assessment of Adult Competencies (PIAAC; OECD, 2009b), and the Progress in International Reading Literacy Study (PIRLS; Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009), and other assessment efforts and reforms (Bennett, 2011; Bennett & Gitomer, 2009; Gordon Commission, 2013).

Adopting a broad construct of reading comprehension embraces its complexity and allows for a focus on the entire *range* of reading processes, from foundational to higher-order processes (Fletcher, 2009; Goldman & Snow, 2015; Snow, 2018). Indeed, the focus in RISE and FRA is mostly on foundational reading skills, whereas the focus in GISA is on higher-level and goal-directed reading comprehension. Targeting this broad range of processes and embracing the complexity of reading also necessitates the integration of important variables that are expected to influence performance. These variables—(a) prior knowledge, (b) metacognitive and self-regulatory strategies, (c) reading strategies, and (d) student motivation and engagement—can affect the interpretation of reading comprehension scores (O'Reilly & Sabatini, 2013). For this reason, these variables were either directly assessed in the context of the assessment (this was the case for prior knowledge) or integrated in the assessment design (this was the case for all four). This approach is a considerable strength of GISA.

Expanding the reading comprehension construct enabled focus not only on higher-level processes during assessment, but also on deeper comprehension (Graesser, 2015; O'Reilly, Sabatini, & Wang, 2018), and thus deeper learning (Goldman & Pellegrino, 2015). As a result, and consistent with the recommendations made by the RRSG (2002), the focus shifted from comprehension to the consequences of reading with comprehension, such as acquiring and applying knowledge. This was accomplished by using a scenario-based assessment design (Bennett & Gitomer, 2009; O'Reilly & Sheehan, 2009), which approaches reading comprehension assessment as learning: it focuses on the consequences of comprehension rather than comprehension itself. The shift in focus from comprehension to learning is the main difference between GISA and most traditional reading comprehension assessments. This shift already has been embraced by several international literacy assessments (e.g., PISA, PIAAC, and PIRLS) and, once widely adopted, will present both a challenge and an opportunity for theory and practice in reading comprehension. *In other words, theory and practice also need to shift in focus from comprehension to learning, an issue that needs to be addressed in the future research agenda*.

Theoretically Based: Component and Process Theories of Reading Comprehension

It has been argued repeatedly that reading comprehension models and theories have not directly informed past assessment efforts, and that new assessments should be based on an elaborated theory of reading comprehension (RRSG, 2002). The ETS-FCRR consortium drew on multiple theoretical frameworks and models to inform their

> The primary goal of the ETS assessment project was to build a theoretically-driven, developmentally sensitive assessment system that spanned pre-K to grade 12. Our subgoal was to design assessments that address an expanding 21st-century reading construct, incorporate reading and learning science in the designs, and enhance instructional relevance, while still maintaining feasibility of implementation and psychometric quality.
>
> *—John Sabatini, Steering Committee Representative from ETS*

assessment efforts. The use of multiple theories (as opposed to a single theory or model) is consistent with the inherent complexity of reading comprehension that makes it challenging for a single theory to describe the full range of cognitive, social, and linguistic processes involved (Perfetti & Stafura, 2014) or to make precise, testable predictions (Kendeou & O'Brien, 2018). Specifically, the consortium drew on both *component* and *process* models of reading comprehension, integrating different theoretical perspectives and views.

Component models focus on the identification of *component skills* that explain reading comprehension performance. Reading component skills are subskills that can be isolated and assessed independently from higher-level reading comprehension (Perfetti & Adlof, 2012). Relevant to the RfU, component models of reading comprehension have been particularly influential for the development of the core assessments FRA and RISE as well as additional assessments, such as CALS-I, ASPP, and ASK (see Figure 3-1). These assessments include several of the component skills known to predict reading comprehension, such as word decoding and its precursors (Ehri, 2014), reading fluency (Fuchs, Fuchs, Hosp, & Jenkins, 2001), syntactic awareness (Cain & Nash, 2011; Crosson & Lesaux, 2013), vocabulary knowledge (Quinn, Wagner, Petscher, & Lopez, 2015), academic language (Snow, Lawrence, & White, 2009; Uccelli et al., 2015a, 2015b), language comprehension (Connor et al., 2014, 2018; Kendeou, van den Broek, White, & Lynch, 2009; Kim, 2016; Storch & Whitehurst, 2002), and perspective taking (LaRusso et al., 2016). Several of these components have been termed "pressure points" (Compton & Pearson, 2016), defined as skills that can result in robust variations in reading comprehension performance (Perfetti & Adlof, 2012). Among the component models in the extant literature, the Simple View of Reading (SVR; Hoover & Gough, 1990), which describes reading comprehension as the product of decoding and language comprehension, has been very influential for the development of both RISE and FRA. In the context of the SVR, decoding includes processes needed to decipher written code, such as phonological processing, orthographic processing, and word recognition, whereas language comprehension includes processes needed to build a coherent mental representation, such as vocabulary, academic language, and inference generation.

Process models focus on the identification of various *processes* involved in the construction of a mental text representation during reading (see McNamara & Magliano, 2009, for a review). An important assumption in most process models is that reading is a purposeful or goal-driven activity (Britt, Rouet, & Durik, 2018; McCrudden, Magliano, & Schraw, 2011). These purposes or goals influence readers' desired level of comprehension or standards of coherence (van den Broek, Bohn-Gettler, Kendeou, Carlson, & White, 2011) and thus comprehension and learning from text. Relevant to the RfU, several process models of reading comprehension have been particularly influential for the development of the core assessment GISA as well as additional assessments, such as BRIDGE-IT, the LARRC inference task, and EBA (see Figure 3-1). Among these models, the Construction-Integration Model (Kintsch & van Dijk, 1978) describes reading comprehension as the activation and integration of text information and relevant background knowledge into a coherent mental representation (i.e., a situation model) (Kintsch, 1988; van den Broek et al., 2005). The Landscape Model (van den Broek, Young, Tzeng, & Linderholm, 1999) specifies how the construction and integration processes are influenced by readers' standards of coherence or criteria for comprehension. The Documents Model Framework (Perfetti, Rouet, & Britt, 1999) and the Multiple-Document Task-based Relevance Assessment and Content Extraction model (MD-TRACE; Rouet & Britt, 2011) describe reading comprehension of multiple documents and texts and identify additional processes that are relevant in this context, including the evaluation and integration of information across sources (Goldman, Greenleaf, et al., 2016).

To the extent that theories of reading comprehension inform the development of reading comprehension assessments, evidence from the use of these assessments can also inform further development of reading comprehension theories. Indeed, the development of theoretically-based assessments has already begun to facilitate this reciprocal relation between theory and assessment. For example, ongoing work by the assessment consortium has produced new insights with respect to the relation of core component skills, such as decoding and reading comprehension. Wang, Sabatini, O'Reilly, and Weeks (2019) provided evidence for the nonlinear relation between decoding and reading comprehension by identifying a *decoding threshold* in grades 5–10 using RISE. Decoding below this threshold was only weakly related to reading comprehension and reading comprehension performance was limited. Decoding above this threshold positively predicted performance in reading. Wang et al. (2019) argued that the Decoding Threshold Hypothesis has the potential to explain differences in prominent reading theories in terms of the role of decoding in reading comprehension across development. *Thus, using evidence from the use of these assessments to further develop current theories of reading comprehension is an important goal in the future research agenda.*

Developmental Sensitivity: A Dynamic Construct

Reading comprehension is a dynamic construct that changes across development (Weeks, 2018). That is because reader characteristics change with age and experience. As a result, the relative contribution of these characteristics to reading comprehension varies across development (van den Broek & Kendeou, 2017). For example, in the early elementary school grades decoding skills (the "reading" in reading comprehension) are a major contributor to reading comprehension, but in later elementary school grades and onward comprehension skills (the "comprehension" in reading comprehension), such as inference generation and oral language, are stronger predictors (Catts, Hogan, & Fey, 2003; Ehri, Nunes, Stahl, & Willows, 2001). This shift coincides with a transition from *learning to read* to *reading to learn* as complex informational texts become more common in the curriculum (Chall, Jacobs, & Baldwin, 1990; Goldman, Snow, & Vaughn, 2016; Snow & Sweet, 2003), and fits with conceptualizations of reading development as a dynamic system (van Geert, 1991). This dynamic nature of the construct itself presents a challenge when the goal is to develop an assessment system that can span stages of development (e.g., K–12). The ETS-FCRR consortium addressed this challenge by taking into account the main determinants of the construct across development.

Specifically, in the FRA assessment there is a clear differentiation between kindergarten through grade 2 and grades 3–12, such that basic decoding skills are initially assessed with tasks that provide direct, specific measurement of letter-sound knowledge, phonological awareness, and spelling, whereas they are later assessed with tasks that provide measurement of the application of decoding skills, such as word recognition and vocabulary knowledge (Fitzgerald et al., 2014; Foorman, Francis, Davidson, Harm, & Griffin, 2004). Similarly, comprehension assessment also shifts from *listening* to *reading*, requiring students to either listen to or read passages (depending on their decoding proficiency). Furthermore, the system is designed to be administered in fall, winter, and spring to effectively track development of these "dynamic" skills in shorter periods.

RISE evaluates components of reading comprehension in grades 3–12, one dimension of the broader construct of reading literacy (O'Reilly et al., 2012; Sabatini, Bruce, & Steinberg, 2013; Sabatini, Weeks, et al., 2019). The range of these component skills—beginning with the recognition or decoding of words, to understanding the meanings of words and sentences, to building meaning from a text—is consistent with the developmental progression of reading comprehension (RRSG, 2002; Snow, 2018). Sabatini et al. (2015) provided initial evidence for unidimensionality of each component/subscale and examined all of the grade-level means and standard deviations, noting that, in general, the means incrementally increased across grades 5–10. Recently, Sabatini, Weeks, et al. (2019) provided further evidence for the unidimensionality of these components in grades 3–12.

GISA evaluates the expanded construct of reading comprehension following the same design across grades 3–12. The available GISA forms are scenario based, framed around different literacy goals, include either science or history topics, and have various numbers (range from 27 to 59) and types (constructed response, graphic organizer, and multiple choice) of questions. Despite these variations, Sabatini and colleagues (Sabatini, O'Reilly, et al., 2019) provided adequate evidence for the scale's unidimensionality. Sabatini, O'Reilly, et al. (2016) also examined all of the grade-level means and standard deviations and noted that they reflected developmental differences in ability across grades. In general, the means increased across all grades. The one exception was in grade 12, where the mean was slightly lower than that in grade 11.

Sabatini, Halderman, O'Reilly, and Weeks (2016) also developed and tested a GISA form for K–3 in a small-scale study to evaluate the feasibility of the scenario-based assessment for younger students. The results showed ability differences across grade levels, even though third graders read silently, second graders read aloud, and kindergarten and first graders listened to the texts. Technical adequacy indices, though, were susceptible to the changes in delivery/modality, making it challenging to further develop GISA forms for K–2; for this reason, GISA begins at grade 3.

Thus, FRA, RISE, and GISA collectively assess a dynamic construct of reading comprehension across grades and demonstrate adequate developmental sensitivity. It is important to note that, with respect to component skills, the assessments extend from kindergarten through grade 12 (RISE is in grades 3–12 and FRA is in kindergarten through grade 12) but with a validation sample only up to grade 10 for FRA. With respect to global reading literacy (GISA), the assessment extends from grades 3–12 but with less developmental sensitivity in grades 11 and 12 (and has no forms in K–2). This pattern of results suggests that at the ends of the developmental spectrum (pre-K, K–2, and grades 11 and 12), the broader construct of reading comprehension is not yet adequately captured by these assessment systems. *Thus, further developing measures of global reading literacy for younger readers while also refining those for older readers is an important goal in the future research agenda.*

The changing nature of the reading comprehension construct for older readers is also reflected in the approaches adopted by the three RfU teams that focused specifically on the development of interventions for adolescent students (Goldman, Snow, & Vaughn, 2016). Specifically, the READI team approached reading comprehension in grades 6–12 as a discipline-specific task (Shanahan & Shanahan, 2008) that requires readers to analyze, synthesize, and evaluate information within and across sources (Goldman, Greenleaf, et al., 2016, 2019; Lee & Goldman, 2015). The focus on sources also expanded the traditional notion of "text" to include print-based texts, images, audio, and video texts. This approach provided new insights on the higher-order and discipline-specific aspects of reading comprehension that are important for readers in the 21st century. Likewise, the PACT team approached reading comprehension in grades 7–12 through the lenses of content learning (Gersten, Baker, Smith-Johnson, Dimino, & Peterson, 2006; Vaughn et al., 2009), focusing primarily on prior knowledge activation, vocabulary building, text-based learning, and team-based learning (Vaughn et al., 2017). Finally, the CCDD team approached reading comprehension in grades 4–7 through discussion and debate, focusing on identifying different perspectives expressed in texts, learning academic vocabulary, and practicing academic language structures orally and in writing (Jones et al., 2019). The team's work was motivated by an analysis of the tasks adolescents are meant to be accomplishing through reading, and how they differ from the tasks typically embedded in traditional comprehension assessments, which often require only relatively shallow inferences. Importantly, these three teams (READI, PACT, and CCDD) also developed measures to evaluate core constructs related to their intervention research, such as EBA (Goldman et al., 2019), knowledge acquisition (ASK knowledge measure; Vaughn et al., 2013), academic language (CALS-I; Phillips Galloway & Uccelli, 2019; Uccelli et al., 2015a, 2015b), and social perspective taking (ASPP; Kim et al., 2018). These measures reflect a few of the additional aspects of the broader reading comprehension construct that are developmentally appropriate for middle and high school readers. *An important goal in the future research agenda would be to continue to identify aspects of the broader reading comprehension construct that are developmentally appropriate for different populations and disciplines.*

Instructional Sensitivity: Reflect the Effects of Intervention

Instructional sensitivity, namely, an assessment's capacity to reflect the effect of instruction or intervention, has been set as a core assessment criterion by the RRSG (2002) that has been realized in several RfU assessments. This is an important goal for any assessment since, historically, traditional measures of reading comprehension rarely show such sensitivity (Denton, Wexler, Vaughn, & Bryan, 2008). This is due, in part, to the fact that different reading comprehension assessments often measure different aspects of the construct (Keenan, Betjemann, & Olson, 2008), and evidence for the impact of an intervention depends on whether the aspects of the construct being assessed are the same as those being trained (O'Reilly, Weeks, Sabatini, Halderman, & Steinberg, 2014). This lack of sensitivity is also a consequence of transfer failure. Transfer (Barnett & Ceci, 2002; Day & Goldstone, 2012) is very difficult to evaluate and achieve in education settings in general, and in reading in particular (Gick & Holyoak, 1980, 1983). As a result, researchers have often distinguished between "proximal" and "distal" measures of reading comprehension in intervention studies (Connor et al., 2014). *Proximal* measures are closely tied to the intervention/instruction (require *near* or no transfer), whereas *distal* measures are generalized outcomes we expect to be influenced by the intervention/instruction (require *far* transfer). What is particularly promising for this new generation of assessments is that the three core assessments—GISA, RISE, and FRA—have shown some sensitivity to instruction, even though they would be considered distal measures.

Specifically, O'Reilly et al. (2014) demonstrated GISA's use as a summative assessment designed to provide evidence for the efficacy of the Reading Apprenticeship intervention (Monte-Sano, 2010). It is important to note that the underlying approach to assessment design in GISA (see O'Reilly & Sabatini, 2013; Sabatini & O'Reilly, 2013; Sabatini, O'Reilly, & Deane, 2013) had several elements in common with the Reading Apprenticeship program, making it more a proximal rather than a distal measure. The program was designed to train disciplinary reading in high school history, science, and literature and three GISA forms were specifically designed to evaluate the outcomes of the intervention in grades 9–12. O'Reilly et al. (2014) concluded that the GISA assessments were promising for use as outcomes in the intervention and sensitive to intervention effects.

Similarly, Goldman et al. (2019) also used GISA as a distal measure in a randomized controlled trial evaluating the efficacy of the READI Science intervention when compared to a business-as-usual control in grade 9. The READI Science intervention aims to improve reading comprehension by training evidence-based argumentation across multiple sources in science. The results showed that GISA was sensitive to the READI intervention effects with the treatment condition scoring significantly higher on GISA than the control condition. Notably, this effect held even after controlling for RISE at pretest.

Kim et al. (2017) included both RISE and GISA with the goal to evaluate the efficacy of the Strategic Adolescent Reading Intervention (STARI). STARI aims to improve reading comprehension by using reciprocal teaching strategies (Palincsar & Brown, 1984) and student discussion and debate within thematic text units in grades 6–8, while also building fluency through carefully selected leveled texts. Kim et al. (2017) reported that the program demonstrated significantly positive effects on the RISE word recognition, morphological awareness, and efficiency subtests.

Foorman, Herrera, Dombek, Schatschneider, and Petscher (2017) also demonstrated the utility of FRA K–2 as a summative measure to provide evidence for the efficacy of interventions. The study consisted of a randomized controlled trial that compared two early literacy interventions—one using standalone materials and one using materials embedded in the existing core reading program. The findings showed that the FRA

K–2 system was sensitive to intervention effects by demonstrating that the standalone intervention significantly improved spelling outcomes relative to the embedded intervention; other student outcomes were similar for the two interventions.

As noted earlier, the RfU also resulted in a set of additional measures developed specifically to evaluate the efficacy of various interventions and other reading-related constructs. These measures cover a range of different constructs and age groups. Specifically, the CCDD team developed the CALS-I (Uccelli et al., 2015a, 2015b) to evaluate the efficacy of the STARI and Word Generation interventions on improving students' *academic language* in grades 4–8 (LaRusso et al., 2016). The team also developed the ASPP measure (Kim et al., 2018) to assess students' social perspective taking since their intervention program hypothesized *social perspective-taking* performance as a core mechanism of learning. Both CALS-I and ASPP have shown evidence for instructional sensitivity (Kim et al., 2017; Phillips Galloway & Uccelli, 2019; Uccelli et al., 2015a, 2015b) and predictive relationships to comprehension (LaRusso et al., 2016).

The LARRC team developed an inference measure to assess global and local inference-making skills during listening comprehension in children pre-K through grade 3 (LARRC & Muijselaar, 2018) in the context of the language-based comprehension instruction Let's Know! (LARRC, Jiang, & Logan, 2019). LARRC (2015) provided evidence for the validity of a discourse skills factor that included this inference task along with four other measures of discourse skills. LARRC (2017) also provided evidence for the validity of a listening comprehension factor that included this inference task along with two other listening comprehension tests. Even though the team suggests that the measure could be used to evaluate the effects of language comprehension intervention or instruction, such evidence has not been provided yet.

The PACT team developed the ASK measure to evaluate middle and high school students' learning of U.S. history. ASK includes two subtests, one that measures content knowledge relevant to the intervention and one that measures reading comprehension. ASK has been used successfully to evaluate the efficacy of the PACT intervention in improving students' social studies content knowledge in grade 8 (Vaughn et al., 2013, 2015). The measure was also used to evaluated PACT's efficacy for English learners in grade 8 (Vaughn et al., 2017; Wanzek, Swanson, Vaughn, Roberts, & Fall, 2016), students with disabilities in grade 8 (Swanson et al., 2016; Wanzek, Swanson, Vaughn, Roberts, & Fall, 2016), and students in grade 11 (Wanzek et al., 2015).

The READI team developed the EBA measure to evaluate adolescents' ability to make evidence-based arguments from multiple sources in science. The EBA measure was used to evaluate the efficacy of the READI intervention that was designed to engage students in evidence-based argumentation from multiple text-based sources in grade 9 life sciences (Goldman, Greenleaf, et al., 2016). Goldman et al. (2019) showed that the multiple-choice component of the EBA measure was sensitive to instruction, with the intervention group performing significantly higher compared to the control group. EBAs were also developed for history and literature, along with rubrics for evaluating them. These were used in the context of the iterative design-based research conducted with a small number of teachers in each discipline. The EBAs in history and literature remain to be validated with larger samples of students.

The READI team also developed epistemic cognition scales in history, science, and literature. The science and history scales emphasized two dimensions of epistemic cognition for multiple sources in history and science: the importance of corroborating across documents (history) and data sets and experiments (science), and the complexity and uncertainty of historical and scientific knowledge. The Literature Epistemic Cognition Scale (LECS; Yukhymenko-Lescroart et al., 2016) emphasized three dimensions: the multiple meanings of any literary work, the relevance of literature to life, and the importance of multiple readings of a literary work. The READI literature intervention (Goldman, Greenleaf, et al., 2016) used the LECS in the context of a 2-year longitudinal study of adolescents during their sophomore and junior years. Two of the subscales (multiple meanings and relevance to life) were significantly correlated with students' perceptions of the instructional context for literature (e.g., encouraging them to consider readings from multiple perspectives and to think about why writers and characters they create do what they do) as well as their self-reports of how frequently they analyzed their readings from different perspectives and considered how others interpreted readings.

With respect to the classroom survey measures, there is also increasing evidence for their instructional sensitivity. The PACT team developed the CReSS to evaluate four constructs related to students' reading strategy use in grades 7–12 (Denton, Wolters, et al., 2015). These constructs included *evaluation and integration strategies* (integrating current text information with previous text information and prior knowledge), *note-taking strategies* (identification of important text information), *regulation strategies* (adjustment of reading in response to difficulty), and *help-seeking strategies* (asking for help in response to difficulty). The survey is designed so that students respond to items targeting the use of comprehension strategies in four imagined reading situations. The first scenario involves reading a social studies textbook to prepare for a small group discussion and class presentation. The second scenario involves reading a story from an English language arts book to prepare for a quiz. The third scenario involves reading a self-selected nonfiction book in social studies in preparation for a written short report. The fourth scenario involves reading two articles from the internet to prepare for a class report. Denton, Wolters, et al. (2015) reported higher use of evaluation/integration and regulation strategies by adequate than struggling comprehenders, while the use of help seeking and note taking did not differ between these groups. Students at higher grade levels also reported greater use of evaluation/integration and regulation strategies than those in lower grades.

Finally, the READI team developed a self-report survey to assess teachers' *attitudes*, *self-efficacy*, and *argument/multiple source practices* (Goldman et al., 2019) in an effort to evaluate the impact of teacher professional development activities. Although developed and piloted in all three content areas (history, literature, and science), it was only validated in life sciences. Goldman et al. (2019) provided evidence that READI science intervention teachers scored significantly higher than those in the control condition on argument/multiple source practices at the conclusion of the intervention although there were no differences between the groups prior to the intervention. Even though there were no significant differences in attitude and self-efficacy from pre- to post-intervention, intervention teachers consistently scored higher than control teachers on the post-intervention administration. The READI team also developed a classroom observation protocol for the life sciences efficacy study and used it to evaluate teacher and student activities. Goldman et al. (2019) reported that, of the six constructs on that protocol, READI intervention teachers improved from the first to the second observation but control teachers did not. Furthermore, at the end of the intervention, READI intervention teachers scored higher than control teachers on all six constructs, with significant differences and large effect sizes on two of the constructs (support for reading and collaboration).

Collectively, the measures developed in the context of the RfU show increased instructional sensitivity. This is true for the measures that by design were well aligned with the outcomes of the respective interventions, but also for the core measures (GISA, RISE, and FRA), as well as the classroom survey measures. Thus, the RfU has contributed to the literature a set of instructionally sensitive measures of knowledge, skills, and processes that contribute to using information obtained through reading single and multiple texts to address important questions. These include more generic skills such as use of academic language and perspective taking, as well as discipline-specific knowledge and skills. This "toolbox" of measures enriches the range of possibilities available to researchers in the field of reading and enables measurement of aspects of reading comprehension that are contemporary and innovative—aspects that were not possible to adequately measure before (e.g., global literacy, evidence-based argumentation, academic language, and social perspective taking). *An important goal in a future research agenda would be further development, calibration, and scale-up of these measures to evaluate their practical utility. In doing so, access to these measures by the scientific community would be necessary.*

Instructional Value:
Identify Student Strengths and Weaknesses to Inform Instruction

Teachers need information with respect to students' strengths and weaknesses in reading, as well as specific instructional recommendations that can address these weaknesses (Denton, Enos, et al., 2015; Pellegrino, DiBello, & Goldman, 2016). Indeed, effective teachers systematically collect and share student assessment data to make instructional decisions that improve student performance (Lipson, Mosenthal, Mekkelsen, & Russ, 2004) by as much as 0.20 standard deviations (Kingston & Nash, 2011). Effective evaluation of students' reading skills and instruction planning, however, requires high-quality formative assessments that assess both comprehension processes and their products (Kendeou, McMaster, & Christ, 2016; van den Broek & Kendeou, 2014). Until recently, there were only a few high-quality formative assessments of reading comprehension (Afflerbach, Cho, & Kim, 2015), a need that has also been highlighted in the research agenda by the RRSG (2002).

The three core assessments produced in the context of the RfU partly address this need, particularly with respect to *component skills* of reading comprehension. Sabatini et al. (2015) suggested that RISE could be used to identify students' strengths and weaknesses in conjunction with GISA. For example, RISE can be used to detect whether foundational reading skills are possible barriers to achieving higher levels of reading comprehension performance as reflected in GISA performance. Sabatini et al. (2014a) provided "proof of concept" of this approach in a small-scale study where they used RISE to create four subgroups of students (proficient, high basic, low basic, and below basic) and subsequently explored the extent to which each RISE subtest predicted unique variance in GISA across these four ability groups. In this sample, each RISE subtest added significant unique variance predicting GISA scores, and together accounted for approximately 69 percent of the variance. Part of the residual variance unaccounted for presumably comprises the complex, deep comprehension required in GISA that cannot be captured by the individual subtests themselves. The results from this proof-of-concept study suggest that scores for each RISE subtest provide evidence for readers' strengths and weaknesses, and the combination of RISE and GISA assessments can provide useful insights on understanding students' reading ability. *It remains an open question whether training of these specific component skills improves reading comprehension. That would be an important next step in examining RISE's diagnostic accuracy.*

Foorman et al. (2015a, 2015b) provided strong evidence that the FRA assessment is an effective screening and diagnostic system of foundational reading comprehension skills. *Screening* in kindergarten through grade 2 is accomplished by evaluating foundational skills, such as phonological awareness, letter sounds, word reading, spelling, vocabulary, and following directions. In grades 3–12, screening is accomplished by evaluating word recognition, vocabulary knowledge, and reading comprehension. In each system, these tasks produce the Probability of Literacy Success (PLS) score following a weighted formula. The PLS score indicates the likelihood that a student will reach end-of-year expectations in literacy. For the purposes of the FRA, reaching expectations is defined as performing at or above the 50th percentile on the Stanford Achievement Test, Tenth Edition. The PLS is also color coded, providing the teacher with actionable information: red indicates the student is at high risk and needs targeted intervention (PLS < .50), yellow indicates the student may be at risk and needs supplemental instruction (PLS > .50 and ≤ .70), and green indicates the student is likely not at risk (PLS > .70). Foorman et al. (2015a, 2015b) provided strong evidence for the predictive power of the PLS cutoff score in kindergarten through grade 10. The FRA team indicated that even though in the initial studies they also included grades 11 and 12, the sample was skewed toward lower-performing students in Florida, so they described it as having a K–10 proficiency range.

It is important to note that, despite the progress made with RISE and FRA, there are currently no formative assessments that evaluate the *actual processes* during reading comprehension. As outlined earlier in this report, current models of reading comprehension assume that comprehension involves the construction of a coherent mental representation of a text or "situation model" (Kintsch & van Dijk, 1978). These models differentiate between the actual processes that give rise to a mental product. An important next step in the development of assessments with instructional value is measures that can provide insights into the cognitive processes "in the moment." For example, the BRIDGE-IT measure (Barth et al., 2015) developed by the PACT team, a computerized inference measure for students in grades 6–12, is a good example of how one core comprehension process—inference making—can be evaluated in the moment. The test evaluates inference making by asking students to judge whether a continuation sentence is consistent or inconsistent with prior text; both accuracy and response times are considered as evidence for inference making. It is during these moment-by-moment processes that comprehension succeeds or fails (e.g., Kintsch, 1998). *Thus, the development, calibration, and scale-up of process assessment measures should be an important goal in the future research agenda. Technological advancements (e.g., eye-tracking methodologies) and trace or log data recorded in digital environments can be particularly helpful in this context.*

Increased Complexity: Texts and Tasks

Most published, standardized reading comprehension assessments in the United States include a set of independent texts each with related literal and/or inferential multiple-choice questions with a task or goal to simply perform in the context of the assessment (Rupp, Ferne, & Choi, 2006). The majority of these assessments are also paper-based and do not include aspects of online reading or digital literacy (Kiili et al., 2018; Sabatini et al., 2015). The assessment consortium took a contemporary approach that expanded the types of tasks, texts, and questions associated with these texts.

With respect to text types, GISA shifted from the traditional set of independent texts to a *set of interrelated texts* that includes different sources and interactive communications (Sabatini et al., 2015). This was enabled with the adoption of a scenario-based design (Bennett, 2011; Bennett & Gitomer, 2009; O'Reilly & Sheehan, 2009). A scenario-based design provides test takers with a specific purpose for reading, a set of materials, and relevant questions. With respect to types of questions, GISA depends heavily on *multiple-choice questions*, but also incorporates two other types*, constructed-response items* summarizing text content and *graphic organizer items* organizing text content. These additional item types require strategies such as integration, synthesis, and application. With respect to task or context, GISA includes aspects of technology and digital environments by design (e.g., simulated peers, multiple sources), making the students' experience akin to learning (rather than testing). The tasks call for students to analyze, evaluate, synthesize, and report information and ideas.

The inclusion of various texts, questions, and tasks begins to address the RRSG (2002) call for assessments to evaluate the performance of an individual across activities with varying tasks and text types. In GISA, this was accomplished by using a scenario-based design. This was also accomplished in additional measures, such as EBA. EBA aligned tasks and texts with their disciplinary reading context (Lee & Goldman, 2015). Even though these are important steps forward, more work is needed to better understand how increased difficulty or complexity can be accomplished by taking into account various combinations of tasks and texts, and how to best utilize the affordances of digital environments in doing so. *Thus, exploring the extent to which scenario-based assessments can be used to introduce increased complexity in the assessment of reading comprehension is an important question in the future research agenda.*

Prior Knowledge: An Integral Component

The inherent influence of prior knowledge on reading comprehension has always been a challenge for reading comprehension assessments. Traditionally, reading comprehension assessments aimed to eliminate rather than integrate prior knowledge, by including content that reduced knowledge demands (Francis et al., 2009; RRSG, 2002). This approach is less than optimal because prior knowledge is not only an integral component of reading comprehension, it is also one of the factors that carries the largest variability (Ahmed et al., 2016; Kendeou et al., 2016; McNamara & Kintsch, 1996) in middle and high school students (Goldman, Snow, & Vaughn, 2016). Prior knowledge is an integral component because at various points during reading, the reader draws on *different sources of knowledge*, including linguistic knowledge and general world knowledge (Perfetti & Stafura, 2014). The accuracy of that knowledge is also important:

accurate knowledge can facilitate reading comprehension, whereas inaccurate knowledge can severely disrupt it (Kendeou & O'Brien, 2015).

Rather than controlling prior knowledge, GISA included it as one of the important moderators in the assessment design (O'Reilly & Sabatini, 2013). This was accomplished by (a) measuring prior knowledge directly, (b) providing access to additional content during the assessment (e.g., videos, audio, definitions, diagrams) that supported students' prior knowledge, and (c) structuring the sequence of sources to facilitate knowledge acquisition. To measure prior knowledge, students were presented with a list of words/terms from a natural language processing database that provided a topical-association index for each word to a topic (Deane, 2012), and were asked to decide whether a term was related to the topic of the text. Students responded "Yes," "No," or "I don't know." O'Reilly et al. (2014) showed that this task was a quick and valid indicator of topic knowledge. By integrating a measure of prior knowledge into the assessment, one can investigate directly how student proficiency might interact with prior knowledge and whether students learn new content after taking the assessment. Indeed, GISA included prior knowledge measurement in a selected number of forms that functioned as a proof of concept for this approach (McCarthy et al., 2018; O'Reilly, Sabatini, & Wang, 2019).

In an elegant analysis, O'Reilly, Wang, and Sabatini (2019) used scores in this prior knowledge assessment to identify a *knowledge threshold*. Below the threshold, the relation between knowledge and performance on GISA was weak (β = 0.18), whereas above the threshold, the relation between knowledge and performance was strong (β = 0.81). These results show that integrating prior knowledge assessment into reading comprehension not only is feasible, but may also help identify what is the minimum knowledge required to comprehend information on a topic. *An important goal in the future research agenda is to evaluate at a larger scale the utility of integrating this type of prior knowledge test into assessment, and to understand better the implications for score interpretation.*

Technical Adequacy

Following measurement theory and sound testing practices are key criteria for the construction of new assessments. In this report, the evaluation of the technical quality of the RfU assessments focused specifically on validity, reliability/precision, fairness in testing, and intended use of scores as outlined by the *Standards for Educational and Psychological Testing* (AERA et al., 2014). To meet these standards, the assessment consortium used sophisticated methodologies associated with test development and statistical analyses (e.g., measurement theory, classical test theory, and item response theory).

The calibration and validation studies for the three core assessment systems (GISA, RISE, and FRA) were *extensive.* More than 100,000 students in grades 3–12 participated from the Midwest, Northeast, Southern, and Western United States for RISE and GISA (Sabatini, 2017; Sabatini, Weeks, et al., 2019), and more than 70,000 students participated from kindergarten through grade 10 for FRA from the southern United States (Foorman et al., 2015a, 2015b). These studies not only included large national samples but were also *iterative* in item design and sample selection, resulting in significant improvements over the course of 5 years of development. Detailed technical reports have been produced for each assessment system that allow researchers and teachers to evaluate whether each assessment fits their assessment needs from a technical adequacy perspective and help understand the scales, scores, and samples used to create them.

Across these assessments, the validity argument (Kane, 2013a, 2013b) integrated three types of evidence: (1) *evidence based on internal structure*, namely, the extent to which the relations among the test components conform to the hypothesized construct (in all instances, unidimensionality was hypothesized and tested); (2) *evidence based on test content*, namely, alignment of content to student learning standards; and (3) *evidence based on relations with other variables*, and specifically the extent to which test scores and other measures intended to assess similar constructs provided convergent and predictive evidence. Reliability/precision evidence was based primarily on *internal-consistency coefficients*. Finally, fairness in testing was based primarily on evidence for *lack of measurement bias* using differential item functioning. The latter was in line with the RRSG (2002) call for assessments that would *not* reflect social, linguistic, or cultural variation in reading comprehension performance.

Taken together, the results of the iterative and extensive calibration and validation studies suggest that the GISA, RISE, and FRA assessments have defensible psychometric properties. Given the large number of features that were novel in the design of these assessment (e.g., expanded construct, item types, being web based, automated scoring, and being computer adaptive), this is no small feat. This is particularly important for GISA, which uses a scenario-based assessment design in a digital environment, and various themed texts and types of items across forms.

It is also important to note that the additional assessments developed by the other RfU teams (LARRC, CCDD, PACT, and READI) and reviewed in this report also met basic technical adequacy standards. These assessments were developed primarily to evaluate reading-related constructs and intervention effects, so the validation and calibration studies were not extensive.

Standardization and Efficiency

The three core assessments are characterized by standardization and efficiency. Specifically, both GISA and RISE (grades 3–12) are web administered and automatically scored (including selected constructed-response items). GISA takes 45 minutes to complete, whereas RISE takes 45–60 minutes. Both assessments make reference to reporting support that has the potential to be scalable at the classroom, school, or district level. It remains unclear, however, how researchers and practitioners can gain access to each of these assessments.

The FRA system consists of a K–2 system and a grades 3–12 system administered at three periods (fall, winter, and spring). Each system takes 45 minutes to complete. The K–2 system consists of screening, comprehension, and diagnostic tests that the teacher administers to students individually. The grades 3–12 system consists of screening and comprehension tests using web-based administration. The systems include reporting support that is scalable at the classroom, school, or district level. Notably, the FRA system is a computer-adaptive system; namely, the selection, order, and number of items administered depend on a student's response to the first item and each subsequent item of the assessment. Students receive harder or easier items based on their performance, and the system stops administering items once it has enough information about the student's ability (i.e., a small enough amount of error or uncertainty associated with a student's score). Thus, this adaptive assessment maximizes precision efficiency—the maximum of precision of information with a minimum of time spent gaining it (Mitchell, Truckenmiller, & Petscher, 2015).

The efficiency of these new assessments with respect to administration and scoring "raises the bar" for current testing practices. *The future research agenda needs to continue to explore approaches, methodologies, and technologies that will increase further standardization and efficiency of reading comprehension assessments.*

CONCLUSION

Historically, the assessment of reading comprehension is one of the most important outcomes of reform movements (Pearson & Hamm, 2005). Our evaluation is that the RfU research initiative had a profound impact on assessment, akin to that of reform movements. Collectively, the three core assessments developed—RISE, GISA, and FRA—can be characterized as a *new generation* of reading assessments. These assessments have a strong theoretical basis, reflect a broader and more authentic conceptualization of reading comprehension, are developmentally sensitive, emphasize instructional sensitivity and value, and have defensible psychometric properties. The calibration and validation studies were extensive, iterative, and undertaken across the United States. The result is a set of *forward-thinking assessments* that not only meet the standards of educational and psychological testing, but also promise to advance both research and practice in reading comprehension for years to come.

What the RfU has also contributed to the literature is a set of *additional measures of reading-related constructs* that are sensitive to high-quality instruction designed to improve different aspects of reading comprehension. These assessments also have a strong theoretical basis, reflect various aspects of reading comprehension or reading-related constructs, emphasize instructional sensitivity, and have defensible psychometric properties. This toolbox of measures enriches the range of possibilities available to researchers in the field of reading comprehension and enables measurement of aspects of reading comprehension that are contemporary and innovative (e.g., evidence-based argumentation, academic language, social perspective taking, online inference making). An important goal in a future research agenda would be further development, calibration, and scale-up of these measures to evaluate their practical utility.

The multiyear iterative efforts to develop these assessments also produced an incredible volume of empirical research that has used these assessments in small-scale, proof-of-concept studies; intervention studies; and large-scale calibration studies. The findings from the use of these assessments in various populations and contexts can inform further development and refinement of reading comprehension theories and models, help evaluate with better precision additional aspects of reading comprehension in younger and older readers, and help understand more deeply the implications of integrating important moderators (such as prior knowledge) into assessment design. An important goal in the future research agenda would be to use these assessments in place of more traditional standardized reading comprehension measures.

Advances in assessment influence instruction, and this new generation of assessments has the potential to transform current assessment practices and, thus, significantly influence instructional practices in reading comprehension. Finally, because these new assessments reflect some of the inherent complexities of the comprehension process that only now have been realized in assessment, they open new possibilities in the future research agenda that can significantly advance the field of reading comprehension.

REFERENCES

Ackerman, B. P. (1984). The effects of storage and processing complexity on comprehension repair in children and adults. *Journal of Experimental Child Psychology, 37*, 303–334.

Adlof, S. M., Catts, H. W., & Little, T. D. (2006). Should the Simple View of Reading include a fluency component? *Reading and Writing, 19*, 933–958. doi: 10.1007/s11145-006-9024-z.

AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (2014)*. Standards for educational and psychological testing*. Washington, DC: AERA.

Afflerbach, P., Cho, B. Y., & Kim, J. Y. (2015). Conceptualizing and assessing higher-order thinking in reading. *Theory into Practice, 54*(3), 203–212.

Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology, 44–45*, 68–82.

Albrecht, J. E., & O'Brien, E. J. (1991). Effects of centrality on retrieval of text-based concepts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 932–939.

Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1061–1070.

Armbruster, B. B., Anderson, T. H., & Meyer, J. L. (1991). Improving content-area reading using instructional graphics. *Reading Research Quarterly, 26*, 394–416. doi: 10.2307/747895.

Barnes, M. A., Faulkner, H., Wilkinson, M., & Dennis, M. (2004). Meaning construction and integration in children with hydrocephalus. *Brain and Language, 89*, 47–56.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637.

Barr, C. D., Uccelli, P., & Phillips Galloway, E. (2019). Specifying the academic language skills that support text understanding in the middle grades: The design and validation of core academic language skills construct and instrument. *Language Learning, 69*, 978–1021.

Barth, A. E., Barnes, M., Francis, D., Vaughn, S., & York, M. (2015). Inferential processing among adequate and struggling adolescent comprehenders and relations to reading comprehension. *Reading and Writing*, *28*(5), 587–609.

Bean, T. W., Singer, H., Sorter, J., & Frazee, C. (1986). The effect of metacognitive instruction in outlining and graphic organizer construction on students' comprehension in a tenth-grade world history class. *Journal of Reading Behavior, 18*(2), 153–169. doi: 10.1080/10862968609547562.

Bean, T. W., & Steenwyk, F. L. (1984). The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. *Journal of Reading Behavior, 16*(4), 297–306. doi: 10.1080/10862968409547523.

Beck, I. L., & McKeown, M. G. (1991). Conditions of vocabulary acquisition. In R. Barr, M. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 789–814). New York: Longman.

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

Beck, I. L., McKeown, M.G., & Kucan, L. (2008). *Creating robust vocabulary: Frequently asked questions and extended examples.* New York: Guilford.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*, 70–91. doi: 10.1080/15366367.2010.508686.

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

Bennett, R. E., & Gitomer, D. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). Dordrecht: Springer. doi: 10.1007/978-1-4020-9964-9_3.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

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

Bråten, I., Strømsø, H. I., & Britt, M. A. (2009). Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly, 44*(1), 6–28.

Britt, M. A., Rouet, J.-F., & Durik, A. M. (2018). *Literacy beyond text comprehension: A theory of purposeful reading.* New York: Routledge.

Britt, M. A., & Sommer, J. (2004). Facilitating textual integration with macro-structure focusing tasks. *Reading Psychology, 25*, 313–339. doi: 10.1080/02702710490522658.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cain, K., & Nash, H. M. (2011). The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology, 103*(2), 429–441. doi: 10.1037/a0022824.

Cain, K., & Oakhill, J. V. (1999). Inference making and its relation to comprehension failure. *Reading and Writing: An Interdisciplinary Journal, 11*, 489–503. doi: 10.1023/A:1008084120205.

Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is vocabulary more important for some aspects of comprehension? *L'Année Psychologique, 114*, 647–662.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96*(1), 31–42. doi: 10.1037/0022-0663.96.1.31.

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology, 96*(4), 671–681.

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing, 12*(3), 169–190.

Carrow-Woolfolk, E. (1995). *Oral and written language survey (OWLS).* Circle River, MN: American Guidance Service.

Carrow-Woolfolk, E. (2008). *Comprehensive assessment of spoken language*. Torrance, CA: Western Psychological Services.

Catts, H. W., Hogan, T. P., & Fey, M. E. (2003). Subgrouping poor readers on the basis of individual differences in reading-related abilities. *Journal of Learning Disabilities, 36*, 151–164. doi: 10.1177/002221940303600208.

Chall, J. S., Jacobs, V. A., & Baldwin, L. E. (1990). *The reading crisis: Why poor children fall behind*. Cambridge, MA: Harvard.

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.

Coleman, D., & Pimentel, S. (2011). *Revised publishers' criteria for the Common Core State Standards in English language arts and literacy, Grades 3–12.* Washington, DC: National Governors Association and Council of Chief State School Officers. Retrieved from http://www.corestandards.org/assets/Publishers_Criteria_for_3-12.pdf.

Compton, D. L., & Pearson, P. D. (2016). Identifying robust variations associated with reading comprehension skill: The search for pressure points. *Journal of Research on Educational Effectiveness, 9*, 223–231.

Conley, A. M., Pintrich, P. R., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology, 29*(2), 186–204.

Connor, C., Phillips, B., Kaschak, M., Apel, K., Kim, Y.-S., Al Otaiba, S., et al. (2014). Comprehension tools for teachers: Reading for understanding from prekindergarten through fourth grade. *Educational Psychology Review, 26*(3), 379–401. doi: 10.1007/s10648-014-9267-1.

Connor, C., Phillips, B., Kim, Y.-S., Lonigan, C., Kaschak, M., Crowe, E., Dombek, J., & Al Otaiba, S. (2018). Examining the efficacy of targeted component interventions on language and literacy for third and fourth graders who are at risk of comprehension difficulties. *Scientific Studies of Reading, 22*, 462–484.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Crosson, A. C., & Lesaux, N. K. (2013). Connectives: Fitting another piece of the vocabulary instruction puzzle. *The Reading Teacher, 67*(3), 193–200. doi: 10.1002/TRTR.1197.

Currie, N. K., & Cain, K. (2015). Children's inference generation: The role of vocabulary and working memory. *Journal of Experimental Child Psychology, 137*, 57–75. doi: 10.1016/j.jecp.2015.03.005.

Dahlberg, L. L., Toal, S. B., & Behrens, C. B. (1998). *Measuring violence-related attitudes, beliefs, and behaviors among youths: A compendium of assessment tools.* Atlanta, GA: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control.

Dawson, T. L. (2002). A comparison of three development stage scoring systems. *Journal of Applied Measurement, 3*, 146–189.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist, 47*(3), 153–176.

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

Denton, C. A., Enos, M., York, M. J., Francis, D. J., Barnes, M. A., Kulesz, P. A., et al. (2015). Text-processing differences in adolescent adequate and poor comprehenders reading accessible and challenging narrative and informational text. *Reading Research Quarterly, 50*, 393–416.

Denton, C. A., Wexler, J., Vaughn, S., & Bryan, D. (2008). Intervention provided to linguistically diverse middle school students with severe reading difficulties. *Learning Disabilities Research & Practice, 23*, 79–89.

Denton, C. A., Wolters, C. A., York, M. J., Swanson, E., Kulesz, P. A., & Francis, D. J. (2015). Adolescents' use of reading comprehension strategies: Differences related to reading proficiency, grade level, and gender. *Learning and Individual Differences, 37*, 81–95.

Diazgranados, S., Dionne, M. O., & Selman, R. L. (2011). *The social perspective taking acts measure (SPTAM).* Unpublished manuscript, Graduate School of Education, Harvard University, Cambridge, MA.

Diazgranados, S., Selman, R. L., & Dionne, M. (2015). Acts of social perspective taking: A functional construct and the validation of a performance measure for early adolescents. *Social Development, 25*, 572–601. doi: 10.1111/sode.12157.

Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the mini-mental state examination: An application of the Mantel Haenszel and standardization procedures. *Medical Care, 44*, S107–S114. doi: 10.1097/01.mlr.0000245182.36914.4a.

Douglas, K. M., & Albro, E. R. (2014). The progress and promise of the reading for understanding research initiative. *Educational Psychology Review, 26*(3), 341–355.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test-IV.* Circle Pines, MN: American Guidance Service.

Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading, 18*, 5–21. doi: 10.1080/10888438.2013.819356.

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 71*, 393–447. doi: 10.3102/00346543071003393.

Ehri, L. C., Nunes, S., Willows, D., Schuster, B., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly, 36*, 250–287.

Elleman, A. M. (2017). Examining the impact of inference instruction on the literal and inferential comprehension of skilled and less skilled readers: A meta-analytic review. *Journal of Educational Psychology, 109*, 761–781.

Ferguson, C. (1994). Dialect, register, and genre: Working assumptions about conventionalism. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 15–30). New York: Oxford University Press.

Fischer, K. W., & Bidell, T. R. (2006). Dynamic development of action, thought and emotion. In W. Damon & R. M. Lerner (Eds.), *Theoretical models of human development: Handbook of child psychology* (6th ed., Vol. 1, pp. 313–399). New York: Wiley.

Fitzgerald, J., Elmore, J., Koons, H., Hiebert, E. H., Bowen, K., Sanford-Moore, E., & Stenner, A. J. (2014). Important text characteristics for early-grades text complexity. *Journal of Educational Psychology, 93*, 3–22. doi: 10.1037/a0037289.

Fletcher, J. M. (2009). Measuring reading comprehension. *Scientific Studies of Reading, 10*, 323–330.

Foorman, B., Espinosa, A., Wood, C., & Wu, Y. (2016). *Using computer-adaptive literacy assessments to monitor the progress of English language learner students*. (REL 2016-149). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

Foorman, B. R., Francis, D. J., Davidson, K. C., Harm, M., & Griffin, J. (2004). Variability in text features in sex grade 1 basal reading programs. *Scientific Studies of Reading, 82*(2), 167–197.

Foorman, B., Herrera, S., Dombek, J., Schatschneider, C., & Petscher, Y. (2017). *The relative effectiveness of two approaches to early literacy intervention in grades K–12.* Retrieved from https://ies.ed.gov/ncee/edlabs/regions/southeast/pdf/REL_2017251.pdf.

Foorman, B. R., Petscher, Y., & Schatschneider, C. (2015a). *Florida Center for Reading Research (FCRR) Reading Assessment (FRA): Grades 3 through 12 technical manual.* Retrieved from http://www.fcrr.org/for-researchers/fra.asp.

Foorman, B. R., Petscher, Y., & Schatschneider, C. (2015b). *Florida Center for Reading Research (FCRR) Reading Assessment (FRA): Kindergarten to grade 2 technical manual.* Retrieved from http://www.fcrr.org/for-researchers/fra.asp.

Foorman, B. R., Petscher, Y., Stanley, C., & Truckenmiller, A. (2017). Latent profiles of reading and language and their association with standardized reading outcomes in kindergarten through tenth grade. *Journal of Research on Educational Effectiveness, 10*(3), 619–645. doi: 1080/19345747.2016.1237597.

Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2009). Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading, 10,* 301–322.

Freed, J., & Cain, K. (2017). Assessing school-aged children's inference making: The effect of test story format in listening comprehension. *International Journal of Language and Communications Disorders, 52*, 95–105. doi: 10.1111/1460-6984.12260.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*, 45–58.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239–256.

Gersten, R., Baker, S. K., Smith-Johnson, J., Dimino, J., & Peterson, A. (2006). Eyes on the prize: Teaching complex historical content to middle school students with learning disabilities. *Exceptional Children, 72*(3), 264–280.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*(3), 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*(1), 1–38.

Gillam, R. B., & Pearson, N. A. (2004). *Test of Narrative Language–Receptive.* Austin, TX: Pro-Ed.

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

Goldman, S., & Snow, C. E. (2015). Adolescent literacy: Development and instruction. In A. Pollatsek & R. Treiman (Eds.), *Handbook on reading* (pp. 463–478). New York: Oxford University Press.

Goldman, S., Snow, C. E., & Vaughn, S. (2016). Common themes in teaching reading for understanding: Lessons from three projects. *Journal of Adolescent and Adult Literacy, 60*, 255–264. doi: 10.1002/jaal.586.

Goldman, S. R., Greenleaf, C., & Yukhymenko-Lescroart, M. (with Brown, W., Ko, M., Emig, J., George, M., Wallace, P., Blum, D., Britt, M. A., & Project READI). (2019). Explanatory modeling in science through text-based investigation: Testing the efficacy of the READI intervention approach. *American Educational Research Journal, 56*(4), 1148–1216. doi: 10.3102/0002831219831041.

Goldman, S. R., McCarthy, K. S., & Burkett, C. (2014). Interpretive inferences in literature. In E. J. O'Brien, A. E. Cook, & R. F. Lorch (Eds.), *Inferences during reading* (pp. 386–415). New York: Cambridge University Press.

Goldman, S. R., & Pellegrino, J. W. (2015). Research on learning and instruction implications for curriculum, instruction, and assessment. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 33–41.

Goldman, S. R., & Rakestraw, J. S. (2000). Structural aspects of constructing meaning from text. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 311–335). Mahwah, NJ: Lawrence Erlbaum.

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

Graesser, A. C. (2015). Deeper learning with advances in discourse science and technology. *Policy Insights from the Behavioral and Brain Sciences, 2*, 42–50. doi: 10.1177/2372732215600888.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371–395.

Griffin, C. C., Malone, L. D., & Kameenui, E. J. (1995). Effects of graphic organizer instruction on fifth-grade students. *Journal of Educational Research, 89*, 98–107. doi: 10.1080/00220671.1995.9941200.

Haberman, S. J. (2005). *When can subscores have value?* (Research Report No. RR-05-08). Princeton, NJ: Educational Testing Service.

Halliday, M. A. K. (2004). *The language of science: Collected works of M.A.K. Halliday* (Vol. 5). New York: Continuum.

Hill, M. (1991). Writing summaries promotes thinking and learning across the curriculum—but why are they so difficult to write? *Journal of Reading, 34*(7), 536–639.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research, 67*(1), 88–140.

Hogan, T., Bridges, M. S., Justice, L. M., & Cain, K. (2011). Increasing higher level language skills to improve reading comprehension. *Focus on Exceptional Children, 44*(3), 1–20.

Hoover, W. A., & Gough, P. B. (1990). The Simple View of Reading. *Reading and Writing: An Interdisciplinary Journal, 2*, 127–160.

Hsin, L., & Snow, C. (2017). From theory of mind to social perspective-taking: A benefit of bilingualism in academic writing. *Reading and Writing: An Interdisciplinary Journal, 30*(6), 1193–1214.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. doi: 10.1080/10705519909540118.

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

Jones, S. M., LaRusso, M., Kim, J., Kim, H. Y., Selman, R., Uccelli, P., et al. (2019). Experimental effects of word generation on vocabulary, academic language, perspective taking, and reading comprehension in high poverty schools. *Journal of Research on Educational Effectiveness, 12*(3), 448–483.

Kane, M. T. (2013a). The argument-based approach to validation. *School Psychology Review, 42*(4), 448–457.

Kane, M. T. (2013b). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

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

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading comprehension: Core components and processes. *Policy Insights from the Behavioral and Brain Sciences, 3*, 62–69.

Kendeou, P., & O'Brien, E. J. (2015). Prior knowledge: Acquisition and revision. In P. Afflerbach (Ed.), *Handbook of individual differences in reading: Text and context* (pp. 151–163). New York: Routledge.

Kendeou, P., & O'Brien, E. J. (2018). Theories of text processing: A view from the top-down. In M. Schober, D. N. Rapp, & M. A. Britt (Eds.), *Handbook of discourse processes* (2nd ed.). New York: Routledge.

Kendeou, P., van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology, 101*(4), 765–778. doi: 10.1037/a0015956.

Kiili, C., Leu, D. J., Utriainen, J., Coiro, J., Kanniainen, L., Tolvanen, A., Lohvansuu, K., & Leppanen, P. H. T. (2018). Reading to learn from online information: Modeling the factor structure. *Journal of Literacy Research, 50*, 304–334.

Kim, H. Y., LaRusso, M., Hsin, L., Selman, R., & Snow, C. (2018). Social perspective-taking performance: Construct, measurement, and relations with academic performance and engagement. *Journal of Applied Developmental Psychology, 57*, 24–41.

Kim, J. S., Hemphill, L., Troyer, M., Thomson, J. M., Jones, S. M., LaRusso, M. D., & Donovan, S. (2017). Engaging struggling adolescent readers to improve reading skills. *Reading Research Quarterly, 52*(3), 357–382. doi: 10.1002/rrq.171.

Kim, Y. S. G. (2016). Direct and mediated effects of language and cognitive skills on comprehension of oral narrative texts (listening comprehension) for children. *Journal of Experimental Child Psychology, 141*, 101–120. doi: 10.1016/j.jecp.2015.08.003.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L. & Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Research Branch Report 8-75. Millington, TN: Naval Technical Training, U.S. Naval Air Station, Memphis, TN.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*, 164–182.

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

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Klingner, J. K. (2004). Assessing reading comprehension. *Assessment for Effective Intervention, 29*(4), 59–70.

Kolen, M. J., & Brennan, R. L. (2004). *Testing equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

LARRC (Language and Reading Research Consortium). (2015). The dimensionality of language ability in young children. *Child Development, 86*, 1948–1965. doi: 10.1111/cdev.12450.

LARRC. (2017). Oral language and listening comprehension: Same or different constructs*? Journal of Speech, Language, and Hearing Research, 60*, 1273–1284. doi: 10.1044/2017_JSLHR-L-16-0039.

LARRC & Muijselaar, M. (2018). The dimensionality of inference making: Are local and global inferences distinguishable? *Scientific Studies of Reading, 22*, 117–136.

LARRC, Jiang, H., & Logan, J. (2019). Improving reading comprehension in the primary grades: Mediated effects of a language-focused classroom intervention. *Journal of Speech, Language, and Hearing Research, 62*(8), 2812–2828.

LaRusso, M., Kim, H. Y., Selman, R., Uccelli, P., Dawson, T., Jones, S., et al. (2016). Contributions of academic language, perspective taking, and complex reasoning to deep reading comprehension. *Journal of Research on Educational Effectiveness, 9*(2), 201–222.

Lee, C. D., & Goldman, S. R. (2015). Assessing literary reasoning: Text and task complexities. *Theory into Practice, 54*(3), 213–227.

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

Leslie, L., & Caldwell, J. S. (2011). *Qualitative reading inventory* (5th ed.). Boston: Pearson.

Lipson, M. Y., Mosenthal, J. H., Mekkelsen, J., & Russ, B. (2004). Building knowledge and fashioning success one school at a time. *The Reading Teacher,* 534–542.

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

MacGinitie, W. H., & MacGinitie, R. K. (1988). *Gates MacGinitie Reading Test.* Itasca, IL: Riverside.

Madnani, N., Burstein, J., Sabatini, J., & O'Reilly, T. (2013). Automated scoring of a summary writing task designed to measure reading comprehension. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the eighth workshop on innovative use of NLP for building educational applications* (pp. 163–168). Atlanta, GA: Association for Computational Linguistics.

Martin, J., Sokol, B., & Elfers, T. (2008). Taking and coordination perspectives: From pre-reflective interactivity, through reflective inter-subjectivity, to meta-reflective sociality. *Human Development, 51*, 294–317. doi: 10.1159/000170892.

McCarthy, K., Guerrero, T., Kent, K., Allen, L., McNamara, D., Chao, S., et al. (2018). Comprehension in a scenario-based assessment: Domain and topic-specific background knowledge. *Discourse Processes, 55*(5–6), 510–524. doi: 10.1080/0163853X.2018.1460159.

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

McCrudden, M. T., & Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review, 19*, 113–139. doi: 10.1007%2Fs10648-006-9010-7.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*(3), 247–288. doi: 10.10807/s10648-006-9010-7.

McNamara, D. S., & Magliano, J. (2009). Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation* (pp. 297–384). New York: Academic Press.

Mead, G. H. (1938). *Mind, self and society: From the standpoint of a social behaviorist.* Chicago: University of Chicago Press.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–306). Westport, CT: American Council on Education/Praeger.

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

Mislevy, R. J., & Sabatini, J. P. (2012). How research on reading and research on assessment are transforming reading assessment (or if they aren't, how they ought to). In J. P. Sabatini, E. R. Albro, & T. O'Reilly (Eds.). *Measuring up: Advances in how we assess reading ability* (pp. 119–134). Lanham, MD: Rowman & Littlefield Education.

Mitchell, A. M., Truckenmiller, A., & Petscher, Y. (2015). Understanding computer adaptive assessments: Fundamentals and considerations for school psychologists. *Communique, 43*, 8.

Monte-Sano, C. (2010). Disciplinary literacy in history: An exploration of the historical nature of adolescents' writing. *Journal of the Learning Sciences, 19*, 539–568. doi: 10.1080/10508406.2010.481014.

Mullis, I. V. S., Martin, M. O, Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework.* Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. Retrieved from https://files.eric.ed.gov/fulltext/ED512410.pdf.

Nagy, W. E., & Scott, J. A. (2000). Vocabulary processes. In M. L. Kamil, P. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 269–284). Mahwah, NJ: Lawrence Erlbaum.

NCSS (National Council for the Social Studies). (2013). *The College, Career, and Civic life (C3) framework for social studies state standards: Guidance for enhancing the rigor of K–12 civics, economics, geography, and history.* Silver Spring, MD: NCSS.

NELP (National Early Literacy Panel). (2008). *Developing early literacy: Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy. Retrieved from https://lincs.ed.gov/publications/pdf/NELPReport09.pdf.

NGA & CCSSO (National Governors Association Center for Best Practices and Council of Chief State School Officers). (2010). *Common Core State Standards: College and career readiness standards for reading, writing and communication.* Washington, DC: Author. Retrieved from http://www.corestandards.org/ELA-Literacy.

NICHD (National Institute of Child Health and Human Development). (2000). *Report of the National Reading Panel. Teaching children to read: Report of the subgroups* (NIH Publication No. 00-4754). Washington, DC: U.S. Department of Health and Human Services.

NRC (National Research Council). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

NRC. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: The National Academies Press.

Oakhill, J., & Cain, K. (2012). The precursors of reading comprehension and word reading in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading, 16*, 91–121, doi: 10.1080/10888438.2010.529219.

OECD (Organisation for Economic Co-operation and Development). (2009a). *PISA 2009 assessment framework—Key competencies in reading, mathematics, and science.* Paris: Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd.org/document/44/0,3746,en_2649_35845621_44455276_1_1_1_1,00.html.

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

O'Reilly, T., & Sabatini, J. (2013). *Reading for understanding: How performance moderators and scenarios impact assessment design* (ETS Research Report No. RR-13-31). Retrieved from https://www.ets.org/research/topics/reading_for_understanding/publications.

O'Reilly, T., Sabatini, J., Bruce, K., Pillarisetti, S., & McCormick, C. (2012). Middle school reading assessment: Measuring what matters under a RTI framework. *Reading Psychology, 33*(1–2), 162–189. doi: 10.1080/02702711.2012.631865.

O'Reilly, T., Sabatini, J., & Wang, Z. (2018). Using scenario-based assessments to measure deep learning. In K. Millis, D. Long, J. Magliano, & K. Weimer (Eds.), *Deep comprehension: Multi-disciplinary approaches to understanding, enhancing, and measuring comprehension* (pp. 197–208). New York: Routledge.

O'Reilly, T., Sabatini, J., & Wang, Z. (2019). What you don't know won't hurt you, unless you don't know you're wrong. *Reading Psychology, 40*, 638–677.

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

O'Reilly, T., Wang, Z., & Sabatini, J. (2019). How much knowledge is too little? When a lack of knowledge becomes a barrier to comprehension. *Psychological Science, 30*, 1344–1351.

O'Reilly, T., Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing reading comprehension assessments for reading interventions: How a theoretically motivated assessment can serve as an outcome measure. *Educational Psychology Review, 26*, 403–424. doi: 10.1007/s10648-014-9269-z.

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology, 98*, 554–566. doi: 10.1037/0022-0663.98.3.554.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*(2), 117–175.

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

Pearson, P. D., & Hamm, D. N. (2005). The history of reading comprehension assessment. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13–70). Mahwah, NJ: Erlbaum.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81.

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

Perfetti, C. A., Rouet, J.-F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Mahwah, NJ: Lawrence Erlbaum Associates.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*, 22–37. doi: 10.1080/10888438.2013.827687.

Petscher, Y., & Foorman, B. R. (2011). *Summary of the predictive relationship between the FAIR and the FCAT in grades 3–10: 2010–2011.* Tallahassee, FL: Florida Center for Reading Research.

Phillips Galloway, E., & Uccelli, P. (2019). Examining developmental relations between core academic language skills and reading comprehension for English learners and their peers. *Journal of Educational Psychology, 111*(1), 15–31.

Pike, M. M., Barnes, M. A., & Barron, R. W. (2010). The role of illustrations in children's inferential comprehension. *Journal of Experimental Child Psychology, 105*, 342–355.

Quinn, J. M., Wagner, R. K., Petscher, Y., & Lopez, D. (2015). Developmental relations between vocabulary knowledge and reading comprehension: A latent change score modeling study. *Child Development, 86*(1), 159–175. doi: 10.1111/cdev.12292.

Ravid, D., & Tolchinsky, L. (2002). Developing linguistic literacy: A comprehensive model. *Journal of Child Language, 29*, 419–448.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*(2), 31–74.

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 Publishing.

RRSG (RAND Reading Study Group). (2002). *Reading for understanding: Toward a research and development program in reading comprehension.* Santa Monica, CA: RAND Corporation.

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

Sabatini, J., Bruce, K., Steinberg, J., & Weeks, J. (2015). *SARA reading components tests, RISE forms: Technical adequacy and test design, 2nd edition* (ETS Research Report No. RR-15-32). Princeton, NJ: Educational Testing Service.

Sabatini, J., O'Reilly, T., & Deane, P. (2013). *Preliminary reading literacy assessment framework: Foundation and rationale for assessment and system design* (ETS Research Report No. RR-13-30). Princeton, NJ: Educational Testing Service.

Sabatini, J., O'Reilly, T., Weeks, J., & Steinberg, J. (2016, April). *The validity of scenario-based assessment: Empirical results.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

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

Sabatini, J., Weeks, J., O'Reilly, T. Bruce, K. Steinberg, J., & Chao, Z. (2019). *SARA reading components tests, RISE forms: Technical adequacy and test design, 3rd edition* (Research Report No. RR-15-32). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12076.

Sabatini, J. P. (2017). *Reading for Understanding assessment network year 7 annual report.* Princeton, NJ: ETS.

Sabatini, J. P., & Bruce, K. (2009). *PIAAC reading component: A conceptual framework* (OECD Education Working Papers No. 33). Paris: OECD Publishing.

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

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

Sabatini, J. P., O'Reilly, T., Halderman, L. K., & Bruce, K. (2014a). Integrating scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disability Research & Practice, 29*(1), 36–43. doi: 10.1111/ldrp.12028.

Sabatini, J. P., O'Reilly, T., Halderman, L., & Bruce, K. (2014b). Broadening the scope of reading comprehension using scenario-based assessments: Preliminary findings and challenges. *L'Année psychologique, 114*, 693–723. doi: 10.4071/S0003503314004059.

Schatschneider, C., Petscher, Y., & Williams, K. M. (2008). How to evaluate a screening process: The vocabulary of screening and what educators need to know. In L. Justice & C. Vukelic (Eds.), *Every moment counts: Achieving excellence in preschool language and literacy instruction* (pp. 304–317). New York: Guilford Press.

Schraw, G. (2000). Reader beliefs and meaning construction in narrative text. *Journal of Educational Psychology, 90*, 705–715. doi: 10.1037/0022-0663.92.1.96.

Semel, E., Wiig, E., & Secord, W. (2003). *Clinical evaluation of language fundamentals—4 (CELF-4)*. San Antonio, TX: PsychCorp.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content area literacy. *Harvard Educational Review, 78*(1), 40–59.

Shin, J., Deno, S. L., & Espin, C. A. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *Journal of Special Education, 34*, 164–172. doi: 10.1177/002246690003400305.

Sinatra, G. M., Kienhues, D., & Hofer, B. K. (2014). Addressing challenges to public understanding of science: Epistemic cognition, motivated reasoning, and conceptual change. *Educational Psychologist, 49*(2), 123–138.

Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*, 21–28. doi: 10.1111/j.1745-3992.2007.00105.x.

Snow, C., Lawrence, J., & White, C. (2009). Generating knowledge of academic language among urban middle school students. *Journal of Research on Educational Effectiveness, 2*(4), 325–344.

Snow, C. E. (2018). Simple and not-so-simple views of reading. *Remedial and Special Education, 39*, 313–316. doi: 10.1177/0741932518770288.

Snow, C. E., & Sweet, A. P. (2003). Reading for comprehension. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 1–11). New York: Guilford Press.

Snow, C. E., & Uccelli, P. (2009). The challenge of academic language. In D. R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 112–133). New York: Cambridge University Press.

Snyder, L., Caccamise, D., & Wise, B. (2005). The assessment of reading comprehension: Considerations and cautions. *Topics in Language Disorders, 25*, 33–50.

Stanford Achievement Test Series Tenth Edition (SAT-10). *Pearson Assessment.* Pearson.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38*(6), 934–947. doi: 10.1037//0012-1649.38.6.934.

Swanson, E. A., Wanzek, J., McCulley, L., Stillman-Spisak, S., Vaughn, S., Simmons, D., et al. (2016). Literacy and text reading in middle and high school social studies and English language arts classrooms. *Reading & Writing Quarterly, 32*(3), 199–222.

Swanson, E. A., Wanzek, J., Vaughn, S., Roberts, G., & Fall, A.-M. (2015). Improving reading comprehension and social studies knowledge among middle school students with disabilities. *Exceptional Children, 81*, 426–442.

Taylor, B. M. (1982). Text structure and children's comprehension and memory for expository material. *Journal of Educational Psychology, 74*, 323–340.

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

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). *Test of word reading efficiency—second edition (TOWRE-2).* Austin, TX: Pro-Ed.

Uccelli, P., Barr, C. D., Dobbs, C. L., Phillips Galloway, E., Meneses, A., & Sanchez, E. (2015a). Core Academic Language Skills (CALS): An expanded operational construct and a novel instrument to chart school-relevant language proficiency in pre-adolescent and adolescent learners. *Applied Psycholinguistics, 5*, 1075–1107. doi: 10.1017/S0142716400006X.

Uccelli, P., Phillips Galloway, E., Aguilar, G., & Allen, M. (2019). Amplifying and affirming students' voices through CALS-informed instruction. *Theory into Practice, 59*, 75–88.

Uccelli, P., Phillips Galloway, E., Barr, C., Meneses, A., & Dobbs, C. (2015b). Beyond vocabulary: Core Academic Language Skills (CALS) that support text comprehension. *Reading Research Quarterly, 50*(3), 261–359. doi: 10.1002/rrq.10.

van den Broek, P., Bohn-Gettler, C., Kendeou, P., Carlson, S., & White, M. J. (2011). When a reader meets a text: The role of standards of coherence in reading comprehension. In M. T. McCrudden, J. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 123–140). Charlotte, NC: Information Age Publishing.

van den Broek, P., & Kendeou, P. (2014). Special issue guest editors' preface problems in reading comprehension: Connecting theory and practice. *Learning Disabilities Research & Practice, 29*(1), 2.

van den Broek, P., & Kendeou, P. (2017). Development of reading comprehension: Change and continuity in the ability to construct coherent representations. In K. Cain, D. Compton, & R. Parrila (Eds.), *Theories of reading development* (pp. 283–308). Amsterdam: John Benjamins.

van den Broek, P., Kendeou, P., Kremer, K., Lynch, J. S., Butler, J., White, M. J., & Lorch, E. P. (2005). Assessment of comprehension abilities in young children. In S. Stahl & S. Paris (Eds.), *Children's reading comprehension and assessment* (pp. 107–130). Mahwah, NJ: Erlbaum.

van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The landscape model of reading: Inferences and the 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: Lawrence Erlbaum Associates.

van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review, 98*, 3–53. doi: 10.1037/0033-295X.98.1.3.

Vaughn, S., Martinez, L. R., Linan-Thompson, S., Reutebuch, C. K., Carlson, C. D., & Francis, D. J. (2009). Enhancing social studies vocabulary and comprehension for seventh-grade English language learners: Findings from two experimental studies. *Journal of Research on Educational Effectiveness, 2*(4), 297–324.

Vaughn, S., Martinez, L. R., Wanzek, J., Roberts, G., Swanson, E. A., & Fall, A.-M. (2017). Improving content knowledge and comprehension for English language learners: Findings from a randomized control trial. *Journal of Educational Psychology, 109*, 22–34.

Vaughn, S., Roberts, G., Wexler, J., Vaugn, M., Fall, A.-M., & Schnakenberg, J. (2015). High school students with reading comprehension difficulties: Results of a randomized control trial of a two-year reading intervention. *Journal of Learning Disabilities, 48*(5), 546–558.

Vaughn, S., Swanson, E. A., Roberts, G., Wanzek, J., Stillman-Spisak, S., Solis, M., & Simmons, D. (2013). Improving reading comprehension and social studies knowledge in middle school. *Reading Research Quarterly, 48*, 77–93.

Vellutino, F., Tunmer, W., Jaccard, J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading, 11*, 3–32. doi: 10.1207/s1532799xssr1101_2.

von Davier, H., & Xu, X. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76*(2), 318–336. doi: 10.1007/s11336-011-9202-z.

Wagner, R.K., Torgesen, J., Rashotte, C.A., & Pearson N. (2010). *Test of Sentence Reading Efficiency and Comprehension.* Austin, TX: Pro-Ed.

Wang, Z., Sabatini, J., O'Reilly, T., & Weeks, J. (2019). Decoding and reading comprehension: A test of the decoding threshold hypothesis. *Journal of Educational Psychology, 111*, 387–401.

Wanzek, J., Swanson, E., Vaughn, S., Roberts, G., & Fall, M-A. (2016). English learner and non-English learner students with disabilities: Content acquisition and comprehension. *Exceptional Children, 82*, 428–442.

Wanzek, J., Swanson, E. A., Vaughn, S., Roberts, G., & Kent, S. C. (2015). Promoting acceleration of comprehension and content through text in high school social studies classes. *Journal of Research on Educational Effectiveness, 8*, 169–188.

Weeks, J. P. (2018). An application of multidimensional vertical scaling. *Measurement: Interdisciplinary Research and Perspectives, 16*, 139–154. doi: 10.1080/15366367.2018.1502005.

Wellborn, J. G., & Connell, J. P. (1987). *Manual for the Rochester Assessment Package for Schools.* Rochester, NY: University of Rochester.

Wigfield, A., Guthrie, J. T., Perencevich, K. C., Taboada, A., Klauda, S. L., McRae, A., & Barbosa, P. (2008). Role of reading engagement in mediating the effects of reading comprehension instruction on reading outcomes. *Psychology in the Schools, 45*, 432–445. doi: 10.1002/pits.20307.

Wood, P., & Kardash, C. (2002). Critical elements in the design and analysis of studies of epistemology. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 231–260). Mahwah, NJ: Lawrence Erlbaum Associates.

Woodcock, R. W., McGraw, K. S., & Mather, N. (2001). *Woodcock-Johnson III Test*. Itasca, IL: Riverside Publishing.

Yukhymenko-Lescroart, M. A., Briner, S. W., Magliano, J. P., Lawless, K., Burkett, C., McCarthy, K. S., et al. (2016). Development and initial validation of the Literature Epistemic Cognition Scale (LECS). *Learning and Individual Differences, 51*, 242–248.

Appendix 3-1[1]
Brief Reviews of Reading for Understanding Assessments

CORE ASSESSMENTS

- Global Integrated Scenario-Based Assessment (GISA)
- Reading Inventory and Scholastic Evaluation (RISE)
- Florida Center for Reading Research Reading Assessment (FRA)
- FRA K–2 System
- FRA Grades 3–12 System

ADDITIONAL ASSESSMENTS

- LARRC Inference Task
- Core Academic Language Skills Instrument (CALS-I)
- Assessment of Social Perspective-Taking Performance (ASPP)
- ASK Knowledge Acquisition Measure
- BRIDGE-IT Measure
- READI Literature Epistemic Cognition Scale (LECS)
- READI Evidence-Based Argument (EBA) Measure

GLOBAL INTEGRATED SCENARIO-BASED ASSESSMENT (GISA)

Conceptual Framework

GISA is designed to measure *global reading literacy*, the second dimension of the broader construct of *reading literacy* (O'Reilly et al., 2012; Sabatini & Bruce, 2009; Sabatini, Bruce, & Steinberg, 2013). Global reading literacy is defined as "the deployment of a constellation of cognitive, language, and social reasoning skills, knowledge, strategies, and dispositions, directed towards achieving specific reading purposes" (Sabatini, O'Reilly, & Deane, 2013, p. 7).

This assessment system uses a scenario-based design (Bennett, 2011; Bennett & Gitomer, 2009) that measures various levels of reading comprehension in a range of reading situations. Specifically, a scenario-based design provides test takers with a specific purpose for reading, a set of materials, and relevant questions. The scenario-based design is consistent with that of Cognitively-Based Assessment for, of, and as Learning (CBAL), a large Educational Testing Service (ETS) initiative that has been focusing on building innovative assessments in English language arts, math, and science (Bennett, 2010).

[1] These brief reviews were based on technical reports and other publications for each measure provided by the research teams at the time of writing this publication. The primary sources should be consulted for complete and detailed information.

GISA builds heavily on process models of reading comprehension that attempt to identify core processes that explain reading comprehension performance. In this context, the assessment team identified three principles to guide the assessment design (principles 4, 5, and 6; Sabatini, O'Reilly, & Deane, 2013). These principles state that reading is viewed as a purposeful activity (McCrudden, Magliano, & Schraw, 2011) that involves the construction of meaning at multiple levels, from literal to text base and situation models (Kintsch, 1998; McNamara & Kintsch, 1996); skilled reading includes proficiency in evaluating and synthesizing information across multiple texts in a digital environment (Britt & Sommer, 2004; Rouet & Britt, 2011); and reading growth involves the expansion of both knowledge and skills (RRSG, 2002). The focus in GISA is on higher-level reading comprehension rather than foundational reading skills. In this context, a set of moderators is expected to influence performance. These moderators—(1) background knowledge, (2) metacognitive and self-regulatory strategies, (3) reading strategies, and (4) student motivation and engagement—affect the interpretation of reading comprehension scores (O'Reilly & Sabatini, 2013). For this reason, the moderators are either directly assessed in the context of the assessment (this is the case for background knowledge and reading strategies) or integrated in the assessment design (this is the case for all four moderators).

Background knowledge provides an indicator of students' knowledge on the topic of the texts in the assessment. This is an important moderator because it can be used as an indicator of students' ability to learn, update, and apply information. In the assessment design this is accomplished by (a) measuring background knowledge directly, (b) providing access to additional content during the assessment (e.g., videos, audio, definitions, diagrams) that supports the test taker's background knowledge, and (c) structuring the sequence of sources (from general to specific) to facilitate knowledge building.

Metacognitive and self-regulatory strategies and behavior provide an indicator of students' ability to monitor their understanding and their ability to repair gaps, errors, and misconceptions. This is an important moderator because it can be used as an indicator of the accuracy of students' judgments of learning and ability to use available resources to solve problems and correct mistakes. In the assessment design this is accomplished by (a) setting goals for reading, (b) sequencing sources, (c) providing feedback/hints after an error, (d) evaluating peer responses in a simulated peer collaboration, and (e) accessing and using supplemental resources.

Reading strategies provide an indicator of students' strategic use of text. This is an important moderator because it can be used as an indicator of students' ability to use strategies such as paraphrasing and summarization. In the assessment design this is accomplished by including items that require students to (a) paraphrase, (b) summarize, and (c) graphically organize information.

Motivation and engagement provide an indicator of students' willingness to expend sufficient effort to understand text. This is an important moderator because it can be used as an indicator of students' interest on the topics and texts and engagement with specific tasks. In the assessment design this is accomplished by including goal-directed, authentic scenarios.

GISA includes a number of different item types, designed to integrate in the design information about the aforementioned moderators. These include constructed-response (CR), graphic organizer (GO), and multiple-choice (MC) item types. The CR items involve constructing summaries according to a rubric that requires the following: the first sentence must be about the entire text; the next three sentences must be about each of the paragraphs in the text; and students must use their own words and exclude their own opinions (Madnani, Burstein, Sabatini, & O'Reilly, 2013). The choice to include a summary type of CR item was motivated by evidence demonstrating that summarization enhances both comprehension (Bean & Steenwyk, 1984; Hill, 1991; Taylor, 1982) and metacognition (Thiede & Anderson, 2003). The GO items involve visualizing and understanding the organizational structure of a text. These items are always partially completed to help students understand the structure of each text (e.g., a 3 × 4 cell). The choice to include GO items was motivated by evidence suggesting that these tasks help construct coherent models of text content (Armbruster, Anderson, & Meyer, 1991; Bean, Singer, Sorter, & Frazee, 1986; Griffin, Malone, & Kameenui, 1995). The MC items involve higher-order processes such as evaluating sources, questions, perspectives, and quality of information.

In some test forms, it is possible to assess relevant *background knowledge* before, after, or before and after to evaluate learning from the assessment. For the background knowledge test students are asked to decide whether a term is *related* to the topic of the text (e.g., farming). Students can choose "Yes," "No," or "I don't know" and the instructions make clear that the section will not count toward their total reading score. Previous work has shown that this task is a quick and valid indicator of students' prior knowledge of the topic (Deane, 2012; O'Reilly et al., 2014).

Description

GISA measures a broader conception of reading literacy ability, consistent with cognitive models of reading comprehension. GISA is a 45-minute, web-administered, scenario-based assessment that includes authentic reading situations. Specifically, test takers are provided with a specific purpose for reading (e.g., studying for a test, preparing for a class presentation, etc.), a set of materials (e.g., websites, blogs, newspaper articles, etc.), and progress through the materials in a structured and scaffolded way. GISA examines the test taker's proficiencies in (a) constructing different levels of mental model representations (Kintsch, 1998), (b) familiarity with text structure and genre differences (Goldman & Rakestraw, 2000), (c) deployment of executive/metacognitive processes (Schraw, 2000), and (d) application of strategies for attaining a literacy goal (McCrudden & Schraw, 2007; van den Broek et al., 2011). Currently, there are 19 test forms available for grades 3–12 that include either science or history/language topics. The number of items varies across forms and ranges from 27 to 59. The types of items also vary across forms and include CR, GO, and MC. Each test form follows the same structure as described below.

GISA Forms (Sabatini, O'Reilly, Weeks, & Steinberg, 2016)

First, prior to reading the texts, students are presented with a scenario. For example (from the *Organic Farming* test form; Sabatini et al., 2014a):

> Your class has decided to create a website about organic farming to help members of the community become more familiar with the subject. The website will provide information to answer the following questions: What are the natural methods used in organic

> farming? How are these methods different from the methods used on non-organic, or conventional, farms? What are the pros and cons of organic farming? You will work with three classmates on the project.

The sources provided in this test form were texts on techniques used in organic farming, simulated results of a web search, advantages/disadvantages of organic farming, a simulated web discussion, cartoons, charts, and graphic organizers. Text readability ranged from grades 4–9 based on the Flesch-Kincaid readability formula (Kincaid, Fishburne, Rogers, & Chissom, 1975). This specific test form includes a total of 35 items as follows: 2 CR items summarizing text content; 7 GO items organizing text content; 3 MC items demonstrating detailed text understanding; 3 MC items demonstrating web source evaluation; 2 MC items and 4 GO items demonstrating evaluation of the advantages and disadvantages of organic farming; 5 MC items demonstrating word and sentence understanding by choosing synonyms in sentence context; and 9 MC items evaluating perspective taking and information quality in a simulated web discussion.

In this test form, relevant *background knowledge* is also being assessed. Students are asked to decide whether a term is *related* to the topic of farming. Students can choose "Yes," "No," or "I don't know" and the instructions make clear that the section will not count toward their total reading score.

GISA Administration and Scoring

GISA includes a computerized administration, automated scoring, and reporting support scalable at the classroom, school, or district level. A single score is produced that is subsequently scaled using 2PL/GPCM (Bock & Zimowski, 1997) and then rescaled. The scores are rescaled to have a mean of 1000 and a standard deviation of 100.

Sample

Evidence reported next (unless otherwise noted) is based on a large-scale field study that recruited students from all four regions of the United States: Midwest, Northeast, South, and West (Sabatini, O'Reilly, Weeks, & Steinberg, 2016). In this study, a total of 12,317 students in grades 3–12 participated. Specifically, there were 1,107 students in grade 3, 1,089 in grade 4, 1,178 in grade 5, 1,355 in grade 6, 1,403 in grade 7, 1,231 in grade 8, 1,401 in grade 9, 1,388 in grade 10, 1,153 in grade 11, and 1,012 in grade 12. In terms of ethnicity, 31.8 percent were Hispanic/Latino. In terms of race, 1.1 percent were American Indian/Native Alaskan, 2.9 percent Asian, 11.9 percent Black, 0.6 percent Native Hawaiian/Pacific Islander, 33.8 percent White, and 17.2 percent other/not reported. Also, 51.4 percent were female, 48.5 percent male, and 1 percent not reported. No other demographic information was reported other than the sample median of students receiving English-language learners' services (5 percent). Tests were administered in school computer labs and proctored by trained school staff members.

Reliability/Precision

Reliability/precision evidence for GISA was based on *internal-consistency coefficients*. Sabatini, O'Reilly, et al. (2016) computed Cronbach's alpha coefficients (Cronbach, 1951) for each test form across all grades (grades 3–12). The range of reliabilities were generally within acceptable range (.72. to .89).

Additional reliability coefficients have also been reported in several papers that evaluated specific forms of GISA. For example, Sabatini, O'Reilly, Halderman, and Bruce (2014b) performed a component reliability analysis for the Organic Farming GISA form, which included 35 items. A sample of 426 grade 6 students completed this form. Cronbach's alpha coefficient was α (426) = .89. The split-half reliability was $r(426) = .76$, with each half of the test also showing adequate alpha reliability ($\alpha = .80$ and $\alpha = .82$, respectively). Also, a subsample of 283 students was administered the same form again at the beginning of the next school year. Test-retest reliability was $r(283) = .87$. Reliabilities for items related to reader mental models (16 items; $\alpha = .78$), digital literacy (13 items; $\alpha = .78$), and other/vocabulary (8 items; $\alpha = .64$) were within acceptable range.

Validity

The validity argument for GISA integrates *evidence based on internal structure*, namely, the extent to which the relations among the test components conform to the hypothesized construct, and *evidence based on relations with other variables*, specifically the extent to which test scores and other measures intended to assess similar constructs provide convergent evidence, and the extent to which there was test-criterion predictive evidence.

Evidence Based on Internal Structure

Sabatini, O'Reilly, Weeks, and Steinberg (2016) theorized that the underlying literacy construct assessed by GISA is unidimensional. They evaluated unidimensionality in a large-scale study. To enable the creation of a vertical scale, a nonequivalent groups common item design (Kolen & Brennan, 2004) was used, which included at least two parallel forms in each grade (to be used as alternate forms in subsequent test administrations) and a linking form. Unidimensionality was evaluated in two ways: (1) factor analysis, and (2) item response theory (IRT) analysis. To evaluate unidimensionality using factor analysis, Sabatini et al. fit and compared three theoretically driven models: a unidimensional model, a two-factor exploratory model, and a two-factor simple-structure model where items associated with science passages loaded on one factor and items associated with history/language arts passages loaded on the second factor. Results from this comparison were mixed. The analysis showed that the unidimensional model fit better than either of the multidimensional models but only when the Bayesian information criterion (BIC) was used. Differences between the three models, however, were small overall. Although the indices provide mixed information, the penalty term is greater in the BIC compared to the Akaike information criterion (AIC). Due to the penalty difference, the BIC is a more conservative estimate and was deemed more appropriate for model selection. Subsequently, the unidimensional model was retained. Thus, the construct measured by the GISA across grades appears to be unidimensional.

On this basis, a unidimensional vertical scale was created using IRT analysis. To evaluate unidimensionality using IRT analysis, the item response curve for the two-parameter logistic (2PL) model (Birnbaum, 1968) was used to create a common scale.

The end result was a set of unidimensional vertical scales spanning grades 3–12. The item parameters for each scale were estimated using marginal maximum likelihood via a multigroup extension of the 2PL model (Bock & Zimowski, 1997), whereas the ability parameters were estimated using expected a posteriori. The item and ability parameters were estimated using the software program MDLTM (von Davier & Xu, 2011). As a final step, scores were rescaled to have a mean of 1,000 and a standard deviation of 100. Sabatini, O'Reilly, et al. (2016) examined all of the grade-level means and standard deviations and noted that they reflected developmental differences in ability across grades. In general, the means increased across all grades. The one exception was in grade 12, where the mean was slightly lower than the grade 11 mean.

Evidence Based on Relations to Other Variables

In the year 7 annual report, Sabatini (2017) reported preliminary findings from an integrated study design on the convergent evidence among RISE, GISA, CBAL, and the Gates-MacGinitie Reading Test (GMRT). The design was very complex and involved approximately 8,000 students. Preliminary results indicated that the correlation between the GISA and the GMRT was in the expected range. Specifically, the average correlation across different GISA forms and GMRT was r = .69 with the correlations ranging from .54 to .75. In the same report, Sabatini et al. also reported preliminary findings from an integrated study (*aka* the Mississippi Study) on the relations of GISA, FRA Reading Comprehension, and GMRT in elementary (grades K–2), middle (grades 3–5), and high school (grades 6–10) students. The correlations between GISA and FRA Reading Comprehension were in the moderate to high range (.483 to .777), whereas the correlations between GISA and GMRT were in the low to high range (.399 to .770). These moderately high correlations are very encouraging because they indicate that these reading comprehension tests are measuring a similar (but not identical) construct: reading.

Sabatini, O'Reilly, Halderman, and Bruce (2014a) provided preliminary evidence for predictive evidence in a small-scale study. In this study, n = 237 students in grade 6 were given the RISE battery which measured core reading skills such as word recognition, decoding, vocabulary, and morphology, as well as a pilot GISA form test (i.e., Organic Farming). The pattern of correlations among measures showed relatively strong relations (range r = .704 to .773), suggesting that all the component skills measured by the RISE are related to comprehension on GISA. As expected, the highest correlation was between the RISE Reading Comprehension subtest and GISA (r = .773), with RISE Reading Efficiency a close second (r = .762). A hierarchical, multiple regression analysis predicting GISA total scores from RISE subtest scores showed that each subtest added significant unique variance with an adjusted total of 69 percent of the variance in GISA accounted for by all the RISE subtests. Overall, the RISE and GISA robust correlations suggested that both batteries measure some overlapping aspects of reading comprehension across the ability range. However, there was also evidence that adequate lower-level skills may be necessary, but not sufficient prerequisites, to higher levels of reading performance as indicated by the regression analysis, providing further evidence that GISA measures complex comprehension.

Fairness

Sabatini, O'Reilly, et al. (2016) followed ETS Standards for fairness. In this context, every item was independently reviewed by ETS staff specifically trained in ensuring the fairness of test items. Evidence for fairness was also based on *lack of measurement bias*. Sabatini et al. (2016) examined effects of potential differential item functioning (DIF) as a function of gender across grades and forms. The DIF procedure determines whether any differential item performance exists between two groups matched for ability above and beyond expectations. The criteria for assessing the presence of DIF are based on Dorans and Kulick (2006) and have three levels based on values of the Mantel–Haenszel chi square statistic: A (negligible), B (moderate), and C (significant). The analysis showed very little presence of significant DIF (7 out of 708 items).

Proposed Intended Use of Scores

The review of GISA demonstrated evidence of careful test construction consistent with current conceptual frameworks of reading comprehension processes, appropriate administration and scoring, adequate score reliability, adequate evidence for validity based on test content, internal structure, and on relations with other variables, and attention to fairness with an emphasis on minimizing measurement bias.

GISA forms have some features that are different from existing standardized assessments. Among the most striking differences in design are the following. First, all assessments are contextualized within a scenario that provides a purpose for integrating multiple sources. Second, all assessments are delivered on computer, which allows for the assessment of "digital literacy" (Coiro, 2009). Third, the assessment uses simulated peers that provide instruction and guidance in "collaborating" with the test taker, making it a more authentic reading situation. Fourth, the assessments taps higher-level skills such as integration, evaluation, and application.

GISA Domain-Specific Assessments for Intervention Studies (O'Reilly et al., 2014)

With respect to its intended purposes, O'Reilly et al. (2014) demonstrated GISA's use as a summative assessment designed to provide evidence for the efficacy of reading interventions. While GISA forms have been developed and evaluated for different grade bands, topics, and skill foci, the GISA forms reported in this study were specifically designed to evaluate the outcomes of a specific intervention in mind: Reading Apprenticeship. The Reading Apprenticeship intervention views reading as an inquiry-based, problem-solving activity that builds knowledge about text content. For instance, reading in history involves evaluating facts and interpretations, the quality of sources (e.g., primary versus secondary), the corroboration of evidence, and an evaluation of the context in which information was collected (Monte-Sano, 2010). Reading in science involves using representations, models, and principles to reason and express key relationships among variables (Goldman, 2012). Reading in literature involves understanding human experience (Lee & Spratley, 2010). The underlying approach to assessment design in GISA (see O'Reilly & Sabatini, 2013; Sabatini & O'Reilly, 2013; Sabatini, O'Reilly, & Deane, 2013) had several elements in common with the Reading Apprenticeship program. Despite the similarities, the biggest difference between the

Reading Apprenticeship intervention and the GISA designs was the strong focus on content and disciplinary reading in high school history, science, and literature. Thus, three GISA summative forms were developed, intended to measure students' abilities to read and understand in each domain. Texts and tasks were sourced from topics in each domain. Each form also included an integrated background knowledge assessment following Deane (2012), in which students are asked to decide whether a term is *related or unrelated* to the topic of the text. Students can choose "Yes," "No," or "I don't know."

O'Reilly et al. (2014) analyzed data from a sample of 12,715 high school students in grades 9–12 from 43 schools in California and Pennsylvania. The three domain-specific forms exhibited good reliability (Cronbach's alpha range .84 to .88), adequate score variation, and positive correlations with measures of background knowledge. Furthermore, the results of a bifactor model suggested that there was a general reading comprehension factor underlying the domain-specific tests. Scores on the specific factors under the bifactor model were correlated at around .70 with the scores on the simple structure model. Thus, from a pure measurement standpoint, O'Reilly et al. (2014) concluded that the GISA assessments were adequate for use as outcomes in equations comparing treatment versus control students in the intervention.

Goldman et al. (2019) also used GISA as a distal measure in a randomized controlled trial evaluating the efficacy of the READI Science intervention to improve reading comprehension in grade 9 science when compared to a business-as-usual control. The results showed that GISA was sensitive to the intervention effects.

READING INVENTORY AND SCHOLASTIC EVALUATION (RISE)

Conceptual Framework

RISE is designed to measure foundational *components of reading*, one dimension of the broader construct of *reading literacy* (O'Reilly et al., 2012; Sabatini & Bruce, 2009; Sabatini, Bruce, & Steinberg, 2013; Sabatini, Weeks, et al., 2019). Reading component skills are subskills of reading that can be isolated and assessed independently from higher-level reading comprehension (Perfetti & Adlof, 2012; Sabatini, Bruce, & Steinberg, 2013). RISE builds heavily on component models of reading comprehension that attempt to identify core linguistic and cognitive skills to explain reading comprehension performance. In this context, the assessment team identified three principles to guide the assessment design (principles 1, 2, and 3; Sabatini, O'Reilly, & Deane, 2013). The first principle states that print skills and language comprehension are each considered necessary components of reading proficiency, though neither individually is sufficient to ensure proficiency (Adlof, Catts, & Little, 2006; Vellutino, Tunmer, Jaccard, & Chen, 2007). The second principle states that both breadth and depth of vocabulary knowledge are essential for understanding (Nagy & Scott, 2000; Ouellette, 2006). The third principle states that readers construct mental models of text meaning at multiple levels, from literal to gist to complex situation models (Kintsch, 1998; McNamara & Kintsch, 1996).

Consistent with these principles, components of the RISE inventory include foundational skills such as word recognition and decoding (Ehri, 2014), reading fluency (Fuchs et al., 2001), vocabulary knowledge (Beck & McKeown, 1991; Quinn et al., 2015), morphology (Carlisle, 2000; Hogan, Bridges, Justice, & Cain, 2011), syntax (Perfetti &

Adlof, 2012), and lower-level reading comprehension—at the sentence and single text level (Kintsch, 1998; McNamara & Kintsch, 1996). The range of these foundational skills—beginning with the recognition or decoding of words, to understanding the meanings of words and sentences, to building meaning from a passage—is consistent with the developmental progression of reading comprehension (RRSG, 2002).

This assessment system was designed for a finite set of purposes, including screening (i.e., identify students at risk of meeting grade-level expectations), diagnosis (i.e., identify students' strengths and weaknesses), formative assessment (i.e., provide actionable information for teachers), and summative assessment (i.e., provide accountability or outcome information) (Sabatini et al., 2015).

Description

RISE (Sabatini et al., 2015) is a 45- to 60-minute web-administered assessment of foundational reading skills in grades 3–12. The RISE is part of a larger reading assessment system called Study Aid and Reading Assistant (SARA). It contains six subtests, each of which targets a specific component of reading that may be affecting a student's progress toward higher levels of reading comprehension proficiency. Reading components are defined here as foundational subskills related to reading comprehension performance. Specifically, the RISE subtests target (a) decoding and recognizing words in isolation; (b) recognizing meaning or semantic relationships of individual words; (c) using knowledge of word parts to identify which word fits the meaning and syntax of a sentence; (d) building meaning from sentences by understanding causal connectors, pronouns, and relationships among terms; (e) reading for basic understanding with fluency; and (f) comprehending the basic meaning of passages. The initial scaling of RISE (Sabatini et al., 2015) had four forms in grade 5, six forms in grades 6–9, and three forms in grade 10. In the final scaling (Sabatini, Weeks, et al., 2019) these forms were reused with a national sample of students in grades 3–12; an additional form for grade 3 students was also developed. Thus, the final scaling includes one form in grade 3, four forms in grades 3–5, six forms in grades 6–9, and three forms in grades 9–12. Each subtest is described in more detail next.

The RISE Word Recognition and Decoding subtest uses three item types to measure a student's ability both to recognize sight words and to decode nonwords:

1. Real words, including content-area words that middle school students will encounter in their school curricula;
2. Nonwords, including a range of spelling and morphological patterns; and
3. Pseudohomophones, including nonwords that sound exactly like real English words.

Students are presented with one item on the screen at a time and are asked to decide if what they see (a) is a real word, (b) is not a real word, or (c) sounds exactly like a real word.

The RISE Vocabulary subtest includes both tier 2 and tier 3 words. Tier 2 words are general academic words, whereas tier 3 words are domain-specific, less frequently used words (Beck, McKeown, & Kucan, 2002, 2008; Coleman & Pimentel, 2011).

Students are presented with a target word on the screen and are asked to select either a synonym or a meaning associate of the target from three choices:

- An example of a synonym item is *data* (information, schedule, star).
- An example of a meaning associate item is *thermal* (heat, bridge, evil).

The RISE Morphology subtest focuses on derivational morphology—those words that have prefixes and/or suffixes attached to a root. The test uses a cloze (fill-in-the-blank) item type. Each item is a sentence. The sentences are designed with straightforward syntactic structures and relatively easy vocabulary so that students focus on the derived words:

- That man treats everyone with respect and __________. (civility, civilization, civilian)

The RISE Sentence Processing subtest focuses on single-sentence semantic and syntactic processing. The focus is on the student's ability to construct basic meaning from print at the sentence level. The cloze (fill-in-the-blank) items in the subtest require the student to process all parts of the sentence to select the correct answer among three choices:

- The dog that chased the cat around the yard spent all night __________. (barking, meowing, writing)

The RISE Efficiency of Basic Reading Comprehension subtest uses the maze selection technique (Fuchs & Fuchs, 1992; Shin, Deno, & Espin, 2000); that is, in each sentence within a passage, one of the words is replaced with three choices, only one of which makes sense in the sentence. Accurately selecting the correct response for each item does require that the reader is comprehending each sentence and likely building a cross-sentence general model of passage gist. Because the task is timed, the simultaneous demand that students read quickly also captures an indicator of silent reading fluency or efficiency. The subtest comprises informational texts. Students have 3 minutes to complete each passage.

- Passage excerpt: During the Neolithic Age, humans developed agriculture—what we think of as farming. Agriculture meant that people stayed in one place to grow their *crops/baskets/rings*. They stopped moving from place to place to follow herds of animals or to find new wild plants to *eat/win/cry.* And because they were settling down, people built permanent *shelters/planets/secrets*.

The RISE Reading Comprehension subtest assesses discourse-level comprehension. Students read a text and answer related question items. The items show a range of difficulties (from a verbatim understanding of the words and phrases to the "gist" understanding of what is being read and low-level inference making):

- Question (Locate/Paraphrase): What did people use to heat water in Neolithic houses? (*hot rocks, burning sticks, the sun, mud*)

- Question (Low-Level Inference): In the sentence "They gave people more protection from the weather and from wild animals." the word "they" refers to: (*permanent shelters, caves, herds, agriculture*)

RISE Administration and Scoring

RISE includes a computerized administration, automated scoring, and reporting support that is scalable at the classroom, school, or district level. Scores for each subtest provide evidence of instructionally malleable targets of readers' strengths and weaknesses. The item parameters for each scale are estimated using marginal maximum likelihood via a multigroup extension of the 2PL model (Bock & Zimowski, 1997). The scores for each scale are then rescaled to have a mean of 250 and a standard deviation of 15.

Sample

The technical quality of the RISE system was initially evaluated based on the findings reported by Sabatini et al. (2015) and subsequently Sabatini, Weeks, et al. (2019). Data were collected in a large, urban school district in the mid-Atlantic region of the United States. A total of n = 17,383 students in grades 5–10 participated. Specifically, there were n = 2,947 in grade 5, n = 3,540 in grade 6, n = 3,477 in grade 7, n = 3,114 in grade 8, n = 2,885 in grade 9, and n = 1,420 in grade 10. In terms of ethnicity, 3.6 percent were Hispanic/Latino. In terms of race, 0.3 percent were American Indian/Native Alaskan, 1.1 percent Asian, 87.7 percent Black, 0.2 percent Native Hawaiian/Pacific Islander, 10.7 percent White, and 0.2 percent other/not reported. Also, 51.4 percent were female, 48.5 percent male, and 1 percent not reported. No exclusions were mandated. In fact, 15.5 percent of the sample was receiving special education services and 1.3 percent English language learner services. Tests were administered in school computer labs and proctored by school staff members who were trained to the protocol. In their year 7 annual report, Sabatini (2017) noted that they conducted a large-scale field study in which they recruited students from all four regions of the United States: Midwest, Northeast, South, and West. Sample size increased to n = 51,391 and grade levels expanded from 4 to 12. Sabatini et al. stated that no meaningful differences were observed compared to those reported previously (Sabatini et al., 2015) and they planned on updating information about the sample as well as the assessment psychometric properties. Indeed, the most recent technical report (Sabatini, Weeks, et al., 2019) includes the updated information. Note that the results of the analyses reported next are based on both the 2015 and 2019 RISE technical reports.

Validity

The validity argument for RISE integrates *evidence based on test content*, namely, the relations between the content of the test and the construct it is intended to measure; *evidence based on internal structure*, namely, the extent to which the relations among the test components conform to the hypothesized construct; and *evidence based on relations with other variables*, and specifically the extent to which test scores and other measures intended to assess similar constructs provide convergent evidence.

Evidence Based on Test Content

Sabatini et al. (2015) theorized that each subtest construct represents a somewhat distinct component or subskill. Drawing on the extant reading literature (RRSG, 2002), it would be predicted that the various subtests would be moderately to strongly related (Mislevy & Sabatini, 2012). Indeed, the analysis showed moderate to strong correlations (Pearson's *r*) between the subtests within each grade level (grade 5 range .450 to .679; grade 6 range .504 to .718; grade 7 range .535 to .699; grade 8 range .522 to .699; grade 9 range .497 to .667; and grade 10 range .570 to .711).

Evidence Based on Internal Structure

Sabatini et al. (2015) used IRT (Lord & Novick, 1968), specifically the item response curve for the 2PL model (Birnbaum, 1968), to create a common scale for each subtest. The result was a set of six unidimensional vertical scales spanning grades 5–10. The item parameters for each scale were estimated using marginal maximum likelihood via a multigroup extension of the 2PL model (Bock & Zimowski, 1997), where the item parameters for the common items were constrained to be equal across groups. As a final step, the scores for all six scales were rescaled to have a mean of 250 and a standard deviation of 15. This analysis provided evidence for the hypothesized unidimensionality of each subscale/construct.

Sabatini, Weeks, et al. (2019) evaluated further the separation between the components across grades 3–12; three factor structures were considered. The first was a unidimensional structure where all the items loaded on a single factor. The second was a six-factor simple structure where the items associated with each component skill loaded only on the respective factor. The third was a two-factor simple view structure where the word reading, vocabulary, and morphology items loaded on one factor (*decoding*) and sentence comprehension, efficiency, and reading comprehension items loaded on the other factor (*comprehension*). The results suggested that both the two-factor and six-factor multidimensional structures had good fit to the data.

Evidence Based on Relations to Other Variables

Sabatini (2017) reported preliminary findings from a large-scale integrated study design on the relations among RISE, GISA, CBAL, and GMRT. Preliminary results indicated that the correlations between the GMRT and the RISE subtests were in the expected range. For instance, the correlation between the Gates-MacGinitie Vocabulary and the RISE Vocabulary subtest was $r(626) = .70$, $p < .01$. Similarly, the correlation between the Gates-MacGinitie Reading Comprehension test and the RISE Reading Comprehension subtest was $r(706) = .61$, $p < .01$. These correlations suggest that the assessments measure related, but not identical, constructs.

Sabatini, Weeks et al. (2019) also reported that the RISE vocabulary and morphology tests were correlated with the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 2012) $r = .36$ to .56, the Peabody Picture Vocabulary Test (PPVT) $r = .52$ to .57, and the Clinical Evaluation of Language Fundamentals (CELF) language measures $r = .38$ to .51. Also, RISE Reading Comprehension and GMRT correlation is .77, whereas RISE Reading Comprehension and GISA correlation is .65.

Reliability/Precision

Reliability/precision evidence was based on *internal-consistency coefficients*. Sabatini et al. (2015) reported Cronbach's alpha coefficients (Cronbach, 1951) for each subtest within each administration, form, and grade. The reliabilities represented as median values within a grade across forms were generally within acceptable range, specifically, for word recognition and decoding (range .899 to .921), for vocabulary (range .830 to .900), for morphology (range .871 to .920), for sentence comprehension (range .826 to .873), for reading efficiency (range .922 to .948), and for reading comprehension (range .604 to .833).

Sabatini et al. (2015) also evaluated subtest scores for consistency (versus the use of a total score) following the approach advocated by Haberman (2005) and Sinharay, Haberman, and Puhan (2007). The input information included Cronbach's alpha reliability values, average raw scores and standard deviations for each subtest, and the correlation between the subtest score and the total score. For purposes of this analysis, the total score was computed as the sum of the six subtest raw scores, and the total reliability coefficient was computed based on all item-level data across subtests merged together by unique student identifier. The analysis provided some evidence for subscore utility. Specifically, across 19 comparisons, 15 (79 percent) met the criteria for subscore utility. The four comparisons that did not meet the criteria involved grades 5 or 6 and three the reading comprehension subtest.

Fairness

Sabatini et al. (2015) followed ETS Standards for fairness. In this context, every item was independently reviewed by ETS staff specifically trained in ensuring the fairness of test items. Evidence for fairness was also based on *lack of measurement bias*. Specifically, Sabatini et al. (2015) examined effects of potential DIF by comparing item-level data for gender and race across grades and forms. The DIF procedure determines whether any differential item performance exists between two groups matched for ability above and beyond expectations. The criteria for assessing the presence of DIF were based on Dorans and Kulick (2006) and had three levels based on values of the Mantel–Haenszel chi square statistic: A (negligible), B (moderate), and C (significant). The analysis showed very little presence of significant DIF, suggesting no differential item performance as a function of gender or race. The updated analysis reported by Sabatini et al. (2019) using the national sample also showed very little presence of significant DIF.

Proposed Intended Use of Scores

The review of RISE demonstrated evidence of careful test construction consistent with current conceptual frameworks of reading comprehension components; appropriate administration and scoring; adequate score reliability; adequate evidence for validity based on test content, on internal structure, and on relations with other variables; and attention to fairness with an emphasis on minimizing measurement bias. With respect to its intended purposes, Sabatini, Weeks, et al. (2019) suggested that RISE could be used for diagnosis (i.e., identify students' strengths and weaknesses), formative assessment (i.e., provide actionable information for teachers), and summative assessment (i.e., provide accountability or outcome information). For example, with respect to diagnosis, RISE can detect whether foundational reading skills are barriers to achieving higher levels of reading comprehension performance. If foundational skills are lacking, then teachers should take this information into account when designing instruction to address student needs. Wang et al. (2019) identified a decoding threshold in RISE word recognition and demonstrated that students who initially fell below the threshold in the early grades showed little to no growth in reading comprehension over time.

With respect to evaluating instructional outcomes, RISE has been used in several large-scale intervention studies, demonstrating its instructional sensitivity. Specifically, Kim et al. (2017) used RISE to evaluate the efficacy of the Strategic Adolescent Reading Intervention (STARI) with low-achieving middle school students. The results showed that students who participated in STARI scored higher than control students on RISE efficiency of basic reading comprehension (Cohen's $d = 0.21$). In other words, the RISE was sensitive to the effects of the reading intervention.

FLORIDA CENTER FOR READING RESEARCH READING ASSESSMENT (FRA)

Conceptual Framework

FRA draws on decades of research about what predicts reading comprehension success in the English language system (NELP, 2008; NICHD, 2000; NRC, 1998; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001; RRSG, 2002). Specifically, in an alphabetic orthography such as English, *mastering the alphabetic principle*, namely, acquiring basic decoding skills, is a necessary skill that needs to be explicitly and systematically taught (Ehri, Nunes, Willows, et al., 2001). However, mastering the alphabetic principle is not a sufficient condition for understanding written text. Understanding written text (i.e., reading comprehension) also requires knowledge of *word meaning or lexical quality* (Perfetti & Stafura, 2014), namely, knowledge of pronunciation, spelling, multiple meanings in a variety of contexts, synonyms, and morphological structure. Understanding written text (i.e., reading comprehension) also requires *syntactic awareness*, namely, understanding of the rules that govern how words are ordered to make meaningful sentences.

The emphasis on achieving the alphabetic principle, lexical quality, and syntactic awareness ensures adequate reading comprehension. However, individual differences in readers' background knowledge, motivation, memory, and attention will also create variability in reading comprehension. Furthermore, because reading comprehension is affected by the interactions of variables related to reader and text characteristics (RRSG, 2002), text genre is also expected to influence performance.

In FRA K–2 the *alphabetic principle* is assessed with tasks that measure letter-sound knowledge, phonological awareness, ability to link sounds to letters, word reading, word building, and spelling. *Knowledge of word meanings* or lexical quality is measured by a word matching task. *Syntactic awareness* is assessed using a following directions task and a sentence comprehension task. *Reading comprehension* is assessed with a listening or reading passage comprehension task that includes both literary and informational passages (Fitzgerald et al., 2014; Foorman et al., 2004).

In FRA Grades 3–12 the *alphabetic principle* is assessed with a word recognition task. *Knowledge of word meanings* or lexical quality is measured by a vocabulary knowledge task that taps morphological awareness and includes words that signal inferential or decontextualized language. *Syntactic awareness* is assessed with syntactic knowledge tasks that taps the meaning and use of connectives (Cain & Nash, 2011; Crosson & Lesaux, 2013). *Reading comprehension* is assessed efficiently with a computer-adaptive reading passage comprehension task that includes both literary and informational passages (Fitzgerald et al., 2014; Foorman et al., 2004).

The FRA system consists of a K–2 system and a Grades 3–12 system administered at three periods (fall, winter, and spring). Each system consists of a series of tasks for which students receive five items at grade level and then additional tasks that the system adapts up or down in grade level based on performance to reach a precise estimate of a student's ability. The K–2 system consists of screening, comprehension, and diagnostic tasks that the teacher administers to students individually. The Grades 3–12 system consists of screening and comprehension tests that students complete online.

FRA is a computer-adaptive assessment system; namely, the selection, order, and number of items administered depend on a student's ability at the time of the assessment. Students receive harder or easier items based on their performance, and the system stops administering items once it has enough information about the student's ability. Thus, adaptive assessments maximize precision of information while minimizing time spent gaining it (Mitchell et al., 2015).

FRA K–2 SYSTEM

Description

The FRA K–2 system (Foorman et al., 2015a) is a 45-minute web-administered assessment of foundational reading skills. In the K–2 system the teacher scores the responses as correct or incorrect. The system is computer adaptive; namely, the selection, order, and number of items administered depend on a student's ability. FRA consists of six computer-adaptive tests, which evaluate students' phonological awareness, letter sounds, word reading, spelling, vocabulary, and following directions. These tasks collectively function as screening and produce the Probability of Literacy Success (PLS) score following a weighted formula. Students whose PLS score predicts that they are at risk of meeting grade-level expectations go on to take *Diagnostic* tasks. These computer-administered tasks are criterion referenced to developmental expectations for beginning readers and are scored for mastery (i.e., 80 percent correct). FRA also assesses comprehension using a listening/reading comprehension task and a sentence comprehension task.

Screening

There are six screening tasks. The *Phonological Awareness* task (K only) requires students to listen to a word that has been broken into parts and then blend them together to reproduce the full word. The *Letter Sounds* task (K only) presents students with a letter and asks them to provide the sound that the letter represents. The *Word Reading* task (grades 1 and 2) displays a word on a screen and students respond by reading the word out loud. The *Spelling Task* (grade 2) aurally presents a word and uses it in a sentence. Students respond by typing the word. The *Vocabulary Pairs* task (K–2) presents and pronounces three words. The student then selects the two words that go together best (e.g., dark, night, swim). The *Following Directions* task (K–2) requires students to listen and attend as they hear directions. Students respond to the directions by clicking on or moving the specified objects on the computer monitor (e.g., put the square in front of the chair and then put the circle behind the chair).

There are two comprehension tasks. The *Listening and Reading Comprehension* task (K–2) requires students to either listen to or read one passage and answer comprehension questions. Students are placed into listening or reading comprehension passages based on their performance on the *Screening* (and specifically the *Word Reading* task). Each passage has five multiple-choice questions. For each passage, the number of questions answered correctly, the number of words read correctly, and the words read correctly per minute are used in conjunction with the student's classroom performance to descriptively inform classroom instruction. The *Sentence Comprehension* task (K–2) requires students to select the one picture out of the four presented that depicts the sentence given by the computer (e.g., click on the picture of the bird flying toward the nest).

Administration and Scoring

In K–2, each task has four stop rules that determine when administration of each task is complete. Specifically: (a) a reliable estimate of the student's abilities is reached (i.e., standard error is less than 0.316); (b) the student has responded to 30 items (29 items in Letter Sounds); (c) the student responds correctly to all of the first eight items; and (d) the student responds incorrectly to all of the first eight items.

FRA produces three different scores. An *ability score* and *a percentile rank score* are provided for each computer adaptive task (Letter Sounds, Phonological Awareness, Word Reading, Vocabulary Pairs, Following Directions, Spelling, and Sentence Comprehension in K) at each time point. A PLS score is provided at each assessment period, which is an aggregate of the individual student's scores. In K the aggregate is based on Letter Sounds, Phonological Awareness, Vocabulary Pairs, and Following Directions. In grade 1 the aggregate is based on Word Reading, Vocabulary Pairs, and Following Directions. In grade 2, the aggregate is based on Word Reading, Vocabulary Pairs, Spelling, and Following Directions.

The PLS score indicates the likelihood that a student will reach end-of-year expectations in literacy. For the purposes of FRA, reaching expectations is defined as performing at or above the 40th percentile on the Stanford Achievement Test, Tenth Edition (SAT-10). The PLS is also color coded: red indicates the student is at high risk and needs targeted intervention, yellow indicates the student may be at risk and needs supplemental instruction, and green indicates the student is likely not at risk.

Ability scores provide an estimate of a student's development in a particular skill. The range is approximately 200 to 1,000, with a mean of 500 and standard deviation of 100. This score has an equal interval scale and is used to determine the degree of growth in a skill for individual students.

Percentile ranks vary from 1 to 99. The median percentile rank on FRA is 50. The percentile rank is an ordinal variable and is used to compare a student's performance to other students within a grade level.

Sample

Evidence reported next (unless otherwise noted) is based on a large-scale field study that recruited students in Florida. A total of 27,862 students in kindergarten through grade 2 across multiple districts in Florida participated in the calibration and validation studies (Foorman et al., 2015a). These studies involved students being administered subsets of items from each task depending on their grade level. Demographic information for the sample approximated that of the state of Florida: 40 percent White, 31 percent Hispanic, 23 percent Black, and 6 percent other; 65 percent eligible for free or reduced-price lunch; and 18 percent limited English proficient.

Validity

The validity argument for FRA K–2 integrates *evidence based on test content*, namely, the relations between the content of the test and the construct is intended to measure; *evidence based on internal structure*, namely, the extent to which the relations among the test components conform to the hypothesized construct; and *evidence based on relations with other variables*, and specifically the extent to which test scores provide convergent evidence and predict criterion performance.

Evidence Based on Test Content

The expectation was that oral language and reading measures would be moderated correlated with higher intercorrelations within each cluster. Indeed, FRA scores in K–2 were moderately interrelated (r = .20 to .78) with the highest correlations observed within oral language measures (e.g., Following Directions and Sentence Comprehension in K, r = .61) and reading measures (e.g., Spelling and Word Reading in grade 2, r = .78).

Evidence Based on Relations to Other Variables

Convergent evidence was provided by correlating performance on the FRA screening tasks with well-known clinical measures. Specifically, the FRA Phonological Awareness task scores in a low-performing sample of 100 English learners correlated r = .36 with the Letter-Word Identification task of the Woodcock-Johnson III Test of Achievement (Woodcock, McGrew, & Mather, 2001). FRA Letter Sounds correlated r = .52 with the Phonemic Awareness task of the Woodcock-Johnson III Test of Achievement (Woodcock et al., 2001). FRA Sentence Comprehension scores correlated r = .48 in K, r = .44 in grade 1, and r = .40 in grade 2 with the Sentence Structure subtest from the

CELF-4 (Semel, Wigg, & Secord, 2003). FRA Vocabulary Pairs scores correlated *r* = .46 in K, *r* = .59 in grade 1, and *r* = .50 in grade 2 with the PPVT-4 (Dunn & Dunn, 2007). FRA Following Directions scores correlated *r* = .58 in K, *r* = .58 in grade 1, and *r* = .64 in grade 2 with the CELF-4 Concepts and Following Directions (Semel et al., 2003).

Test-criterion predictive evidence was obtained in two ways. First, multiple regression analysis was used to estimate the total amount of variance that the linear combination of the FRA predictors explained in SAT-10 Word Reading in K and SAT Reading Comprehension in grades 1 and 2 (Foorman et al., 2015a). The analysis showed that FRA predicted a significant amount of variance at each grade (.46, .43, and .51, respectively). Sabatini (2017) also reported preliminary findings from an integrated study (*aka* the Mississippi Study) on the relations of FRA (K–2) with GISA and GMRT. The correlations between FRA (K–2), GISA, and GMRT were low to moderate (range .291 to .640). A series of regression analyses showed that FRA (K–2) accounted for 24.9 percent of variance in GISA and 54.8 percent of variance in GMRT.

Second, logistic regression analysis was used to estimate the predictive power of the PLS cutoff score. Recall that the PLS score is used to estimate the probability that a student is at risk of meeting grade-level expectations. This analysis focused on negative predictive power (Schatschneider, Petscher, & Williams, 2008), namely, the percentage of students who are identified as "not at risk" on the FRA screening but performing below benchmark on the outcome tests (< 40th percentile on SAT Word Reading and Reading Comprehension). The analysis evaluated PLS cutoff scores of .85 and .70, following previous work (Petscher & Foorman, 2011), and showed that a PLS score of .70 not only reduces false positives (range from .83 to .94), but also increases positive predictive power (range from .52 to .82) and the overall correct classification (range from .66 to .82).

Reliability/Precision

Across all grades and assessment periods, Foorman et al. (2015a) reported *marginal reliability coefficients* for the computer-adaptive tasks ranging from .85 to .96. *Test-retest reliability* was evaluated at three testing points: fall, winter, and spring. Across tasks and grade levels, correlations ranged between .42 to .80 in fall-winter, .44 to .72 in winter-spring, and .23 to .65 in fall-spring. The lowest correlations were consistently for the Vocabulary Pairs task.

Fairness

Evidence for fairness was based on *lack of measurement bias*. Specifically, the PLS cutoff score was evaluated for differential accuracy across different demographic groups. This procedure involved a series of logistic regressions predicting success on the SAT-10 tests (i.e., at or above the 50th percentile). The independent variables included a variable that represented whether students were identified as not at risk (PLS ≥ .70; coded as "1") or at risk (PLS < .70; coded as "0"), a variable that represented a selected demographic group, as well as an interaction term between the two variables. A statistically significant interaction term would suggest differential accuracy. For the combination of FRA screening task scores, differential accuracy was separately tested for Black and Latino students as well as for students identified as English language learners and students who were eligible for free or reduced-price lunch. These analyses showed no significant interactions and thus no differential effects.

Proposed Intended Use of Scores

The review of FRA K–2 demonstrated evidence of careful test construction consistent with current conceptual frameworks of reading comprehension, appropriate administration and scoring, adequate score reliability, adequate evidence for validity based on test content, internal structure, and on relations with other variables, and attention to fairness with an emphasis on minimizing measurement bias. With respect to intended use, Foorman and colleagues provided evidence for score appropriateness in evaluating the efficacy of interventions and identifying profiles of readers with instructional utility.

Evaluating Intervention Effects (Foorman, Herrera, et al., 2017)

In this study, the utility of FRA K–2 as a pre- and postintervention measure was evaluated in a randomized controlled trial in 55 low-performing schools across Florida that compared two pull-out early literacy interventions—one using standalone materials and one using materials embedded in the existing core reading program. The interventions were delivered daily for 45 minutes for 27 weeks in small groups of students at risk of literacy failure in K–2 for 2 consecutive years. A three-level hierarchical linear model with students nested in small groups, nested in schools, was used to estimate treatment effects by grade. The findings showed that the standalone intervention significantly improved grade 2 spelling outcomes relative to the embedded intervention, but impacts on other student outcomes were similar for the two interventions. On average, students in schools that used the standalone intervention and students in schools that used the embedded intervention showed similar improvement in reading and language outcomes. The two interventions also had similar impacts on reading and language outcomes among English learner students.

Identifying Latent Profiles with Instructional Utility (Foorman, Petscher, Stanley, & Truckenmiller, 2017)

This investigation had several aims, one of which was to determine the latent profiles of reading and language skills as measured by FRA and the extent to which these latent profiles were related to important reading outcomes, namely, SAT-10 Reading Comprehension (SESAT Word Reading for K). A total of 7,752 students in kindergarten through grade 10 across multiple districts in Florida participated in this study. Demographic information for the sample approximated that of the state of Florida: 42.18 percent White, 29.10 percent Hispanic, 22.5 percent Black, and 3.59 percent other; 60 percent eligible for free or reduced-price lunch; and 10.39 percent limited English proficient. There were 2,295 students in K–2. Latent profile analysis (LPA) identified five to six classes in the elementary grades. Profiles revealed high and low patterns in addition to interesting heterogeneous patterns (e.g., vocabulary deficit in K; vocabulary and word reading deficit in grade 1; word reading and spelling deficit in grade 2). These profiles have implications for differentiating instruction.

FRA GRADES 3–12 SYSTEM

Description

The FRA Grades 3–12 system is a 45-minute web-administered assessment. It includes four computer-adaptive tests, which evaluate students' word recognition, vocabulary knowledge, syntactic knowledge, and reading comprehension. *Screening* measures are the Word Recognition and the Vocabulary Knowledge tasks. *Diagnostic* measure is the Syntactic Knowledge task. Students are placed on a comprehension passage in the Reading Comprehension task based on their scores on the Word Recognition and Vocabulary Knowledge tasks.

Screening

There are two screening measures. The Word Recognition task presents a word to the student aurally and the student selects the correctly spelled word from three options. The Vocabulary Knowledge task presents one sentence at a time with a word missing. The missing word is replaced with a choice of three morphologically related words. The student selects the word that best completes the sentence.

Diagnostic

The Syntactic Knowledge task presents to the student one sentence (or sentences) aurally. Each sentence is missing one word. The computer also displays the sentence(s) for the student to read along. The student selects the missing word from a dropdown menu of three choices.

Reading Comprehension

The Reading Comprehension task presents students with a sample of one to three passages that are between 200 and 1,300 words in length. Each passage has seven to nine multiple-choice questions. All questions associated with the passage are displayed at the same time and the passage is also available during question answering.

Administration and Scoring

In grades 3–12 each task (except for Reading Comprehension) has four stop rules that determine when administration of each task is complete; specifically, (a) a reliable estimate of the student's abilities is reached (i.e., standard error is less than .50), (b) the student has responded to 30 items, (c) the student responds correctly to all of the first eight items, and (d) the student responds incorrectly to all of the first eight items.

FRA produces three different scores. An *ability score* and a *percentile rank score* are provided for each computer adaptive task (Word Recognition, Vocabulary Knowledge, Syntactic Knowledge, and Reading Comprehension) at each time point. A *probability* *of literacy success score* is provided at each assessment period, which is an aggregate of the individual student's scores. In grades 3–12 the aggregate is based on Word Recognition, Vocabulary Knowledge, and Reading Comprehension. The PLS score indicates the likelihood that a student will reach end-of-year expectations in literacy. For the purposes of FRA, reaching expectations is defined as performing at or above the 40th percentile on the SAT-10. The PLS is also color coded: red indicates the student is at high risk and needs targeted intervention, yellow indicates the student may be at risk and needs supplemental instruction, and green indicates the student is likely not at risk. The *ability score* provides an estimate of a student's development in a particular skill. The range is approximately 200 to 1,000, with a mean of 500 and standard deviation of 100. This score has an equal interval scale and is used to determine the degree of growth in a skill for individual students. Finally, *percentile ranks* vary from 1 to 99. The median percentile rank on FRA is 50. The percentile rank is an ordinal variable and is used to compare a student's performance to other students within a grade level.

Sample

Evidence reported next (unless otherwise noted) is based on a large-scale field study that recruited students in Florida. A total of 44,780 students in grades 3–10[2] across multiple districts in Florida participated in the calibration and validation studies (Foorman et al., 2015b). These studies involved students being administered subsets of items from each task depending on their grade level. Demographic information for the sample approximated that of the state of Florida: 41 percent White, 30 percent Hispanic, 23 percent Black, and 6 percent other; 60 percent eligible for free or reduced-price lunch; and 8 percent limited English proficient.

Validity

The validity argument for FRA Grades 3–12 integrates *evidence based on test content*, namely, the relations between the content of the test and the construct is intended to measure; *evidence based on internal structure*, namely, the extent to which the relations among the test components conform to the hypothesized construct; and *evidence based on relations to other variables*, and specifically the extent to which test scores provide convergent evidence and predict criterion performance.

Evidence Based on Test Content

FRA tasks were expected to be moderately correlated. Indeed, across grades FRA scores were moderately interrelated (range r = .29 to .63) with the highest correlations observed between reading comprehension and the three other measures (Vocabulary Knowledge, Word Recognition, and Syntactic Knowledge).

[2] The FRA team indicated that even though in their initial studies they also included grades 11 and 12, the sample is skewed toward lower-performing students. As a result they describe the sample as having a grade 3–10 proficiency range.

Evidence Based on Internal Structure

A series of parametric factor analyses by grade within each task were conducted. The comparative fit index (CFI), Tucker-Lewis index (TLI), and root mean square error of approximation (RMSEA) were used to evaluate model fit for the Vocabulary Knowledge, Word Recognition, and Syntax Knowledge tasks. CFI and TLI values of at least .90 are considered acceptable as are RMSEA values less than .10. With respect to the Vocabulary Knowledge, Word Recognition, and Syntax Knowledge tasks the results provided support for a unidimensional construct in each case. RMSEA values ranged between 0.000 and 0.028, CFI between 0.89 and 1.00, and TLI between 0.88 and 1.00 across grades. For the Reading Comprehension task, a unidimensional model was compared to a testlet model using the AIC and BIC indices. Results from this comparison were mixed. The AIC suggested that the testlet model should be used while the BIC and adjusted BIC values were smaller for the unidimensional model. Although the indices provided mixed information, the penalty term was greater in the BIC compared to the AIC. Due to the penalty difference, the BIC is a more conservative estimate and was deemed more appropriate for model selection. Subsequently, the unidimensional model was retained.

Evidence Based on Relations to Other Variables

A study that involved n = 1,825 students in grades 3–10 was used to provide convergent evidence. Students were administered the FRA tasks and well-known clinical measures. These measures included the TOWRE (Torgesen, Wagner, & Rashotte, 2012), the PPVT-4 (Dunn & Dunn, 2007), and the Grammaticality Judgment Test of the Comprehensive Assessment of Spoken Language (GJT; Carrow-Woolfolk, 2008). The analyses showed that the average correlation between the FRA Vocabulary Knowledge task and the PPVT-4 was r = .52 (range of .47 to .67); that of the FRA Word Recognition task and the TOWRE Real Word test was r = .33 (range of .24 to .49); that of the FRA Word Recognition task and the TOWRE Non-Word test was r = .38 (range of .30 to .47); and that of the FRA Syntax Knowledge task and the GJT was r = .49 (range of .37 to .61). Convergent evidence was also reported for students with low (< 40th quantile), average (40th to 60th quantile), and high (< 60th quantile) scores using quantile correlation analysis. The quantile correlations demonstrated a trend that higher correlations between the measures were observed for students who scored low or average on each measure.

Discriminant evidence was provided by estimating correlations between the FRA tasks and variables such as sex and birth date. The results showed overall weak relations across grades for both sex (range −.26 to .22) and birthdate (range .01 to .28).

Test-criterion predictive evidence was obtained in two ways (Foorman et al., 2015b). First, multiple regression analysis was used to estimate the total amount of variance that the linear combination of the FRA predictors explained in SAT-10 Reading Comprehension. The analysis showed that FRA predicted a significant amount of variance at each grade (range from .39 to .62). Sabatini (2017) reported preliminary findings from an integrated study (*aka* the Mississippi Study) on the relations of FRA (3–12) with GISA and GMRT. The correlations between FRA (3–12), GISA, and GMRT were moderate to high at middle (range .475 to .716) and high school (range .419 to .777) levels. A series of regression analyses showed that FRA accounted for 52.5 percent and 57.3 percent of variance in GISA middle and high school levels, respectively. Also, FRA accounted for 62.9 percent and 60.2 percent of variance in GMRT middle and high school levels, respectively.

Second, logistic regression analysis was used to estimate the predictive power of the PLS cutoff score. Recall that the PLS score is used to estimate the probability that a student is at risk of meeting grade-level expectations. This analysis focused on negative predictive power (Schatschneider et al., 2008), namely, the percentage of students who are identified as "not at risk" on the screening assessment (FRA) but performing below benchmark on the outcome tests (< 40th percentile on SAT Reading Comprehension). The analysis evaluated cutoff scores of .85 and .70 following previous work (Petscher & Foorman, 2011) and showed that a PLS score of .70 not only reduces false positives (range from .84 to .91), but also increases positive predictive power (range from .45 to .68) and overall correct classification (range from .71 to .86).

Reliability/Precision

Across all grades and assessment periods, Foorman et al. (2015b) reported average *marginal reliabilities* for the computer-adaptive tasks ranging from .86 to .93. *Test-retest reliability* was evaluated at three testing points: fall, winter, and spring. Across tasks and grade levels, correlations ranged between .46 to .85 in fall-winter, .51 to .80 in winter-spring, and .31 to .80 in fall-spring. The lowest correlations were consistently for fall-spring, which was expected as students' performance differentially changes from the beginning to the end of the year.

Fairness

Evidence for fairness was based on *lack of measurement bias*. Specifically, the PLS cutoff score was evaluated for differential accuracy across different demographic groups. This procedure involved a series of logistic regressions predicting success on the SAT-10 test (i.e., at or above the 50th percentile). The independent variables included a variable that represented whether students were identified as not at risk (PLS ≥ .70; coded as "1") or at risk (PLS < .70; coded as "0"), a variable that represented a selected demographic group, as well as an interaction term between the two variables. A statistically significant interaction term would suggest differential accuracy. For the combination of FRA screening task scores, differential accuracy was separately tested for Black and Latino students as well as for students identified as English language learners and students who were eligible for free or reduced-price lunch. These analyses showed only one significant interaction between the PLS cut point and minority status in grade 4 (p = .005) such that White students with a PLS above the cut point had a greater chance of being at or above the 50th percentile on the SAT-10 compared to Black students above the cut point on the PLS. The researchers noted the need to replicate this effect before definitive conclusions can be drawn.

In a subsequent study, Foorman, Espinosa, Wood, and Wu (2016) examined the appropriateness of FRA for English learner students. A sample of n = 102 English learner students in grades 3–5 participated. The students were classified as English levels 1 and 2 based on district-determined ranges of ability scores on the Comprehensive English

Language Learning Assessment for grades 3–5. The study showed that it was feasible for teachers to use FRA score reports and graphs to note students' strengths and weaknesses in oral language and reading and to differentiate instruction. They also used scores to monitor student progress, make instructional adjustments as needed, and report progress to parents.

Proposed Intended Use of Scores

The review of FRA Grades 3–10 demonstrated evidence of careful test construction consistent with current conceptual frameworks of reading comprehension; appropriate administration and scoring; adequate score reliability; adequate evidence for validity based on test content, on internal structure, and on relations to other variables; and attention to fairness with an emphasis on minimizing measurement bias. With respect to intended use, Foorman and colleagues provided evidence for score appropriateness in identifying profiles of readers with instructional utility.

Identifying Latent Profiles with Instructional Utility (Foorman, Petscher, et al., 2017)

This investigation had several aims, one of which was to determine the latent profiles of reading and language skills as measured by FRA and the extent to which these latent profiles were related to important reading outcomes, namely, SAT-10 Reading Comprehension (SESAT Word Reading for K). A total of 7,752 students in kindergarten through grade 10 across multiple districts in Florida participated in this study. Demographic information for the sample approximated that of the state of Florida: 42.18 percent White, 29.10 percent Hispanic, 22.5 percent Black, 3.59 percent other; 60 percent eligible for free or reduced-price lunch; and 10.39 percent limited English proficient. There were 5,457 students in grades 3–10. LPA identified three classes. Profiles in grades 3–10 followed a high, medium, and low pattern. In all grades, the latent profiles were significantly related to the reading outcome scores, explaining from 24 percent to 61 percent of the variance, with the mode being 42 percent. These profiles have possible implications for differentiating instruction.

LARRC INFERENCE TASK

Conceptual Framework

The LARRC team was particularly interested in the dimensionality of language (LARRC, 2015) and aimed to assess different levels of receptive and expressing language (i.e., single word, sentence, and discourse levels). In this context, the team developed the Inference Making task to evaluate discourse-level language comprehension following the work of Cain and Oakhill (1999) and Oakhill and Cain (2012). Thus, the LARRC Inference Making task builds heavily on process models of reading comprehension. Inference making is necessary to establish both local and global coherence (Graesser, Singer, & Trabasso, 1994) during comprehension of written or spoken text (Kintsch & van Dijk, 1978). Local coherence inferences are necessary in order to integrate information from adjacent pieces of text, whereas global coherence inferences are used to fill in details not explicitly stated that are needed to construct a globally coherent representation of text meaning (Cain & Oakhill, 1999, 2014; Currie & Cain, 2015; Freed & Cain, 2017). Inference making in general is a critical skill to successful reading and listening comprehension both concurrently and longitudinally, over and above cognitive factors such as general ability and memory (Cain, Oakhill, & Bryant, 2004; Elleman, 2017; Kim, 2016; Oakhill & Cain, 2012).

Description

The Inference Making task was developed to assess global and local inference-making skills during listening comprehension in children in pre-kindergarten (pre-K) through grade 3. The task includes two stories at each grade level, each one followed by eight questions to assess the ability to generate local and global coherence inferences (four questions each for local and global coherence inferences per text). The stories and questions were based on the work of Cain and Oakhill (1999) and Oakhill and Cain (2012). The second story at each grade level was repeated at the subsequent grade, such that there was one unique story at each grade. Students were read each story and then asked inferential questions.

Story Excerpt:

Today was Grandma's birthday. The family was getting ready for the party. Dad and Josh were putting up the party tent in the back lawn. Mom told them to put on some sunscreen, so that they didn't burn. Mom drove over to pick up Grandma, who lived an hour away. Mom told Linzie to keep an eye on the cake in the oven and to make some fruit punch.

Sample Questions:

- What were the family getting ready for?*
  Answer: Grandma's (birthday) party (2 points); a party (1 point); to go out (0 points)

- What was the weather like?
  Answer: (hot and) sunny (2 points); warm (1 point); rainy (0 points)

Administration and Scoring

The task is individually administered and scored. In this task, children listen to two narrative passages read aloud and are asked a series of inference-based questions. Children's responses are audio-recorded and postscored. Questions are scored as either correct (2 points), partially correct (1 point), or incorrect (0 points) using a rubric. The total score is the average score on all questions.[3]

Sample

The LARRC Inference Making task was evaluated using a sample of participants from the larger longitudinal study on listening and reading comprehension conducted in the context of the Reading for Understanding (RfU) research initiative (LARRC, 2017; LARRC & Muijselaar, 2018). Participants were 416 pre-kindergartners (241 boys, *M* = 5 years and 1 month, *SD* = 4.33 months), 520 kindergartners (289 boys, *M* = 6 years and 1 month, *SD* = 3.93 months), 620 first graders (324 boys, *M* = 7 years and 1 month, *SD* = 4.10 months), 724 second graders (380 boys, *M* = 8 years and 1 month, *SD* = 4.19 months), and 783 third graders (400 boys, *M* = 9 years and 1 month, *SD* = 4.10 months). Children in each grade level were selected from research sites in Arizona, Kansas, Nebraska, and Ohio. It is important to note that the sample was predominantly White (83–94 percent across grades); most with high income level (12.8 percent < $30,000, 25.3 percent $31,000–$60,000, 61.9 percent > $60,000); and 14.6 percent on free or reduced-price lunch.

Validity

Evidence Based on Internal Structure

Several confirmatory factor models that assumed unidimentionality of inference making but accounted for text and coherence factors to various degrees were tested. Three models were directly compared: (1) a one-factor model in which all items loaded on a general inference-making factor; (2) a bifactor model in which all items loaded on a general inference making factor, and in addition, on the text to which they belonged; and (3) multitrait, multimethod (MTMM) model in which each item loaded on a local or global inference factor, in addition to the loadings on the general factor and one of the text factors. The latent factors in all models were specified to be uncorrelated. The fit of the models was evaluated with inspection of three indices: the chi-square goodness-of-fit test statistic, the RMSEA, and the CFI (Kline, 2011). A nonsignificant chi square indicated good overall model fit, whereas a significant chi square showed poor fit. The ratio χ^2/df was also used to evaluate model fit. A χ^2/df ratio < 2 confirmed a good fit. A model with an RMSEA below .05 has a good approximate fit, an RMSEA between .05 and .08 was taken as satisfactory approximate fit, and values above .10 indicated poor approximate model fit (Browne & Cudeck, 1993). A model with a CFI larger than .95 had a good incremental fit to the data, and a CFI larger than .90 was taken as acceptable (Hu & Bentler, 1999). Differences between nested models were tested with the corrected chi-square difference test (with Satorra-Bentler correction) (Kline, 2011). These analyses showed that, across grades, the MTMM model had the best fit to the data. The general factor explained most of the variance in the items, whereas the latent text and inference factors explained little additional variance. This suggests that the

[3] The local and global subscores were also evaluated for reliability and validity but were not deemed adequate (LARRC & Muijselaar, 2018).

construct of inference making is broadly unidimensional. Furthermore, even though it is important to account for text and type of inference, the additional explanatory power of these factors is limited.

Evidence Based on Relations to Other Variables

Convergent evidence was evaluated with a series of correlation analysis between the Inference Making task scores and scores on the Listening Comprehension Measure (LCM) from the Qualitative Reading Inventory–Fifth Edition (QRI-5; Leslie & Caldwell, 2011), the CELF-4 Subtest Understanding Spoken Paragraphs (USP; Semel, Wiig, & Secord, 2003), and the Test of Narrative Language–Receptive (TNL; Gillam & Pearson, 2004). Across grades, correlations of the Inference Making task with LCM from QRI-5 ranged from .48 to .69; with the USP from CELF-4 ranged from .37 to .61; and with the TNL ranged from .40 to .72. These moderate to high correlations suggest that the Inference Making task is a valid measure of listening comprehension.

Reliability/Precision

The *internal consistency coefficients* of the test at each grade level were acceptable. Cronbach's alphas for pre-K = .78, kindergarten = .64, grade 1 = .71, grade 2 = .74, and grade 3 = .69. Test-retest reliability was evaluated with correlations for consecutive years. These were consistently moderate (pre-K to kindergarten, *r* = .63; kindergarten to grade 1, *r* = .58; grade 1 to grade 2, *r* = .56; and grade 2 to grade 3, *r* = .54).

Proposed Intended Use of Scores

This experimenter-developed Inference Making task is a reliable and valid measure to assess discourse listening comprehension in pre-K through grade 3. LARRC and Muijselaar (2018) suggest that the Inference Making task could be used as a measure for general listening comprehension or as a measure of discourse narrative comprehension with a focus on inference making. Indeed, the LARRC team included this measure as one of the main dependent variables in a randomized controlled trial designed to evaluate the efficacy of a language-based comprehension instruction in pre-K through grade 3.

CORE ACADEMIC LANGUAGE SKILLS INSTRUMENT (CALS-I)

Conceptual Framework

The Catalyzing Comprehension Through Discussion and Debate (CCDD) team proposed an expanded operationalization of academic language skills, namely, skills that involve understanding the meanings of words and the syntactic and discourse constructions in which they are embedded (Halliday, 2004; Snow & Uccelli, 2009). The focus on academic language proficiency was driven by evidence that it may be one key source of difficulty in accessing the meaning of texts, particularly in preadolescents and adolescents. The construct Core Academic Language Skills or CALS was defined as "knowledge and deployment of a repertoire of language forms and functions that co-occur with oral and written school learning tasks across disciplines" (Uccelli et al., 2015a, p. 1). Instead of focusing on discipline-specific language proficiency, CALS focus on the high-utility language skills hypothesized to support reading comprehension *across content areas.* Instead of focusing only on English learners, as most prior research on academic language proficiency had, CALS were hypothesized to be significant contributors of reading comprehension also for English-proficient students.

The work on CALS is situated in a sociocultural pragmatics-based view of language, which views language as inseparable from its social context and posits that language continues to develop throughout adolescence and even adulthood as people continue to learn new ways of using language to navigate more social contexts (Uccelli et al., 2015a, 2015b). During adolescence, language development entails developing "rhetorical flexibility" (Ferguson, 1994; Ravid & Tolchinsky, 2002), defined as the ability to use lexicogrammatical and discourse resources appropriately and flexibly in a variety of social contexts.

Description

The CALS Instrument (CALS-I) is a researcher-designed group-administered instrument for students in grades 4–8 that measures CALS. CALS are operationalized as a set of skills that correspond to linguistic features prevalent in academic texts across content areas yet are rare in colloquial conversations. This set of skills was hypothesized to support academic reading across school content areas and to encompass the following nonexhaustive domains:

- *Unpacking dense information*: skill in unpacking dense information in academic texts at the word and sentence levels:
  - decomposing complex words (e.g., decomposing nominalizations: *invasion* > *invade*); and
  - understanding complex sentences (e.g., extended noun phrases, embedded clauses).
- *Connecting ideas*: skill in comprehending connectives used to signal relations between ideas in academic texts (e.g., *consequently*, *in contrast*, *in other words*).
- *Tracking themes*: skill in identifying terms or phrases used to refer to the same participants or themes throughout an academic text, specifically tracking conceptual anaphors, those that refer to a complex concept mentioned in a different part of the text (e.g., *Water evaporates at 100 degrees Celsius. This process . . .*).
- *Organizing argumentative texts*: skill in organizing argumentative texts according to conventional academic structures, especially argumentative texts (e.g., thesis, argument, example, and conclusion).
- *Understanding metalinguistic vocabulary*: skill in understanding metalinguistic vocabulary or words that refer to—or qualify—thinking and reasoning processes (e.g., *hypothesis*, *generalization*, *contradictory*).
- *Understanding a writer's viewpoint*: skill in understanding markers that signal a writer's viewpoint, especially epistemic stance markers, those that signal a writer's degree of certainty in relation to a claim (e.g., *certainly*; *it is unlikely that*).

- *Recognizing academic language*: skill in recognizing more academic language when contrasted with more colloquial language in communicative contexts where academic language is expected (e.g., more colloquial versus more academic dictionary-like noun definitions).

Administration and Scoring

CALS-I is a group-administered 45-minute test that has two vertically equated forms: Form 1 (for grades 4–6) contains 49 items, and Form 2 (for grades 7 and 8) contains 46 items. A total of 29 items are common across both forms. Most items are scored dichotomously as correct (1) or incorrect (0), except for those of one task (i.e., organizing argumentative texts), which can receive partial credit. All partial-credit items are rescaled to be between 0 and 1. Scores include raw scores, percent correct scores, factor scores, and extended CALS-I (or ECALS) scores. The ECALS scores are the original factor scores (which are on a *z* score metric) rescaled to a new scale that has a mean of 500 and a standard deviation of 50 (Barr, Uccelli, & Phillips Galloway, 2019).

Sample

To date, three main studies have provided technical quality evidence for the CALS-I with samples of participants that included English-proficient and bilingual students designated as English learners across grades 4–8. A total of 7,152 students across grades 4–8 from 6 districts and 36 urban public schools in the Northeast and Middle Atlantic regions of the United States participated in the final norming study. The sample was balanced by gender (50.1 percent female), with a majority of students classified as English proficient and 12 percent identified as English learners according to official school records. Students were predominantly from low-income backgrounds (81 percent) as indexed by their eligibility for free or reduced-price lunch.

Validity

Evidence Based on Test Content

CALS-I tasks were expected to be moderately correlated (Uccelli et al., 2015a). Indeed, across grades CALS-I scores were moderately interrelated (range r = .23 to .64), with the lowest correlations observed for the academic register task. Findings revealed also that CALS-I captured developmental trends with upward trends in higher grades, yet considerable individual differences within grade.

Evidence Based on Internal Structure

This was evaluated using confirmatory factor analysis (CFA) and Rasch IRT. The authors assumed unidimentionality of CALS-I. In these analyses, the CALS-I task-specific scores (Unpacking Complex Words, Comprehending Complex Sentences, Connecting Ideas, Tracking Themes, Structuring Argumentative Texts, Identifying Academic definitions, and Producing Academic Definitions) were used. In two separate studies with students in grades 4–8 (Uccelli et al., 2015a, 2015b) the CFA results supported a single factor solution (CFI = .93 and .95, TLI = .92 and .94, RMSEA < .05 and = .06) and offered evidence for unidimensionality. In a third study (Barr, Phillips Galloway, & Uccelli, 2019), several competing models were tested to investigate the dimensionality of the construct assessed and the Rasch unidimensional measurement model was selected as the best model due to theoretical, empirical, and practical considerations.

Evidence Based on Relation to Other Variables

Validity evidence was provided in two studies that included Gates McGinitie (Uccelli, Phillips Galloway, Aguilar, & Allen, 2020) and GISA (Barr, Uccelli, & Phillips Galloway, 2019), respectively. Relations with Gates-MacGinitie Passage Comprehension as indexed by the zero-order correlations were .70 for Form 1 and .75 for Form 2. Relations between the CALS-I and the GISA reading comprehension scores were .69 for Form 1 and .71 for Form 2.

Reliability/Precision

For Form 1 reliability was .90 and for Form 2 it was .86 as indexed by coefficient alpha. Reliability of the CALS-I was also assessed by comparing the test information function and standard error of measurement for each of the two forms. Both forms had adequate test information function to standard error of measurement ratios (Form 1, –2.8 to 2.6; Form 2, –2.3 to 2.8), indicating that both forms offered adequate estimates of student ability across the expected range, with Form 1 scores having a higher reliability for low-performing students and Form 2 scores for students at higher ability levels (Barr et al., 2019).

Proposed Intended Use of Scores

This researcher-developed instrument is a reliable and valid measure to assess academic language skills in grades 4–8. The CCDD team used the CALS-I to model the relation between academic language proficiency and reading comprehension (LaRusso et al., 2016), to track concurrent longitudinal development in academic language and reading comprehension (Phillips Galloway & Uccelli, 2019), and to evaluate the impact of interventions (developed in the RfU initiative) on improving students' academic language proficiency and reading comprehension (Jones et al., 2019). The CALS-I is presently available for use as a research instrument upon request. Results of the CALS-I have been used effectively in teachers' professional development to raise awareness of the importance of paying attention to core academic language skills during instruction. Additional uses of the CALS-I to inform pedagogical practice are being investigated (Uccelli et al., 2020).

THE ASSESSMENT OF SOCIAL PERSPECTIVE-TAKING PERFORMANCE (ASPP) MEASURE

Conceptual Framework

Diazgranados, Selman, and Dionne (2015) identified the functional dimensions of social perspective taking (SPT) in the context of social-relational frameworks (Martin, Sokol, & Elfers, 2008; Mead, 1934). Social-relational frameworks differ from the cognitive-representational approaches (e.g., theory of mind, executive functions) that have largely dominated the literature on perspective taking. They followed a grounded-theory approach to develop a framework that resulted in the development of the Social Perspective Taking Acts Measure (SPTAM; Diazgranados et al., 2015) initially, and a revised version subsequently (ASPP; Kim et al., 2018). This work identified SPT as acts that serve different functions. Specifically, when students were challenged to resolve social situations presented in a scenario, they produced responses that (1) *acknowledged* the existence of different actors; (2) *articulated* the thoughts, feelings, and orientations to action of those actors; and (3) *positioned* these actors according to their characteristics, social roles, or circumstances in the scenario. Responses varied in their levels of *integration*, as participants demonstrated different abilities to acknowledge, articulate, and position the perspectives of multiple actors in the scenario. Kim et al. suggest that the ability to consider multiple perspectives is a critical skill for learning in 21st-century classrooms, facilitating the processing and integration of information from multiple sources.

Description

ASPP is a revised and extended version of SPTAM (Diazgranados et al., 2015), a scenario-based assessment of students' ability to perform SPT acts in response to written texts about specific social situations. ASPP is designed to assess children's ability to *acknowledge*, *articulate*, and *position* the perspectives of multiple stakeholders in a given social conflict and to provide solutions that consider and integrate their different positions. The measure puts students in the shoes of an advisor, who needs to make a recommendation to address social conflicts that occur at the interpersonal, group, and institutional levels. Specifically, students are presented with a subset (typically three) of four scenarios. In each scenario, an actor who is observing a social problem (i.e., a witness to teasing, mockery, or breaking school rules) does not know what to do and is asking different people for advice. Students are prompted to think about the recommendations this observer might receive from the following two types of advisors: (1) someone who was recently teased, whose privacy was recently violated, or is otherwise oriented in opposition to the perpetrator(s); and (2) someone who often socializes with the teasers or rule violators, or is otherwise in sympathy with the perpetrator(s). Then, students answer three questions: (1) What would (the prompted actor) recommend to the observer? (2) Why would (the prompted actor) make that recommendation? and (3) What might go wrong with this recommendation? This structure (four scenarios × two advisors) provides participants with the opportunity to produce open-ended responses to these sets of questions. Answers to all three questions provided by each advisor constitute one unit of analysis, which receives one score for each of the three subscales: *acknowledgment*, *articulation*, and *positioning*. Note that the revised version of ASPP excluded acknowledgment as a core component and focused primarily on *articulation* and *positioning* since both of these depend on actors being acknowledged (Kim et al., 2018). These subscales refer to the *function* of the SPT act, with acknowledgment serving the basic function of introducing a potential actor; articulating that actor's perspective is a more advanced act, while positioning that actor's perspective in light of her social role represents the pinnacle of SPT skill in ASPP.

Administration and Scoring

ASPP is a group-administered 30-minute test identified for students as "The Advice on Making Social Choices Measure." An experimenter reads the instructions and walks participants through the scenarios and questions, providing them with 4 minutes to answer each prompt. If participants complete a section before others, they are allowed to move on at their own pace. Coding follows detailed guidelines with examples that can be found in the coding manual (Diazgranados et al., 2011). The coding system was deemed stable when interrater reliability reached .90 (which reflected the proportion of units on which raters agreed out of the total number of units coded). This coding system results in three subscale scores (acknowledgment, articulation, and positioning). ASPP includes two forms (i.e., with the addition of new social scenarios and changes in elicited perspectives of the two advisor roles; Form A and Form B), and scoring excludes the *acknowledgment* dimension (even though it is initially coded).

Acknowledgment is the act of identifying the various actors involved. It can be determined by counting, only once per unit of analysis, the names and pronouns that refer to any particular actor that is included in the unit of analysis, irrespective of whether anything further is said about that actor.

Articulation is the act of describing the thoughts, feelings, or orientations to action of distinct actors involved. It can be determined by counting, only once per unit of analysis, the actors whose feelings, opinions, beliefs, preferences, and orientations to action are described in the scenarios.

Positioning is the act of identifying the roles, circumstances, or attributes that qualify the position distinct actors hold in a social scenario. It can be determined by counting, only once per unit of analysis, the actors whose roles, attributes, experiences, or circumstances are identified in the scenario as motivations for their beliefs, thoughts, actions, or potential actions.

Separate scores are assigned to each of the dimensions: for acknowledgment, 1 point per (potential) actor named; for articulation, 1 point per perspective described; and for positioning, 1 point per perspective explicitly positioned. In some past research, scores for each dimension have been scaled separately using item response theory to facilitate analysis.

Sample

Diazgranados et al. (2015) evaluated the initial SPTAM measure using a sample of participants from the larger study conducted in the context of the RfU Research Initiative. Participants were n = 459 students in grades 4–8 (50 percent boys), 25 percent in grade 4, 21 percent in grade 5, 18 percent in grade 6, 16 percent in grade 7, and 16 percent in grade 8. Subsequently, Kim et al. (2018) evaluated the ASPP measure in the same context, drawing on n = 1,299 students in grades 4–7. The current participants include 52 percent female students, 14 percent Black, 39 percent White, 4 percent Asian, 39 percent Latino, 3 percent mixed race/other, 79 percent eligible for free or reduced-price lunch, 12 percent English language learners, and 14 percent with special education classification.

Validity

Evidence Based on Internal Structure

Diazgranados et al. (2015) evaluated the SPTAM factor structure using a CFA, which provided support for a three-dimensional model in which SPT is a factor comprising acknowledgment, articulation, and positioning. The results showed that all parameter estimates were positive, statistically significant, and exhibited loadings in the range .62–.71. The three subscales exhibited positive, moderate, and statistically significant correlations with each other (r range .40 to .46).

Kim et al. (2018) tested the two-factor structure of the ASPP, articulation and positioning, using multigroup categorical confirmatory factor analysis (CCFA). The standardized factor loadings of the articulation items on the articulation factor ranged from .55 to .77, and factor loadings of the positioning items on the positioning factor ranged from .49 to .68. This model with two dimensions had a good fit, $\chi^2(135) = 174.72$ (Form A = 67.44, Form B = 107.28), p = .01, RMSEA = .02, 90% CI [.01, .03], CFI = .99, and TLI = .99. These multidimensional models fit the data significantly better than unidimensional models, $\Delta\chi^2(1) = 231.91$, $p < .001$.

Evidence Based on Relations to Other Variables

Diazgranados et al. (2015) examined hypothesized relations between SPT and several other constructs, while controlling for the rest. Specifically, they expected and confirmed that children in higher grades would perform better (for every additional grade level, students scored .37 points more on SPTAM, $p < .001$) and that girls would perform better (girls scored 1.37 points higher on SPTAM than boys, $p < .001$). It was also expected that SPTAM would have a negative and moderate association with the Aggressive Interpersonal Strategies (AINS) measure (Dalhberg, Toal, & Behrens, 1998). Indeed, for every additional unit in the AINS measure, students obtained .33 fewer points on SPTAM ($p < .10$). Finally, SPTAM was expected to have a moderate positive association with the Written Language Scale of the Oral and Written Language Scale (OWLS-II; Carrow-Woolfolk, 1995) because of its high language production demands. Indeed, for every additional point in the OWLS writing test, students scored 11.54 more points on SPTAM ($p < .001$). SPTAM was not related to measures of complex reasoning (LAS; Dawson, 2002; Fischer & Bidell, 2006), academic language (CALS-I; Uccelli et al., 2015a, 2015b), and reading (GMRT; MacGinitie & MacGinitie, 1988).

Kim et al. (2018) evaluated the relation of the two ASPP factors to several academic and engagement outcomes. The results showed that the overall ASPP model explained

52 percent to 54 percent of the variance in Reading Engagement (Wigfield et al., 2008), 33 percent to 34 percent of the variance in Classroom Engagement (Wellborn & Connell, 1987), 62 percent to 67 percent of the variance in ELA, and 62 percent to 64 percent of the variance in the Mathematics state test scores. The model had adequate goodness of fit, χ^2 (913) = 1138.10 (Form A = 501.37, Form B = 636.73), p < .001, RMSEA = .02, 90% CI [.02, .02], CFI = .97, and TLI = .96. When demographic variables were considered, the results also showed that students who scored higher on ASPP were more likely to be in higher grades and female. English language learners and students eligible for special education were likely to score lower on the ASPP.

Reliability/Precision

Diazgranados et al. (2015) reported Cronbach's alpha for each subscale of SPTAM: $\alpha_{\text{acknowledgment}} = .80$, $\alpha_{\text{articulation}} = .83$, and $\alpha_{\text{positioning}} = .70$. The latent factor of SPT exhibited excellent internal consistency (.90).

For each of the two forms of ASPP, Kim et al. (2018) reported both Cronbach alpha coefficients, $\alpha_A = .82$ and $\alpha_B = .78$ for articulation, and $\alpha_A = .67$ and $\alpha_B = .66$ for positioning, and omega reliabilities, $\Omega_A = .86$ and $\Omega_B = .83$ for the articulation scale and $\Omega_A = .74$ and $\Omega_B = .76$ for the positioning scale. In the CCFA context, omega reliabilities should be interpreted as more representative, and for both articulation and positioning they were acceptably high.

Proposed Intended Use of Scores

Diazgranados et al. (2015) and Kim et al. (2018) suggest that SPTAM/ASPP provides researchers with a tool to assess early adolescents' ability to produce SPT acts in an innovative way. This instrument can be particularly useful in the context of intervention programs whose theory of change includes SPT performance as a mechanism of change or outcome. For example, Hsin and Snow (2017) used a modification of the SPTAM coding scheme to examine the incidence of SPT acts in the argumentative essays of language-minority and English-only students in grades 4–6, and then associated the SPT found in students' writing with their ASPP scores. The results showed that language-minority students matched or surpassed the English-only students on perspective taking, and that there was a significant relationship between essay SPT and ASPP scores among language-minority students but not among English-only students.

ASK KNOWLEDGE ACQUISITION MEASURE

Conceptual Framework

Promoting Acceleration of Comprehension and Content through Text (PACT) is a multicomponent treatment aimed at improving content-area knowledge acquisition in social studies/history and also improving reading comprehension, consistent with the Common Core State Standards (CCSS). The CCSS requires teachers to emphasize students' understanding and learning from complex reading materials. Existing research shows that middle school teachers must make adjustments to current instructional practices to provide the reading opportunities and instruction necessary to ensure that students meet the CCSS expectations. To highlight the problem, students engaged in reading texts in only 38 percent of middle and secondary social studies classes and fewer than 20 percent of middle school social studies classes. Text reading consumed only 10.4 percent of social studies instructional time (Swanson, Wanzek, Vaughn, Roberts, & Fall, 2016). Vaughn et al. (2013) identified five components of the PACT intervention informed by the content learning model (Gersten et al., 2006; Vaughn et al., 2009) that focus on improving understanding while reading text, and provide opportunities for students to connect new, text-based learning to previous learning. They also infused the intervention with motivational aspects to further bolster its effectiveness among adolescent learners. These components are (1) a comprehension canopy that contains a motivational springboard and an overarching issue or question, (2) essential words or key vocabulary related to the unit, (3) knowledge acquisition (appropriate text-based instruction and reading), (4) team-based learning (TBL) comprehension checks, and (5) TBL knowledge application. In this context, it was deemed important to develop an appropriate knowledge test, the ASK Knowledge Acquisition measure.

Description

The ASK Knowledge Acquisition measure is one of the two subtests of the ASK assessment (Vaughn et al., 2015). The other subtest is a passage comprehension measure. The assessment is a researcher-developed measure. This knowledge subtest is a 42-item, four-option, untimed multiple-choice test that measures content knowledge in the three units that comprised the intervention (Colonial America, Road to Revolution, and Revolutionary War). The items comprising the test were collected (with permission) from released Texas state social studies tests (Texas Assessment of Knowledge and Skills), released Massachusetts state social studies tests (Massachusetts Comprehensive Assessment System), and released advanced placement tests in social studies from the College Board. Researcher-developed vocabulary items were also included in the item set. The item pool from these released items was further narrowed to align with the content of the Texas and Florida state content standards for the units covered in the PACT intervention. Following the first year of implementation of PACT the psychometric properties of the ASK content items were evaluated. Poor-performing items were removed from the assessment and a final version was created.

Administration and Scoring

The ASK Knowledge Acquisition measure is dichotomously scored (1 for correct, 0 for incorrect responses). The test was administered at pretest, posttest, 4 weeks following intervention, and again 12 weeks following intervention (Wanzek et al., 2015).

Sample

Participants were 1,487 students (male = 712), 39 percent qualified for free or reduced-price lunch, 4.8 percent were classified as limited English proficient, and 7.9 percent of students qualified for special education services. Students' average age was 13.16 in the treatment condition and 13.16 in the comparison condition (Vaughn et al., 2015).

Validity

Evidence Based on Internal Structure

Item response theory was used to analyze initial data from the validation process. IRT parameters for the 42 items reflect a sizable range of underlying knowledge acquisition (−2.12 to +2.67) and good item discrimination (0.05 to 2.13). Vaughn et al. (2013) used confirmatory factor analysis on pretest data to evaluate the degree to which the hypothesized models represented their observed data. Model fit was very good for the ASK Knowledge Acquisition test: $\chi^2 = 1{,}022.69$, $df = 989$, $p = .22$, CFI = .97, RMSEA = .009.

Reliability/Precision

Reliability information from IRT analyses was above .80 from −1.6 to +1.2 thetas. Alpha coefficients for the ASK knowledge acquisition measure was .89 (Vaughn et al., 2013). Vaughn et al. (2017) reported Cronbach's alpha of .93. Wanzek, Swanson, Vaughn, Roberts, and Kent (2015) reported alpha of .90.

Proposed Intended Use of Scores

This experimenter-developed measure is a valid and reliable indicator of middle and secondary school students' U.S. history knowledge and reading comprehension ability in the social studies domain. ASK has been used to evaluate the efficacy of the PACT intervention (developed as part of the RfU initiative) for improving students' social studies content knowledge and text comprehension among typical grade 8 students (Vaughn et al., 2013, 2015), grade 8 English learners (Vaughn et al., 2017; Wanzek et al., 2016), grade 8 students with disabilities (Swanson et al., 2016; Wanzek et al., 2016), and grade 11 students (Wanzek et al., 2015).

BRIDGE-IT MEASURE

Conceptual Framework

Barth, Barnes, Francis, Vaughn, and York (2015) aimed to develop a computerized inference measure drawing on the extant literature in discourse processes. Inferential processes support the integration of text-derived information and general world knowledge (Graesser, Singer, & Trabasso, 1994). These inferential processes involve maintaining local and global coherence during reading. Local and global coherence has been examined by studies that manipulate textual features, including distance in the text that separates two sentences or ideas that need to be integrated (Albrecht & O'Brien, 1991). Shorter distances between sentences may draw on local-coherence processes such as accessing and retrieving information from working memory (Albrecht & O'Brien, 1993). Larger distances are more likely to tap global coherence processes such as integration of information and bridging inferences. Coherence breaks become easier to detect with age (e.g., Ackerman, 1984) and seem to be more difficult to detect over longer distances (Pike, Barnes, & Barron, 2010). Moreover, skilled comprehenders detect coherence breaks more easily than less-skilled comprehenders, especially with larger distances between information units (e.g., Barnes, Faulkner, Wilkinson, & Dennis, 2004; Cain, Oakhill, & Lemmon, 2004).

Description

Bridge-IT was designed to measure the effects of textual distance (i.e., near versus far) on students' ability to generate inferences by judging the consistency or inconsistency of a continuation sentence with prior text. The Bridge-IT consists of 32 five-sentence narrative passages, presented on a computer monitor. Each story consists of five sentences and contains a key sentence important to making the consistency judgment. In the near condition, the key sentence is the final sentence in the story. In the far condition, the first sentence of the story serves as the key sentence. In both conditions, the additional sentences of the story are compatible with either the consistent or inconsistent continuation sentence. Correct judgments in the near condition require that readers evaluate information presented earlier in the text as well as the critical information presented in the final sentence. This information is likely still accessible by the reader. Correct judgments in the far condition require that readers evaluate information they just read as well as critical information in the first sentence of the story, which likely needs to be reactivated from episodic memory.

Administration and Scoring

A "Ready" prompt appears on the computer monitor for one second, followed by a five-sentence story. Instructions prompt students to press the spacebar after they finish the story. The spacebar removes the story and presents an asterisk in the center of the screen to signal the presentation of the test sentence. Participants receive instructions to read the test sentence and then to press a green button if they judge the sentence as a good continuation (i.e., consistent) or a red button if they judge that the sentence is not a good continuation (i.e., inconsistent). Judgments are to be made as quickly and accurately as possible. Students are provided two practice items to ensure familiarity with good and poor continuations and the task procedure.

Students receive a testlet that consists of eight items in each condition (i.e., near-consistent, far-consistent, near-inconsistent, and far-inconsistent). Items are counterbalanced across conditions. For each item, reading time is measured for the passage, and accuracy and response time are measured for continuation sentence judgments. Continuation sentences range from 3 to 12 words in length across items. Word length is consistent across consistent and inconsistent versions of the continuation sentence for each passage.

In terms of scoring, a total accuracy score and condition accuracy scores in all four conditions (i.e., near-consistent, far-consistent, near-inconsistent, and far-inconsistent) are calculated. Accuracy scores represent the proportion of items answered correctly after trimming for outliers.

Sample

Barth et al. (2015) evaluated the Bridge-IT measure using a sample of 1,203 students (*n* = 531 struggling comprehenders, 11 percent in grade 6, 14 percent in grade 7, 19 percent in grade 8, 15 percent in grade 9, 17 percent in grade 10, 15 percent in grade 11, and 10 percent in grade 12; and *n* = 675 adequate comprehenders, 9 percent in grade 6, 14 percent in grade 7, 13 percent in grade 8, 17 percent in grade 9, 17 percent in grade 10, 16 percent in grade 11, and 13 percent in grade 12). Adequate comprehenders were students attaining scale above 2,150 on the Texas Assessment of Knowledge and Skills (TAKS) Reading Test.

Validity

Evidence Based on Test Content

Barth et al. (2015) hypothesized grade-level changes in inferential processes across grades 6–12 for both adequate and struggling comprehenders, especially in the far condition. With regard to accuracy, results indicated that in the near condition, the effect of grade within distance was significant ($p < .001$). Students in grades 6 and 7 were less accurate than students in grades 10–12; students in grades 8 and 9 were less accurate than students in grade 10 ($p < .007$). In the far condition, students in grade 10 were more accurate than students in grades 6–9; and students in grade 12 were more accurate than students in grade 9 ($p < .007$). With regard to response time, students in grade 6 were slower at sentence continuation judgments than students in grades 8–12; students in grade 7 were slower than students in grades 9–12; students in grade 8 were slower than students in grades 10–12; and students in grade 9 were slower than students in grades 11 and 12 ($p < .007$).

Evidence Based on Relation to Other Variables

Barth et al. (2015) also hypothesized that inferential processes would account for unique variance in passage-level comprehension but not single-sentence comprehension, after controlling for working memory and a host of other reading-related variables. Predictive validity evidence was assessed with a series of hierarchical regression analyses. Bridge-IT-near explained 0.7 percent of the variance in the Test of Sentence Reading Efficiency and Comprehension (TOSREC) standard scores, and 3 percent of the variance in Gates MacGinitie reading test-Lexile Score over and above grade level, WJ-III letter word identification, TOWRE, WJ-III numbers reversed, KBIT-2 verbal knowledge, and reader group status.

Bridge-IT-far explained 0.3 percent of the variance in TOSREC (Wagner, Torgesen, & Rashotte, 2010) standard score and 2 percent of the variance in the Gates MacGinitie test-Lexile Score over and above grade level and other linguistic and cognitive measures (Barth et al., 2015).

Reliability/Precision

Average reliability coefficients (Kuder-Richardson 20) were .85 for near-consistent, .87 for near-inconsistent, .83 for far-consistent, and .87 for far-inconsistent continuations.

Proposed Intended Use of Scores

Barth et al. (2015) suggest that Bridge-IT adequately discriminates inference making, local and global, across grade levels 6–12 and comprehension skill (skilled versus less-skilled comprehenders). Thus, Bridge-IT can be used as a process measure of inference making.

READI LITERATURE EPISTEMIC COGNITION SCALE (LECS)

Conceptual Framework

Epistemic cognition has been broadly defined as the knowledge and beliefs people draw from in order to understand particular phenomena (Hofer & Pintrich, 1997; Yukhymenko et al., 2016). Epistemic cognition has been found to be related to students' problem solving, learning, and reasoning about topics in the natural and social sciences (Bråten, Strømsø, & Britt, 2009; Conley, Pintrich, Vekiri, & Harrison, 2004; Sinatra, Kienhues, & Hofer, 2014). Whether epistemic cognition also relates to students' understanding of literature remains less clear.

Literary reading (i.e., understanding literature, response to literature) can be conceptualized as a complex problem-solving task that requires readers to go beyond basic comprehension of the explicit content in literary texts. Readers must make deeper interpretative inferences about the literary text content, such as inferences about the moral and theme of the text (Goldman, McCarthy, & Burkett, 2014). Literary reading is also an ill-defined problem-solving task, as readers do not come to the same interpretations even after reading the same literary text. Thus, literary reading adopts some of the problem-solving characteristics found in the natural and social sciences. In this regard, epistemic cognition may also play an important role in literary reading. The Literature Epistemic Cognition Scale (LECS; Yukhymenko-Lescroart et al., 2016) was developed to measure epistemic cognition in literature.

Note that the READI team also developed epistemic cognition scales in history and science. The science and history scales emphasized two dimensions of epistemic cognition for multiple sources in history and science: the importance of corroborating across documents (history) and data sets and experiments (science), and the complexity and uncertainty of historical/scientific knowledge. These scales have not been validated to the same extent as LECS yet, and thus are not reviewed here.

Description

LECS measures three epistemic constructs for literature in adolescents (grades 6–12): relevance to life, multiple meanings, and multiple readings. *Relevance to life* measures the degree to which readers believe that reading literature can help them understand the human condition. Reading literature in order to understand the human condition is a fundamental assumption in the field of literary studies. *Multiple meanings* refer to readers' tendency to view literary texts as amenable to multiple interpretations. *Multiple readings* reflect readers' belief in the benefit of multiple readings in understanding a literary text. These constructs are thought to be central to adolescents' understanding of literature.

LECS consists of 16 items (the prevalidation version had 29 items) across the three different subscales: 5 items for multiple meanings, 6 items for multiple reading, and 5 items for relevance to life.

Administration and Scoring

LECS is individually administered and scored. Higher scores on the multiple meanings and relevance to life subscales reflect more sophisticated beliefs. Higher scores on the multiple readings subscale reflect less sophisticated beliefs. Prior to analysis, items corresponding to the multiple readings subscale were recoded so that higher scores reflected more sophisticated beliefs.

Sample

LECS was evaluated using a sample of 798 students. Of the total students, 455 were in middle school (*M* = 13.2 years, *SD* = 0.95 years) and 343 were in high school (*M* = 16.1 years, *SD* = 1.27 years). Of the total sample, 53.5 percent were female, and gender was evenly distributed across grades. Students were chosen from 47 classrooms across four middle schools and four high schools in a district located near a large urban Midwestern area. In regards to race, 33.4 percent of participants self-reported as Hispanic or Latino, 24.1 percent as White, 21.4 percent as Asian, 6.8 percent as Black, 1.6 percent as American Indian or Alaska Native, 0.9 percent as Native Hawaiian or Pacific Islander, and 11.7 percent as other.

Validity

Evidence Based on Internal Structure

The 798 surveys were divided into split-half samples after stratifying across gender and grade. The first sample ($n = 399$) was used to perform a confirmatory factor analysis to test the three-factor structure of the original 29-item scale. The model fit was evaluated with inspection of several indices: chi-square index, RMSEA, standardized root mean square residual (SRMSR), CFI, Tucker-Lewis index (TLI), and chi-square to degrees of freedom ratio. The results did not indicate good model fit: $\chi^2(374, n = 399) = 822.2$, $p < .001$ ($\chi^2/\mathrm{df} = 2.20$), CFI = .892, TLI = .882, RMSEA = .055, 90% CI [0.050, 0.060], SRMSR = .068.

The second sample ($n = 399$) was also used to perform a confirmatory factor analysis to test the model fit of an adjusted LECS with 16 items. Results indicated a good model fit: $\chi^2(101, n = 399) = 124.3$, $p = .058$ ($\chi^2/\mathrm{df} = 1.23$), CFI = .987, TLI = .985, RMSEA = .024, 90% CI [0, 0.037], SRMSR = .035. Model fit indices also did not change significantly by grade and gender for models measuring invariance of factor pattern, loadings, and variances, indicating that the model is valid for all genders in middle and high school.

Evidence Based on Relations to Other Variables

Criterion validity was evaluated with correlational analyses between the subscales of LECS, the Speed of Knowledge Acquisition subscale from the Wood and Kardash (2002) epistemology scale, and students' reading habits. The Speed of Knowledge Acquisition subscale measured students' beliefs about the speed of learning that ranged from learning is quick and straightforward to learning is complex and gradual. Speed of knowledge acquisition was predicted to correlate with the multiple meaning and multiple reading subscales. Students' reading habits were assessed by their response on two questions about their reading habit outside of school. Students who read more outside a school setting were thought to find reading more enjoyable, which would be associated with positive ratings on all three of the epistemic cognition constructs. Speed of knowledge acquisition was positively correlated with multiple reading, *r*(397) = .49, p < .001, and multiple meaning, *r*(397) = .50, p < .001. Liking of reading was positively correlated with multiple reading, *r*(397) = .36, p < .001, relevance to life, *r*(397) = .21, p < .001, and multiple meaning, *r*(397) = .17, p = .006.

Reliability/Precision

The omega reliability for each subscale was acceptable: .78 for multiple meaning, .85 for relevance to life, and .89 for multiple reading.

Proposed Intended Use of Scores

LECS is a reliable and valid measure to assess epistemic cognition for literature in adolescents (grades 6–12). As the first measure of epistemic cognition for literature, LECS can be used to explore the relationship between epistemic cognition and literary reading. The READI Literature intervention (Goldman, Greenleaf, et al., 2016) incorporated LECS as a pre- and posttest and the analysis showed that pretest scores on the multiple meaning, multiple reading, and relevance to life subscales predicted posttest scores. Importantly, the multiple meaning and relevance to life beliefs changed as a result of the intervention.

READI EVIDENCE-BASED ARGUMENT (EBA) MEASURE

Conceptual Framework

The ability to identify, evaluate, and synthesize information across multiple sources is a very important literacy skill in the 21st century. The READI team has focused on developing an instructional and curricular intervention that can help adolescents develop evidence-based argumentative skills from multiple sources across academic disciplines (Goldman, Greenleaf, et al., 2016; Goldman et al., 2019). However, what constitutes an evidence-based argument differs according to discipline. As a result, students must learn to engage in different reading practices that reflect different disciplinary epistemologies. This is challenging because students are rarely taught the discipline-specific skills and knowledge required to do so. The READI Evidence-Based Argument (EBA) assessments were designed to evaluate adolescents' ability to make evidence-based arguments from multiple sources in each of three disciplines, one in science, one in literature, and one in history. The science EBA was most extensively tested and was used as a proximal outcome measure in the randomized controlled trial efficacy study. The literature EBA was developed in the context of the design-based classroom research and was administered in these classrooms as well as in the 2-year longitudinal study. The history EBA was likewise developed and tested in the context of the design-based research classrooms. Because the technical qualities of the literature and history EBAs need further testing in larger samples of students, this review focuses primarily on the science EBA.

Science EBA

The READI science EBA was aligned with the learning goals of the science intervention. These include the following: Students need an understanding of what knowledge and knowledge building in science means. They must understand how claims and evidence are established or justified in science as well as the reasoning principles used to connect evidence to claims. Students must be able to understand different types of scientific texts and graphics that present scientific information. Finally, they must be able to understand the technical expressions and language conventions used in the texts and graphics. When students are equipped with this knowledge, they will be able to engage in evidence-based argumentation from multiple sources in science. More specifically, they will be able to use information from scientific texts to construct their own explanations of science phenomena, support their explanations, and critique explanations. These are the skills that are specifically assessed by the READI science EBA measure.

Description

The READI science EBA measure consists of five different tasks that tap evidence-based argumentative skills from multiple sources in science:

- *Reading*: closely reading and annotating scientific texts;
- *Essay*: reading and synthesizing task-relevant information within and across scientific texts;
- *Multiple-choice (9 items)*: reading and synthesizing task relevant information within and across scientific texts;
- *Graphical model comparison*: analyzing two graphic explanatory models related to the topic, selecting the better of the two, critiquing explanatory graphic model; and
- *"Peer" essay evaluation*: critiquing explanatory graphic models.

Administration and Scoring

Each student was provided with a text set on skin cancer or coral bleaching. A text set consisted of one text that provided background information about one of the two topics, two additional texts providing more information about the topic, and two graphics that portrayed an explanation of a phenomenon associated with the topic. The texts in a given set and tasks were chosen so that students would have to read and synthesize information across multiple sources.

The science EBA was administered over 2 school days. On the first day, students were administered a six-question survey that measured their prior knowledge of skin cancer or coral bleaching. Students were then told they were going to read about one of two topics. They were explicitly told that they would have to read and use information across multiple sources to explain a phenomenon related to the topic. Students received a text set and were asked to *read and annotate the texts*. Texts could be read in any order, although students were encouraged to read the background text first. On the second day, students were given the same text set and a booklet in which they would complete the other tasks.

The *essays* were scored according to the number of concepts and connections that students provided. Connections were indicated by the students' use of causal language. Essays were scored sentence by sentence.

The *graphical model comparison* task was given a score of 1 or 0 based on a rubric of acceptable answers. The justification that students provided for the model they selected had to include a variation of the following acceptable language conventions: steps, step-by-step, order, cause and effect, the way it is organized, process, chain reaction, and how they connect to each other.

The *peer evaluation essays* were scored based on the inclusion of six variables of interest that were present in the two peer essays: relevance, coherence, completeness, importance of sourcing, mentioning the graph, and mentioning a concept tied to the graph. A score of 1 was given for each variable if students wrote about the variable in at least one of their two evaluations. The acceptable language conventions for each variable were provided in a rubric.

Sample

Participants were n = 964 students in grade 9 (567 READI) from 95 classrooms (48 READI) in 24 schools (12 READI) and were present for all 4 days of the EBA assessment (two pre and two post).

Validity

The science EBA assessment consisted of several tasks that were designed to assess the skills outlined in the READI science intervention designed to help adolescents develop evidence-based argumentative skills from multiple sources (Goldman et al., 2019). Students need an understanding of what knowledge, knowledge building, reasoning, and knowledge expression (in text and graphics) in science means. When students are equipped with this knowledge, they will be able to engage in evidence-based argumentation from multiple sources in science. Thus, the assessment has a solid theoretical basis with respect to the dimensions of the construct being measured.

Reliability/Precision

Interrater reliability for the scoring of essays was determined by two coders who were trained to code essays on one topic and another coder who was trained to code essays on both topics. The two single-topic coders scored six subsets of essays that made up the total set, while the double topic coder randomly coded 20 percent of each subset. The kappa scores for the coral bleaching essays were .75, .89, .85, .86, .86, and .93 while the kappa scores for the skin cancer essays were .64, .92, .88, .89, .85, and .93.

Interrater reliability for the scoring of model evaluation responses was determined by one coder who scored all the responses in three different subsets and another coder who scored 20 percent of responses within each subset. The kappa scores were .90, .92, and .91.

Interrater reliability for the scoring of the peer evaluation task was determined by one coder who scored all the essays and another coder who scored a small set of evaluations over a period of time. Kappa scores were .86, .80, and .84.

Proposed Intended Use of Scores

The READI science EBA is a reliable assessment of evidence-based argumentation from multiple sources in science for students in grade 9. The science EBA measure was used to evaluate the efficacy of the READI intervention and showed sensitivity to intervention effects. Specifically, on average, the intervention group had 5.7 percent higher scores on the multiple-choice task than the control group (Goldman et al., 2019). In regard to essay task performance, students in the intervention and control groups did not differ significantly in the percentage of nodes and connections included in their essays, although the intervention group's scores were generally higher than those of the control group.