Measuring the Power of Learning.™

# Research Report

ETS RR–15-32

# SARA Reading Components Tests, RISE Forms: Technical Adequacy and Test Design, 2nd Edition


John Sabatini

Kelly Bruce

Jonathan Steinberg

Jonathan Weeks



December 2015

# ETS Research Report Series

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

# SARA Reading Components Tests, RISE Forms: Technical Adequacy and Test Design, 2nd Edition

John Sabatini, Kelly Bruce, Jonathan Steinberg, & Jonathan Weeks

Educational Testing Service, Princeton, NJ

This technical report describes the conceptual foundation and measurement properties of the Reading Inventory and Scholastic Evaluation (RISE). The RISE is a 6-subtest, Web-administered reading skills components battery. The theoretical and empirical foundations of each subtest in the battery are reviewed, as well as item designs. The results included in this report feature a vertical extension of the RISE to span Grades 5–10, psychometric analysis of parallel forms of each subtest, results of item response theory (IRT) scaling studies for each of the subtests across the entire grade span, and evaluation of differential item functioning (DIF) for gender and race/ethnicity.

**Keywords** Reading assessment; reading components; adolescent reading

doi:10.1002/ets2.12076

This second edition of the Reading Inventory and Scholastic Evaluation (RISE) technical report is intended to extend and supersede the report entitled *SARA Reading Components Tests, SARA Form: Technical Adequacy and Test Design*.[1] The conceptual framework and six-subtest battery structure of the RISE assessment remain the same; however, we have updated and revised some of the descriptions of the foundational research to reflect the growing literature in reading. The main changes described in this report include the following:

- vertical extension from the original Grades 6–8 to Grades 5 and 9–10
- development and psychometric analysis of parallel forms of each subtest
- construction of item response theory (IRT) scales for each of the subtests across the entire grade span
- evaluation of differential item functioning (DIF) for gender and race/ethnicity

The expansion of the RISE battery to multiple parallel forms, spanning a wider range of grades, with each subtest on a common scale is intended to enhance its utility and value for use in schools. The RISE assessment is a joint project of Educational Testing Service (ETS) and the Strategic Educational Research Partnership.[2]

The RISE assessment is a 45- to 60-minute, Web-based assessment of foundational reading skills. The RISE is part of a larger componential reading assessment system called the Study Aid and Reading Assessment (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. The enhanced RISE battery described in this report features multiple forms arranged in grade bands and is appropriate for students in Grades 5—10.

## Reading Comprehension and Foundational Reading Skills

Reading comprehension is a complex construct that involves the coordination of a number of theoretically integrated processes (Perfetti & Adlof, 2012). From recent reviews of the research literature (O'Reilly & Sabatini, 2013; O'Reilly & Sheehan, 2009; Sabatini & O'Reilly, 2013; Sabatini, O'Reilly, & Deane, 2013), the Common Core State Standards (Council of Chief State School Officers [CCSSO] & National Governors Association [NGA], 2010), the Partnership for 21st Century Skills (2004, 2008), and other seminal efforts in assessment innovation (Bennett, 2011; Bennett & Gitomer, 2009; Gordon Commission, 2013; Pellegrino, Chudowsky, & Glaser, 2001), Sabatini and O'Reilly (2013) extracted a number of common themes, articulated in six principles to guide development of a reading assessment framework. The first three principles are particularly relevant to the design of the RISE battery:

*Corresponding author:* J. Sabatini, E-mail: JSabatini@ets.org

- Principle 1: Print skills and linguistic comprehension are each necessary components of reading comprehension proficiency, though neither individually is sufficient to ensure proficiency (Adlof, Catts, & Little, 2006; Duke & Carlisle, 2011; Gough & Tunmer, 1986; Hoover & Gough, 1990; Vellutino, Tunmer, Jaccard, & Chen, 2007).
- Principle 2: Both breadth and depth of vocabulary knowledge are essential for understanding (Anderson & Freebody, 1981; Deane, 2012; Nagy & Scott, 2000; Ouellet, 2006).
- Principle 3: Readers construct mental models of text meaning at multiple levels, from literal to gist to complex situation models (Kintsch, 1988, 1998; McNamara & Kintsch, 1996).

What do we mean by foundational skills? Following Principle 1, foundational reading skills enable students to decode printed text, recognize words, and read fluently. Following Principle 2, it is foundational to have an extensive general vocabulary and knowledge of morphological variants or families of words. Finally, following Principle 3, students should be able to build a mental model of text meaning at various levels of sophistication. At a basic comprehension level, students need to be able to understand and encode the meaning of single sentences—which themselves might be quite complex. They should be able to read continuous text fluently and efficiently (at an appropriate rate for their grade levels) to get the gist of the meanings. They should also be able to build more complex mental representations of continuous text that may include identifying main ideas, locating details, or making cross-sentence inferences. These are the skills targeted in the RISE assessment.[3]

Ideally, all U.S.-educated students would have robust foundational reading skills in place by around the end of Grade 3 or the beginning of Grade 4. Grade 4 is an important milestone in U.S. schools because the nature and demands of reading change qualitatively. This grade typically marks what Chall (1967) referred to as the transition from learning to read to reading to learn. From Grade 4 on, U.S. students can expect to see an increasing quantity of content area reading and learning in academic subjects such as literature, science, and social studies.

For the typically progressing, on-grade-level, college-ready/bound learner, the reading load will increase every year through primary, middle, and secondary school. Students will be assigned more pages to read in more diverse topics and content areas. Consequently, they will need to learn a wider range of vocabulary. They will find that sentences have higher linguistic complexity; that is, the sentences are longer and include multiple phrases and clauses, and the syntactic structures will also be more complex. Not only are the texts getting longer and more complex but so also are the tasks and demands placed on students to understand and think critically about the content of those texts.

Remarkably, on-grade-level readers keep up with the accelerating reading demands of school curricula. Unfortunately, those with weak foundational skills are likely to fall further and further behind, unless they are provided appropriate help. The intention of the RISE battery is to identify relative weaknesses in foundational reading skills that may impede expected grade-level progress toward higher standards of reading proficiency.

## Conceptual Framework and Test Design

### *Conceptual Framework*

The sequence of six subtests in the RISE assessment forms a rough continuum of foundational skills beginning with the recognition or decoding of words, to understanding the meanings of words and sentences, to building meaning from passages. Reading and psycholinguistic research has documented the nature of processing and its importance to reading or language comprehension, only some of which is cited in this report (for more comprehensive reviews, see Carlisle & Rice, 2002; Snow, 2002).

Even though these components form a continuum theoretically, it would be a mistake to think of the reading process as strictly hierarchical in practice, nor do the foundational skills develop in isolation. Students do not need to master word recognition and decoding skills fully before they can construct some meaning from text. It takes only recognition of a noun and a verb to understand a meaningful proposition. In fact, individuals reading a passage will likely bring to bear all of their language, reading, and thinking skills, as well as relevant world knowledge, in understanding texts. This interactive reading process that combines bottom-up skills (such as word decoding) and top-down processes (such as making an inference based on one's knowledge of the context) is characteristic of reading at any developmental or ability level (Rumelhart & McClelland, 1982).

One might see it as an advantage that one can leverage skills at any level of processing toward understanding text. Unfortunately, there is a price to pay when some of those skills are weak or inefficient. A substantial body of research supports the basic tenets of Perfetti's (1985, 1993) verbal efficiency theory, which posits that weak lower level skills will diminish cognitive resources that can be applied to higher level comprehension and reasoning (e.g., Walczyk, Marsiglia, Bryan, & Naquin, 2001).

One of the key findings of this line of research was that whereas both skilled and unskilled readers could make inferences from sentence context in identifying (recognizing) a word that was already in their mental lexicons (i.e., a word that they already knew the meaning of when they heard it in speech), skilled readers could recognize the word rapidly, with ease, and with minimal attention, that is, with automaticity (LaBerge & Samuels, 1974), without any context. This efficiency of word recognition conserved processing resources that the skilled reader could deploy for higher level processes such as making inferences or reasoning about the text (Perfetti, 1993). Less skilled readers, on the other hand, relied more on the context, thus expending cognitive effort and attention that was no longer available for higher level reasoning and understanding of text (Perfetti, 1985, 1995). On the basis of the stability over time of this research-based tenet of reading development, we concluded that it would be worthwhile to measure the foundational skills separately from overall reading comprehension ability. We determined that this was especially important for students at risk of falling behind grade-level expectations to isolate whether specific barriers are impeding expected growth in reading comprehension.

Measuring discrete component skills, however, requires designing test items that minimize the individual's ability to borrow skills and knowledge from other strengths the individual may possess. This is somewhat contrary to the expectation of interactive processing in typical reading for understanding, but necessary if one wants to be confident about the level of an individual's foundational subskills. Thus, 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 (morphosyntactics) 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 finally, (f) comprehending the basic meaning of passages.

Note that some overlap of skills is inevitable, especially as each subsequent component skill requires some prerequisite knowledge and skills to support its execution. One cannot build meaning from a sentence if one does not understand or recognize most of the words in the sentence. We have taken some steps to minimize this overlap. For example, the words in sentence items were chosen to be of high frequency, and therefore it is more likely that even poor readers will know all of the words. When the design works, most of the processing will be directed toward the targeted cognitive skill of building sentence meaning, not toward recognizing the words. However, one can expect that the sentence task is also partially measuring the recognition of words, and that will impact overall performance. For this reason, as we describe later, it is best to interpret scores from most distal (i.e., decoding and word recognition) to more proximal (i.e., sentence or basic reading efficiency) to reading comprehension. In this way, one can take into account the impact weak, lower level skills may be exerting on subsequent subtest performances.

In the following sections, each of the RISE subtests is described in more detail, accompanied by a brief explanatory note citing some of the pertinent empirical research.

### *Subtest Content Framework*

Overall, the content of the RISE subtests is modeled on the kinds of academic materials (words, sentences, and passages) that students will encounter in their school curricula as determined by a review of formal and informal curricular materials targeted for this population. Other batteries designed for clinical use (e.g., Woodcock–Johnson III; Woodcock, McGrew, & Mather, 2001) utilize similar subtest constructs and item designs. However, most of these batteries were designed for one-to-one administration and interpretation by educational psychologists for high-stakes diagnostic purposes, such as identification of specific reading disabilities.

In contrast, the RISE assessment was designed to target a wider range of below-grade-level at-risk readers. Its computerized administration, relatively brief subtest and session duration (i.e., 45–60 minutes), and automated scoring and reporting support scalable applications at the classroom, school, or district level. It is not intended to replace clinical instruments, but rather to supplement them by providing evidence of instructionally malleable targets of readers' strengths and weaknesses. It can also be an indicator that a particular student should be referred for further clinical diagnostic testing. In line with this broader purpose, item content is drawn primarily from curricular content that one might find in U.S. schools. The theoretical foundations for each construct were reviewed; however, choices for specific items also took into consideration the likelihood and prevalence that students might encounter reading content similar to that in the RISE subtests. In the following sections, we provide brief reviews of the literature and form of each subtest.

*Subtest 1: Word Recognition and Decoding*

Most models of reading development recognize the centrality of rapid, automatic, visual word recognition skills and knowledge to reading ability (Abadzi, 2003; Adams, 1990; Ehri, 2005; Perfetti, 1985; Verhoeven & Perfetti, 2011). Two basic behavioral skills are indicative of proficiency in word recognition: (a) the accumulation of sight word knowledge of real words in the language and (b) (phonological) decoding, which enables the generation of plausible pronunciations of printed words and, conversely, plausible phonetic spellings of heard words. Decoding has been described as the fundamental word learning mechanism in alphabetic languages (Share, 1997) and therefore an essential component to measure directly.

Many non–reading specialists think of decoding as a simple skill mastered by most children in first or second grade, consisting primarily of mappings of individual letters to sounds. True, the mapping of sight–sound correspondences is a fundamental premise of decoding. However, in English, the underlying cognitive ability needs to be much more computationally complex because of the highly irregular sight–sound correspondence patterns of the English language and the influence on pronunciation of different stress patterns in multisyllabic words (Venezky, 1995, 1999). In fact, it is likely that decoding skills develop across the life-span, as the cognitive system adapts to reading the hundreds of thousands of words in texts such as those borrowed from languages other than English (e.g., *entrée*). In fact, the primary symptom of reading disability or dyslexia is weakness in the accuracy and automaticity of decoding words (Olson, 2007).

We reserve the term *decoding* for sounding out novel words that the reader has never or rarely seen before encountering them in a text. This may include dictionary terms or proper nouns such as product, person, or place names (e.g., Atorvastatin or Benin). In the RISE task, to ensure that the reader has never encountered a word before, we use made-up nonwords (e.g., *plign*).

Reading words that have never been encountered before is one kind of decoding; another is reading a word in one's spoken mental lexicon for the first time. In this instance, the processing goal is to sound out the word based on its spelling and match the pronunciation to a word one knows when one hears it. Because we learn words both from reading and from hearing them, it is beneficial to have skills in matching spellings to sounds of words. In the RISE assessment, we use pseudohomophones to test this ability. Pseudohomophones use nonconventional spellings that would sound like real words when pronounced out loud to oneself (e.g., *maik*–*make*).

We use the term *word recognition* (sight words) when readers have likely encountered the word in print numerous times and have built up a memory representation that allows them to identify the word automatically, without the conscious effort of sounding it out to themselves (Ehri, 2005; Logan, 1988; Rayner, 1997; Reynolds, 2000). Over time, with wide experience reading, many of the frequent words in the language become sight words. This allows word reading to become highly efficient, as the reader does not require context to help in identifying words and can therefore use the additional cognitive resources for comprehension (Tannenbaum, Torgesen, & Wagner, 2006). In the RISE assessment, we chose words likely to be encountered in late elementary and middle grade subject areas or literary texts.

Thus, 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, selected to cover a wide frequency range with a bias toward including the kinds of content area words that middle school students will encounter in their school curricula (examples of real words are *elect*, *mineral*, and *symbolic*)
2. nonwords, selected to cover a range of spelling and morphological patterns (examples of nonwords are *clort*, *plign*, and *phadintry*)
3. pseudohomophones, nonwords that nonetheless when pronounced sound exactly like real English words (examples of pseudohomophones are *whissle*, *brane*, and *rooler*)

Students are presented with one of the item types 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. Students are given practice and examples to understand how to complete the task successfully.

*Subtest 2: Vocabulary*

Knowing the meanings of words is essential to the reading process (Beck & McKeown, 1991; Carroll, 1993; Cunningham & Stanovich, 1997; Daneman, 1988; Hirsch, 2003; Perfetti, 1994), with correlations between vocabulary and reading comprehension assessments ranging from .6 to .7 (Anderson & Freebody, 1981). Individual differences in vocabulary knowledge emerge as early as preschool, and these differences tend to grow over time (Graves, Brunetti, & Slater, 1982; Graves & Slater, 1987; Hart & Risley, 1995). Vocabulary development is a critical part of learning to read well and appears to be a significant aspect of the gap between competent and struggling readers (National Center for Education Statistics, 2012).

In middle school, students begin to encounter general purpose academic words as well as more specialized content area words. Beck, McKeown, and Kucan (2002, 2008) distinguished these in their tier word system, a concept specifically referred to in the Common Core State Standards (CCSSO & NGA, 2010). Specifically, Tier 1 words are those used in everyday conversation, Tier 2 words are general academic words, and Tier 3 words are found in specific domains and less frequently in non-disciplinary-specific usage (Beck et al., 2002, 2008; Coleman & Pimentel, 2011). All three tiers are necessary to academic content learning, but the strategies for learning them may differ. The RISE Vocabulary subtest item set includes both Tier 2 and Tier 3 words. The response sets were designed such that the correct answer was either a synonym of the target or a meaning associate.

Another challenge of academic reading is the prevalence of polysemous words, that is, words with more than one meaning (Gernsbacher & Faust, 1991; Kang, 1993; McNamara & McDaniel, 2004). Papamihiel, Lake, and Rice (2005) specifically discussed the difficulties of content-specific polysemous words, where the more common meaning may lead to misconceptions when using that meaning to infer the more specific content meaning (e.g., *prime* meaning "high quality" versus referring to prime numbers in mathematics). RISE vocabulary items often probe these secondary meanings.

Learning word meanings is not entirely distinct from learning their spellings and pronunciations. Perfetti and Hart (2001) described word knowledge as a complex assemblage of representations that vary both in the information they contain and in the degree to which they have been fully specified (i.e., in terms of orthographic, phonemic, syntactic, and semantic quality), which they refer to as the *lexical quality hypothesis*. Thus, an expected relationship exists between the word recognition and decoding subtest and the vocabulary subtest.

As noted, in the vocabulary subtest, the response sets were designed such that the correct answer was either a synonym of the target or a meaning associate:[4]

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

Students are given practice and examples to understand how to complete the task successfully.

*Subtest 3: Morphology*

Morphemes are the basic building blocks of meaning in any language. Anglin (1993) and Nagy and Anderson (1984) estimated that more than half of English words are morphologically complex; that is, they are made up of more than one morpheme.

Morphological awareness is the extent to which students recognize the role that morphemes play in words—both in a semantic and a syntactic sense. A growing body of research suggests that morphological awareness, and morphology knowledge and skills more generally, is related to reading comprehension and the subskills that underlie reading (e.g., Carlisle, 2000; Carlisle & Stone, 2003; Fowler & Lieberman, 1995; Hogan, Bridges, Justice, & Cain, 2011; Kuo & Anderson, 2006; Tong, Deacon, Kirby, Cain, & Parrila, 2011). Nagy, Berninger, and Abbott (2006) concluded that the results of various studies are "consistent with a model of written word learning in which we draw on computations of the interrelationships among phonological, morphological, and orthographic word forms and their parts" (p. 136).

Poor morphological awareness, knowledge, or skills can be a source of reading comprehension difficulties among native speakers of English (Berninger, Abbott, Nagy, & Carlisle, 2010; Carlisle, 2000; Deacon & Kirby, 2004; Nagy et al., 2006; Stahl & Nagy, 2006) and even more so among English learners (Carlo et al., 2004; Kieffer & Lesaux, 2007, 2008). Morphological learning activities should address both roots and affixes and can occur both in isolation and in a reading context where meaning can be derived or guessed (Proctor et al., 2011). Evidence supports the teaching of morphological structure, especially with English-language learners (Carlo et al., 2004; Kieffer & Lesaux, 2007; Lesaux, Kieffer, Faller, & Kelley, 2010; Proctor et al., 2011).

The RISE Morphology subtest focuses on derivational morphology—those words that have prefixes and/or suffixes attached to a root. We use the cloze (fill in the blank) item type for this subtest. Thus one might also consider these items morphosyntactic, in that some items can be answered correctly by understanding how a suffix alters the part of speech of a word and how that would fit a sentence context grammatically. However, understanding how the affixes affect the meaning of the word in the sentence context is always sufficient for answering the item correctly.

The sentences we designed featured straightforward syntactic structures and relatively easy ancillary vocabulary so that the students would concentrate on the derived words. See the following examples.

The target derived form is high frequency:
For many people, birthdays can be times of great __________.
(happiness, unhappy, happily)
The target derived form is medium frequency:
She is good at many sports, but her __________ is basketball.
(specialty, specialize, specialist)
The target derived form is low frequency:
That man treats everyone with respect and __________.
(civility, civilization, civilian)

Students are given practice and examples to understand how to complete the task successfully.

*Subtest 4: Sentence Processing*

A variety of research studies show that the sentence is a natural breakpoint in the reading of continuous text (e.g., Kintsch, 1998). A skilled reader will generally pause at the end of each sentence to encode the propositions of the sentence, make anaphoric inferences, relate meaning units to background knowledge and to previous memory of the passage as it unfolds, and decide which meaning elements to hold in working memory. Thus, every sentence requires some syntactic and semantic processing. In middle school, students encounter texts that contain sentences of a variety of lengths and syntactic structures.

Carlisle and Rice (2002) noted several ways in which compound and complex sentences may pose difficulty for struggling readers. Perhaps most obviously, complex sentences are often longer, and this places increased demands on working memory. Also, complex sentences often have multiple embedded phrases and clauses that increase the distance between subjects and predicates, a feature known to increase processing demands (e.g., Mann, Shankweiler, & Smith, 1984). Key to understanding complex sentences is efficient processing of connectors. Relationships that are signaled may be temporal (e.g., before), causal (e.g., because), adversative (e.g., although), or conditional (e.g., if). Empirical studies have been conducted examining the difficulties learners often have in adequately processing these kinds of relations (e.g., McClure & Steffensen, 1985).

In the sentence processing subtest, we chose to focus 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. Some examples follow.

The dog that chased the cat around the yard spent all night __________.
(barking, meowing, writing)
Shouting in a voice louder than her friend Cindy's, Tonya asked Joe to unlock the door, but __________ didn't respond.
(he, she, they)

Students are given practice and examples to understand how to complete the task successfully.

*Subtest 5: Efficiency of Basic Reading Comprehension—MAZE*

Skilled reading is rapid, efficient, and fluent (silent or aloud). In recent research, a silent reading assessment task design—known as the maze selection technique—has gained empirical support as an indicator of basic reading efficiency and comprehension (Fuchs & Fuchs, 1992; Shin, Deno, & Espin, 2000; Wayman, Wallace, Wiley, Ticha, & Espin, 2007). The design uses a forced-choice cloze paradigm—that is, in each sentence within a passage, one of the words has been replaced with three choices, only one of which makes sense in the sentence.

Fuchs and Fuchs (1992) found correlations of .83 between scores on maze and a read-aloud task and .77 between scores on maze and the reading comprehension subtest of the Stanford Achievement Test (Gardner, Rudman, Karlsen, & Merwin, 1982). In their extensive review of curriculum-based measures, Wayman et al. (2007) concluded that the evidence supported the use of the maze-style task structure with older middle school students, whereas word identification and reading aloud were more appropriate for younger readers.

While the empirical support for the maze selection task has been strong, less has been written about the underlying construct the task represents. This partially stems from its utilitarian origins as a quick, efficient progress monitoring indicator of whether students in special education programs were responding to instruction or needed further support. Our analysis of the task demands has led us to label the task as efficiency of basic reading comprehension and position it as an aspect of building models of text at various levels of sophistication. In the case of the maze task, this level of sophistication is shallow. 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. However, because the task is timed, the simultaneous demand that students read quickly also captures an indicator of silent reading fluency or efficiency. In fact, Espin, Deno, Maruyama, and Cohen (1989) reported correlations to oral reading fluency of .77–.86 for third to fifth graders.

The RISE Efficiency of Basic Reading Comprehension subtest comprises expository texts. Students have 3 minutes to complete each passage. The following is an excerpt from a passage:

> 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.

Students are given practice and examples to understand how to complete the task successfully.

*Subtest 6: Reading Comprehension*

Kintsch's (1998) construction integration model focuses on three levels of understanding: the surface level (a verbatim understanding of the words and phrases), the textbase (the "gist" understanding of what is being read), and the situation model (McNamara & Kintsch, 1996), which is the deepest level of understanding. In the reading literacy assessment framework developed by Sabatini et al. (2013), five dimensions of reading are described: print, verbal, discourse, conceptual, and social. The reading comprehension subtest targets the discourse level. That is, an attempt was made to limit the number of deeper conceptual or social reasoning questions on the subtest. That does not mean that all the questions are easy. In fact, the items show a range of difficulties. However, the reading comprehension subtest does not attempt to cover the broader range of task demands that are addressed in scenario-based assessments (O'Reilly & Sabatini, 2013).

In the reading comprehension subtest, the task focuses on the first two levels of understanding. An excerpt from the "Permanent Housing" passage and two related questions follow:

> To build their houses, the people of this Age often stacked mud bricks together to make rectangular or round buildings. At first, these houses had one big room. Gradually, they changed to include several rooms that could be used for different purposes. People dug pits for cooking inside the houses, and they may have filled the pits with water and dropped in hot stones to boil it. You can think of these as the first kitchens.
>
> The emergence of permanent shelters had a dramatic effect on humans. They gave people more protection from the weather and from wild animals. Along with the crops that provided more food than hunting and gathering, permanent housing allowed people to live together in larger communities.
>
> Example Question 1 (Locate/Paraphrase): What did people use to heat water in Neolithic houses? (hot rocks, burning sticks, the sun, mud)
> Example Question 2 (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)

# Methods

## Equivalent Form Design

The RISE forms are suitable for students in Grades 5 to 10, with the number of forms available for each grade or grade band are shown in Table 1.

To deliver forms that could be equated within and across grades, the tests were constructed from a large pool of items that had already been piloted with students from the appropriate grades. The items on each form, for each subtest, included a subset of items that were administered on more than one form. The items administered on a single form are referred to as unique items, and the items administered on two or more forms are referred to as common items. Figure 1 illustrates the item linking design.

**Table 1** Number of Forms per Grade/Grade Band

| Grade/grade band | No. of forms |
|---|---|
| 5 | 4 |
| 6–9 | 6 |
| 10 | 3 |

## Administration Procedures and Participant Characteristics

### *Test Administration Procedures*

Data were collected in a large, urban school district in the mid-Atlantic region of the United States. The initial test administration occurred in Winter–Spring 2011 and continued for the next fall and spring through Fall 2012 for a total of four waves of data collection for this report. See Table 2 for the grades tested during each wave.

No exclusions (e.g., for language proficiency or special education status) were mandated. Tests were administered in school computer labs and were proctored by school staff members who had been trained in standard test administration procedures.

### *Participant Characteristics*

Participant characteristics are reported in Tables 3–6, which show mean values across the four waves of administration.

# Psychometric Analyses

## Instrument Key

The following abbreviations are used in tables for each of the six subtests: WRDC, word recognition and decoding; VOC, vocabulary; MOR, morphology; SEN, sentence processing; MZ (maze) efficiency of basic reading comprehension; and RC, reading comprehension.

## Item Response Theory Analysis and Scaling

To compare the results across test forms, it is important that they be reported on a common scale. IRT (Lord & Novick, 1968) is commonly used for this purpose. In contrast to classical methods, which essentially aggregate scored responses, IRT is a probabilistic approach that relies on the pattern of item responses and item characteristics to obtain estimates of examinee ability. Let the variable $X_{ij}$ represent the response of examinee $i$ to item $j$, where $X_{ij} = 1$ for a correct item response and $X_{ij} = 0$ for an incorrect response. The item response curve for the two-parameter logistic model (2PL; Birnbaum, 1968) takes the following form:

$$P\left(X_{ij} = 1 | \theta_i, a_j, b_j\right) = \frac{\exp\left[1.7a_j\left(\theta_i - b_j\right)\right]}{1 + \exp\left[1.7a_j\left(\theta_\mathrm{i} - b_j\right)\right]}, \tag{1}$$

| WRDC | | Grade 5 | | | | Grades 6 - 9 | | | | | | Grade 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 |
| Grade 5 | F1 | 50 | | | | | | | | | | | | |
| | F2 | 24 | 50 | | | | | | | | | | | |
| | F3 | 24 | 22 | 50 | | | | | | | | | | |
| | F4 | 24 | 22 | 26 | 50 | | | | | | | | | |
| Grades 6 - 9 | F5 | 27 | 15 | 16 | 16 | 50 | | | | | | | | |
| | F6 | 14 | 9 | 10 | 10 | 25 | 50 | | | | | | | |
| | F7 | 13 | 8 | 6 | 6 | 25 | | 50 | | | | | | |
| | F8 | | 2 | | | | 25 | 25 | 50 | | | | | |
| | F9 | | 1 | | | | | 20 | 20 | 50 | | | | |
| | F10 | 10 | 7 | 5 | 6 | 21 | | 21 | | | 49 | | | |
| Grade 10 | F11 | 6 | 9 | 5 | 5 | 14 | 8 | 7 | 1 | | 6 | 52 | | |
| | F12 | 5 | 4 | 4 | 4 | 11 | 9 | 6 | 4 | 1 | 4 | 22 | 52 | |
| | F13 | 5 | 4 | 4 | 4 | 11 | 9 | 6 | 4 | 1 | 4 | 22 | 25 | 52 |

| VOC | | Grade 5 | | | | Grades 6 - 9 | | | | | | Grade 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 |
| Grade 5 | F1 | 38 | | | | | | | | | | | | |
| | F2 | 28 | 38 | | | | | | | | | | | |
| | F3 | 19 | 19 | 38 | | | | | | | | | | |
| | F4 | 18 | 18 | 18 | 38 | | | | | | | | | |
| Grades 6 - 9 | F5 | 15 | 13 | 13 | 13 | 38 | | | | | | | | |
| | F6 | 8 | 6 | 6 | 6 | 19 | 38 | | | | | | | |
| | F7 | 7 | 7 | 7 | 7 | 19 | | 38 | | | | | | |
| | F8 | | | | | | 19 | 19 | 38 | | | | | |
| | F9 | | | | | | | 19 | 19 | 38 | | | | |
| | F10 | 7 | 7 | 7 | 7 | 19 | | 19 | | | 38 | | | |
| Grade 10 | F11 | 2 | 1 | 1 | 1 | 14 | 7 | 7 | | | 7 | 35 | | |
| | F12 | 2 | 1 | 1 | 1 | 11 | 9 | 6 | 4 | 1 | 5 | 18 | 35 | |
| | F13 | 2 | 1 | 1 | 1 | 11 | 9 | 6 | 4 | 1 | 5 | 18 | 22 | 35 |

| MOR | | Grade 5 | | | | Grades 6 - 9 | | | | | | Grade 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 |
| Grade 5 | F1 | 32 | | | | | | | | | | | | |
| | F2 | 26 | 32 | | | | | | | | | | | |
| | F3 | 15 | 14 | 32 | | | | | | | | | | |
| | F4 | 15 | 14 | 15 | 32 | | | | | | | | | |
| Grades 6 - 9 | F5 | 17 | 15 | 15 | 15 | 32 | | | | | | | | |
| | F6 | 7 | 6 | 6 | 6 | 16 | 32 | | | | | | | |
| | F7 | 10 | 9 | 10 | 9 | 16 | | 32 | | | | | | |
| | F8 | | | 1 | | | 16 | 16 | 32 | | | | | |
| | F9 | | | 1 | | | | 16 | 16 | 32 | | | | |
| | F10 | 10 | 9 | 9 | 9 | 16 | | 16 | | | 32 | | | |
| Grade 10 | F11 | 8 | 6 | 6 | 6 | 16 | 8 | 8 | | | 8 | 30 | | |
| | F12 | 3 | 2 | 3 | 3 | 11 | 5 | 6 | | | 6 | 16 | 30 | |
| | F13 | 3 | 2 | 3 | 3 | 11 | 5 | 6 | | | 6 | 16 | 16 | 30 |

| SEN | | Grade 5 | | | | Grades 6 - 9 | | | | | | Grade 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 |
| Grade 5 | F1 | 26 | | | | | | | | | | | | |
| | F2 | 24 | 26 | | | | | | | | | | | |
| | F3 | 14 | 13 | 26 | | | | | | | | | | |
| | F4 | 14 | 13 | 14 | 26 | | | | | | | | | |
| Grades 6 - 9 | F5 | 23 | 21 | 13 | 13 | 26 | | | | | | | | |
| | F6 | 12 | 12 | 9 | 9 | 13 | 26 | | | | | | | |
| | F7 | 11 | 9 | 4 | 4 | 13 | | 26 | | | | | | |
| | F8 | | | | | | 13 | 13 | 26 | | | | | |
| | F9 | | | | | | | 13 | 13 | 26 | | | | |
| | F10 | 11 | 9 | 4 | 4 | 13 | | 13 | | | 26 | | | |
| Grade 10 | F11 | 8 | 6 | 7 | 7 | 10 | 6 | 4 | | | 4 | 25 | | |
| | F12 | 9 | 7 | 8 | 8 | 11 | 7 | 4 | | | 4 | 12 | 25 | |
| | F13 | 9 | 7 | 8 | 8 | 11 | 7 | 4 | | | 4 | 12 | 14 | 25 |

| MZ | | Grade 5 | | | | Grades 6 - 9 | | | | | | Grade 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 |
| Grade 5 | F1 | 39 | | | | | | | | | | | | |
| | F2 | 25 | 40 | | | | | | | | | | | |
| | F3 | 25 | 25 | 39 | | | | | | | | | | |
| | F4 | 10 | 25 | 10 | 39 | | | | | | | | | |
| Grades 6 - 9 | F5 | 10 | 10 | 10 | 10 | 36 | | | | | | | | |
| | F6 | 10 | 10 | 10 | 10 | 23 | 39 | | | | | | | |
| | F7 | | | | | 13 | | 46 | | | | | | |
| | F8 | | | | | | 16 | 33 | 49 | | | | | |
| | F9 | 10 | 10 | 10 | 10 | 23 | 10 | 13 | | 39 | | | | |
| | F10 | | | | | 13 | 16 | 13 | 16 | 13 | 45 | | | |
| Grade 10 | F11 | | | | | 13 | 13 | | | | | 43 | | |
| | F12 | | | | | 13 | 13 | | | | 16 | 27 | 43 | |
| | F13 | | | | | 13 | 13 | | | | | 29 | 13 | 42 |

| RC | | Grade 5 | | | | Grades 6 - 9 | | | | | | Grade 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 |
| Grade 5 | F1 | 18 | | | | | | | | | | | | |
| | F2 | 13 | 19 | | | | | | | | | | | |
| | F3 | 13 | 13 | 18 | | | | | | | | | | |
| | F4 | 8 | 14 | 8 | 19 | | | | | | | | | |
| Grades 6 - 9 | F5 | 8 | 8 | 8 | 8 | 22 | | | | | | | | |
| | F6 | 8 | 8 | 8 | 8 | 14 | 20 | | | | | | | |
| | F7 | | | | | 8 | | 20 | | | | | | |
| | F8 | | | | | | 6 | 12 | 18 | | | | | |
| | F9 | 8 | 8 | 8 | 8 | 16 | 8 | 8 | | 22 | | | | |
| | F10 | | | | | 8 | 6 | 8 | 6 | 8 | 20 | | | |
| Grade 10 | F11 | | | | | 6 | 6 | | | | | 19 | | |
| | F12 | | | | | 6 | 6 | | | | 6 | 13 | 19 | |
| | F13 | | | | | 6 | 6 | | | | | 12 | 6 | 17 |

**Figure 1** Linking design. The total number of items for each form is located on the diagonal. The cells with nonzero values on the off-diagonal are the number of common items between a given pair of forms. The cells with the light shading correspond to common items across forms within a given grade range. *Note.* WRDC = word recognition and decoding; RC = reading comprehension; MOR = morphology; SEN = sentence processing; MZ = (MAZE) efficiency of basic reading; VOC = vocabulary.

where $\theta_i$ is the individual's ability on a single construct, $a_j$ is the item discrimination (slope), and $b_j$ is the item difficulty.

The forms for each of the six subtests (WRDC, VOC, MOR, SEN, MZ, and RC) were scaled using the 2PL. The end 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 (Bock & Zimowski, 1997) where the item parameters for the common items were constrained to be equal across groups. Each examinee group (Winter/Spring 2011, Fall 2011, Spring 2012, Fall 2012) by form was treated as a separate group for the purpose of the item parameter estimation. The Grade 7 test from the Fall 2012 administration was treated as the reference point. After item parameters were estimated, examinee abilities were estimated using the expected a posteriori method. The item and ability parameters were estimated using the software program PARSCALE. As a final step, the scores for all six scales were rescaled to have a mean of 250 and a standard deviation of 15. The scale is also constrained to have a minimum value of 190 and a maximum value of 310. The grade-by-scale sample sizes, means, standard deviations, and standard errors of measurement based on median values across waves and forms are reported in Table 7.

**Table 2** Grades Tested During Each Wave

| Wave | Grades |
|---|---|
| 1. WINSPR 2011 | 6–9 |
| 2. FALL 2011 | 6–9 |
| 3. SPRING 2012 | 6–9 |
| 4. FALL 2012 | 5–10 |

**Table 3** Participant Characteristics: By Grade and Gender

| Grade | Total students | Female (%) | Male (%) | Not reported (%) |
|---|---|---|---|---|
| 5 | 2,947 | 49.2 | 50.8 | 0.0 |
| 6 | 3,540 | 48.8 | 51.2 | 0.1 |
| 7 | 3,477 | 48.4 | 51.6 | 0.1 |
| 8 | 3,114 | 49.0 | 51.0 | 0.1 |
| 9 | 2,885 | 54.3 | 45.6 | 0.1 |
| 10 | 1,420 | 58.5 | 41.4 | 0.1 |

**Table 4** Participant Characteristics: By Grade, Ethnicity, and Race

| | | Ethnicity (%) | Race (%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| Grade | Total students | Hispanic/ Latino | American Indian/ Native Alaskan | Asian | Black/ African American | Native Hawaiian/ Pacific Islander | White | Other/ not reported |
| 5 | 2,947 | 5.2 | 0.1 | 0.6 | 87.2 | 0.1 | 11.7 | 0.3 |
| 6 | 3,540 | 4.3 | 0.3 | 1.0 | 86.0 | 0.2 | 12.3 | 0.2 |
| 7 | 3,477 | 4.1 | 0.2 | 0.8 | 86.4 | 0.2 | 12.2 | 0.1 |
| 8 | 3,114 | 3.6 | 0.3 | 1.1 | 86.5 | 0.3 | 11.6 | 0.1 |
| 9 | 2,885 | 2.3 | 0.4 | 1.1 | 89.3 | 0.2 | 8.8 | 0.1 |
| 10 | 1,420 | 2.0 | 0.3 | 1.7 | 90.6 | 0.0 | 7.3 | 0.1 |

A wide range of sample sizes may be a function of schools coming in and out of the test administrations. The corresponding means, standard deviations, and standard errors of measurement reflect developmental differences in ability with respect to performance across subtests. For example, scores are lowest with generally the smallest amount of variation in Grade 5, whereas scores are highest with generally the largest amount of variation in Grades 9 and 10.

To provide a sense of the variability in scale scores within grade levels, Table 8 shows scale scores at the 10th, 25th, 50th, 75th, and 90th percentiles.

## Reliability

Cronbach's alpha coefficients (Cronbach, 1951) were computed for each subtest within each administration, form, and grade. Table 9 signifies the range of reliabilities represented as median values within a grade across forms and administrations. Particularly in Grades 6–9, the pattern is such that reliability increases as grade level increases, which is indicative of more consistent responses. All values are generally within acceptable ranges given the number of items in each subtest.

## Validity

As noted in the theoretical descriptions, it would be predicted that the various subtests would be moderately to strongly correlated with each other. Each subtest construct represents a somewhat distinct component or subskill. Conversely, each would be expected to have some dependency on other components, and one would expect that individuals would exhibit some comparability in performance across the subtests, as all are measuring aspects of reading ability. Correlation coefficients were computed between subtest scores within grade across forms and administrations, and where appropriate, ranges are reported (see Tables 10–15). The values in the lower triangle in these tables are the observed correlations; the values in the upper triangle are the correlations after correcting for attenuation.

**Table 5** Participant Characteristics: By Grade and Limited English Proficiency Status for Academic Year 2011–2012

| Grade | Total students | Receiving English-language learner services (%) | Not receiving English-language learner services (%) | Exited services within past 2 years (not currently receiving services) (%) | Not reported (%) |
|---|---|---|---|---|---|
| 5 | 2,947 | 1.4 | 96.1 | 2.4 | 0.1 |
| 6 | 3,540 | 1.9 | 94.8 | 1.3 | 2.1 |
| 7 | 3,477 | 1.5 | 95.7 | 1.0 | 1.8 |
| 8 | 3,114 | 1.4 | 95.5 | 0.9 | 2.3 |
| 9 | 2,885 | 0.7 | 95.8 | 0.7 | 2.8 |
| 10 | 1,420 | 0.6 | 95.8 | 0.6 | 3.0 |

**Table 6** Participant Characteristics: By Grade and Special Education Status for Academic Year 2011–2012

| Grade | Total students | Receiving special education services (%) | Not receiving special education services (%) | Code 504 (%) | Exited special education and placed in Code 504 (%) | Exited services within past 2 years[a] (%) | Not reported (%) |
|---|---|---|---|---|---|---|---|
| 5 | 2,947 | 15.3 | 78.4 | 2.4 | 0.2 | 0.8 | 2.8 |
| 6 | 3,540 | 17.3 | 76.4 | 2.9 | 0.3 | 1.1 | 2.1 |
| 7 | 3,477 | 17.0 | 77.3 | 2.7 | 0.4 | 0.9 | 1.8 |
| 8 | 3,114 | 16.0 | 78.0 | 2.4 | 0.4 | 1.2 | 2.3 |
| 9 | 2,885 | 13.8 | 80.0 | 1.9 | 0.2 | 1.2 | 2.0 |
| 10 | 1,420 | 13.2 | 81.0 | 1.5 | 0.1 | 1.2 | 3.0 |

[a]Not currently receiving services.

## Subscore Utility

Since it has been established that each subtest has adequate reliability and apparent discrimination from the other subtests (i.e., disattenuated intercorrelations remain below .81), it is worthwhile to examine the overall utility of each subtest within the component battery. Haberman (2005) and Sinharay, Haberman, and Puhan (2007) are the seminal works in demonstrating general subscore utility in place of just reporting a total score. With respect to reading components assessment, this has been done previously by McCormick, Sabatini, Bruce, Sinharay, and O'Reilly (2012) in a different school district than the one discussed in this report. As the components were the same, the analyses were replicated using Fall 2012 data and done for each form within each grade, representing 19 total comparisons. The input information included Cronbach's alpha reliability values for each subtest, 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 was computed based on all item-level data across subtests merged together by unique student identifier.

Across all 19 comparisons, 15 of these (79%) met the criteria for subscore utility. Within the four comparisons that did not meet the criteria, there is a relationship to grade level, as all of these involved Grade 5 or Grade 6. In three instances, the reading comprehension subtest did not meet the criteria, which might be expected given it has the fewest number of items of the six subtests (and often the lowest subtest reliability). In addition, in two instances in Grade 6, the vocabulary subtest did not meet the criteria, and in one instance in Grade 5, the criteria was not met for morphology.

## Differential Item Functioning

In validating the use of any assessment, it is important to examine effects of potential DIF. To accomplish this goal, item-level data are needed along with demographic information. In this section, we discuss results using Fall 2012 demographic information provided by the school district under study, consisting of gender (male, female) and race/ethnicity (American Indian/Alaskan Native, Asian, African American, White, Hispanic). DIF analyses consist of comparing individual item performance between two groups matched based on a specified criterion, which in this case is the total

**Table 7** Statistics for Each RISE Subtest by Grade

| Grade | *N* | Mean | *SD* | *SEM* |
|---|---|---|---|---|
| Subtest 1: Word recognition and decoding | | | | |
| 5 | 1,460 | 253.3 | 9.9 | 3.0 |
| 6 | 894 | 256.1 | 11.5 | 3.5 |
| 7 | 854 | 259.4 | 13.0 | 3.7 |
| 8 | 776 | 262.2 | 13.9 | 3.9 |
| 9 | 742 | 265.0 | 14.9 | 4.2 |
| 10 | 1,285 | 264.6 | 13.9 | 4.4 |
| Subtest 2: Vocabulary | | | | |
| 5 | 1,459 | 252.7 | 8.8 | 2.8 |
| 6 | 890 | 253.6 | 9.3 | 3.8 |
| 7 | 852 | 256.5 | 11.4 | 4.3 |
| 8 | 772 | 260.5 | 14.2 | 4.9 |
| 9 | 739 | 265.0 | 15.9 | 5.3 |
| 10 | 1,284 | 267.9 | 17.3 | 6.5 |
| Subtest 3: Morphology | | | | |
| 5 | 1,457 | 252.1 | 9.9 | 3.5 |
| 6 | 885 | 257.0 | 13.7 | 4.6 |
| 7 | 849 | 260.9 | 15.6 | 4.9 |
| 8 | 770 | 265.5 | 17.3 | 5.1 |
| 9 | 731 | 270.0 | 18.6 | 5.3 |
| 10 | 1,279 | 274.0 | 18.9 | 5.7 |
| Subtest 4: Sentence processing | | | | |
| 5 | 1,456 | 255.0 | 11.7 | 4.5 |
| 6 | 883 | 257.9 | 11.0 | 4.6 |
| 7 | 842 | 260.3 | 11.9 | 4.8 |
| 8 | 768 | 262.9 | 12.8 | 5.0 |
| 9 | 727 | 265.2 | 14.4 | 5.5 |
| 10 | 1,270 | 271.7 | 19.5 | 6.9 |
| Subtest 5: Efficiency of basic reading comprehension | | | | |
| 5 | 1,445 | 255.2 | 13.1 | 3.7 |
| 6 | 864 | 258.3 | 14.4 | 3.9 |
| 7 | 830 | 262.9 | 16.9 | 4.2 |
| 8 | 756 | 267.7 | 19.2 | 4.6 |
| 9 | 703 | 272.0 | 20.9 | 4.8 |
| 10 | 1,236 | 273.2 | 21.6 | 4.9 |
| Subtest 6: Reading comprehension | | | | |
| 5 | 1,430 | 241.0 | 10.8 | 6.8 |
| 6 | 841 | 245.0 | 12.3 | 6.5 |
| 7 | 812 | 247.1 | 14.1 | 6.7 |
| 8 | 723 | 250.2 | 15.9 | 6.8 |
| 9 | 658 | 252.8 | 17.1 | 7.0 |
| 10 | 1,201 | 252.9 | 17.7 | 7.3 |

raw test score for each subtest within each form. One group is chosen as the reference group and the other is chosen as the focal group. Typically, the reference group is a set of students representing the majority within a population or a group generally known to have higher ability (e.g., male students, White students). Therefore the focal group would be female students or those from a racial/ethnic minority group, such as African American students, as examples. DIF analyses based on gender and race were carried out with assignments of reference and focal groups done in these typical ways.[5]

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 Maentel–Haentszel chi square statistic: A (negligible), B (moderate), and C (significant). Any items in Category C were closely examined for any construct-irrelevant factors that would cause such disparities to exist and could be considered for removal from the assessment and scoring. Negative values indicate that the item was easier for the reference group than expected, whereas positive values indicate the item was easier for the

**Table 8** Key Percentiles for Each RISE Subtest by Grade

| Grade | Pctl | WRDC | VOC | MOR | SEN | MZ | RC |
|---|---|---|---|---|---|---|---|
| 5 | 10 | 243 | 244 | 242 | 242 | 243 | 230 |
| | 25 | 246 | 247 | 245 | 247 | 247 | 235 |
| | 50 | 252 | 252 | 250 | 255 | 252 | 240 |
| | 75 | 259 | 257 | 258 | 262 | 261 | 247 |
| | 90 | 266 | 262 | 265 | 270 | 272 | 253 |
| 6 | 10 | 244 | 244 | 242 | 246 | 244 | 232 |
| | 25 | 248 | 247 | 247 | 251 | 247 | 238 |
| | 50 | 254 | 252 | 254 | 257 | 255 | 243 |
| | 75 | 263 | 258 | 264 | 263 | 265 | 250 |
| | 90 | 271 | 265 | 277 | 271 | 276 | 260 |
| 7 | 10 | 245 | 245 | 244 | 246 | 244 | 233 |
| | 25 | 250 | 250 | 249 | 253 | 250 | 238 |
| | 50 | 258 | 255 | 259 | 259 | 260 | 245 |
| | 75 | 267 | 261 | 270 | 267 | 272 | 255 |
| | 90 | 276 | 271 | 283 | 274 | 287 | 265 |
| 8 | 10 | 246 | 246 | 245 | 247 | 245 | 234 |
| | 25 | 252 | 251 | 252 | 254 | 252 | 240 |
| | 50 | 261 | 259 | 264 | 261 | 265 | 248 |
| | 75 | 270 | 267 | 277 | 269 | 279 | 258 |
| | 90 | 279 | 279 | 288 | 279 | 297 | 270 |
| 9 | 10 | 248 | 248 | 247 | 250 | 246 | 235 |
| | 25 | 255 | 254 | 256 | 257 | 255 | 241 |
| | 50 | 264 | 261 | 267 | 263 | 269 | 250 |
| | 75 | 274 | 272 | 279 | 272 | 287 | 261 |
| | 90 | 284 | 287 | 288 | 285 | 310 | 274 |
| 10 | 10 | 249 | 248 | 248 | 249 | 245 | 234 |
| | 25 | 255 | 254 | 260 | 257 | 254 | 241 |
| | 50 | 262 | 265 | 272 | 272 | 273 | 249 |
| | 75 | 273 | 278 | 283 | 285 | 289 | 264 |
| | 90 | 283 | 295 | 295 | 302 | 306 | 274 |

*Note*. Pctl = percentile; WRDC = word recognition and decoding; VOC = vocabulary; MOR = morphology; SEN = sentence processing; MZ = (MAZE) efficiency of basic reading; RC = reading comprehension.

**Table 9** Reliability for Each RISE Subtest by Grade

| Grade | WRDC | VOC | MOR | SEN | MZ | RC |
|---|---|---|---|---|---|---|
| 5 | .909 | .900 | .871 | .851 | .922 | .604 |
| 6 | .907 | .830 | .889 | .826 | .925 | .719 |
| 7 | .919 | .860 | .902 | .836 | .937 | .773 |
| 8 | .921 | .879 | .912 | .846 | .942 | .816 |
| 9 | .919 | .889 | .920 | .854 | .947 | .833 |
| 10 | .899 | .860 | .909 | .873 | .948 | .831 |

*Note*. WRDC = word recognition and decoding; VOC = vocabulary; MOR = morphology; SEN = sentence processing; MZ = (MAZE) efficiency of basic reading; RC = reading comprehension.

focal group than expected. The analyses were conducted on students in Grades 6–9, and because there were four forms, the total number of comparisons is the number of items times 4 represented by $k$ in Table 16 along with the range of percentages by DIF category and subtest across forms.

The findings/data in Table 16 show very little presence of significant DIF. The largest number of items for any one form of any one subtest was two. The authors did not find any content in the items that was deemed construct irrelevant or biased. This procedure would need to be replicated with other data from this school district or from other school districts to substantiate the claim that these items are in fact generally free from DIF.

**Table 10** Correlations Between Each RISE Subtest, Grade 5

| Subtest | WRDC | VOC | MOR | SEN | MZ | RC |
|---|---|---|---|---|---|---|
| WRDC | — | *.678* | *.747* | *.578* | *.631* | *.607* |
| VOC | .613 | — | *.743* | *.593* | *.644* | *.627* |
| MOR | .665 | .658 | — | *.742* | *.758* | *.711* |
| SEN | .509 | .519 | .639 | — | *.739* | *.650* |
| MZ | .578 | .586 | .679 | .655 | — | *.756* |
| RC | .450 | .462 | .516 | .466 | .564 | — |

*Note*. WRDC = word recognition and decoding; VOC = vocabulary; MOR = morphology; SEN = sentence processing; MZ = (MAZE) efficiency of basic reading; RC = reading comprehension. Values in lower triangle are observed; in upper (italics) are corrected for attenuation.

**Table 11** Correlations Between Each RISE Subtest, Grade 6

| Subtest | WRDC | VOC | MOR | SEN | MZ | RC |
|---|---|---|---|---|---|---|
| WRDC | — | *.783* | *.758* | *.616* | *.661* | *.647* |
| VOC | .679 | — | *.836* | *.660* | *.717* | *.730* |
| MOR | .680 | .718 | — | *.742* | *.753* | *.738* |
| SEN | .533 | .546 | .635 | — | *.712* | *.654* |
| MZ | .605 | .628 | .683 | .622 | — | *.765* |
| RC | .522 | .564 | .589 | .504 | .624 | — |

*Note*. WRDC = word recognition and decoding; VOC = vocabulary; MOR = morphology; SEN = sentence processing; MZ = (MAZE) efficiency of basic reading; RC = reading comprehension. Values in lower triangle are observed; in upper (italics) are corrected for attenuation.

**Table 12** Correlations Between Each RISE Subtest, Grade 7

| Subtest | WRDC | VOC | MOR | SEN | MZ | RC |
|---|---|---|---|---|---|---|
| WRDC | — | *.750* | *.729* | *.611* | *.650* | *.642* |
| VOC | .666 | — | *.794* | *.647* | *.710* | *.738* |
| MOR | .664 | .699 | — | *.732* | *.742* | *.737* |
| SEN | .535 | .549 | .635 | — | *.696* | *.678* |
| MZ | .603 | .637 | .682 | .616 | — | *.765* |
| RC | .541 | .602 | .615 | .545 | .651 | — |

*Note*. WRDC = word recognition and decoding; VOC = vocabulary; MOR = morphology; SEN = sentence processing; MZ = (MAZE) efficiency of basic reading; RC = reading comprehension. Values in lower triangle are observed; in upper (italics) are corrected for attenuation.

## Conclusion

The six-subtest RISE assessment was designed to address a practical educational need by applying a theory-based approach to assessment development. The need was for better assessment information of struggling middle-grades students — those students who typically score below proficient on state English-language arts tests. The theoretical and empirical literature has suggested that overall reading comprehension skills are built upon a foundation of componential reading skills, such as decoding, word recognition, vocabulary, morphology, sentence processing, and efficiency of basic reading. Weaknesses in one or more of these skills could underlie poor reading comprehension performance. Such componential score information is not derivable from traditional reading comprehension tests. We designed subtests targeting these six components.

Further design considerations were imposed to meet practicality and feasibility constraints, specifically, the need for efficient administration (e.g., a 45- to 60-minute limit) and rapid, inexpensive turnaround of scores. Together, the presence of these constraints supported the argument for electronic delivery and scoring.

The results of extensive field testing demonstrate that the RISE battery exhibits adequate subtest reliability and utility, moderate to strong correlations between the subtests, and minimal DIF for each of the grade levels. The sample included multiple waves of students in a school district comprising a mixture of racial/ethnic groups. Evidence of

**Table 13** Correlations Between Each RISE Subtest, Grade 8

| Subtest | WRDC | VOC | MOR | SEN | MZ | RC |
|---|---|---|---|---|---|---|
| WRDC | — | *.737* | *.719* | *.591* | *.624* | *.630* |
| VOC | .663 | — | *.781* | *.620* | *.677* | *.740* |
| MOR | .659 | .699 | — | *.703* | *.725* | *.718* |
| SEN | .522 | .535 | .618 | — | *.701* | *.673* |
| MZ | .581 | .616 | .672 | .626 | — | *.768* |
| RC | .546 | .627 | .619 | .559 | .673 | — |

*Note*. WRDC = word recognition and decoding; VOC = vocabulary; MOR = morphology; SEN = sentence processing; MZ = (MAZE) efficiency of basic reading; RC = reading comprehension. Values in lower triangle are observed; in upper (italics) are corrected for attenuation.

**Table 14** Correlations Between Each RISE Subtest, Grade 9

| Subtest | WRDC | VOC | MOR | SEN | MZ | RC |
|---|---|---|---|---|---|---|
| WRDC | — | *.710* | *.664* | *.561* | *.615* | *.635* |
| VOC | .642 | — | *.736* | *.608* | *.670* | *.732* |
| MOR | .611 | .666 | — | *.666* | *.693* | *.683* |
| SEN | .497 | .530 | .590 | — | *.666* | *.655* |
| MZ | .574 | .615 | .646 | .599 | — | *.751* |
| RC | .556 | .630 | .598 | .553 | .667 | — |

*Note*. WRDC = word recognition and decoding; VOC = vocabulary; MOR = morphology; SEN = sentence processing; MZ = (MAZE) efficiency of basic reading; RC = reading comprehension. Values in lower triangle are observed; in upper (italics) are corrected for attenuation.

**Table 15** Correlations Between Each RISE Subtest, Grade 10

| Subtest | WRDC | VOC | MOR | SEN | MZ | RC |
|---|---|---|---|---|---|---|
| WRDC | — | *.809* | *.744* | *.643* | *.655* | *.708* |
| VOC | .711 | — | *.788* | *.697* | *.711* | *.744* |
| MOR | .673 | .696 | — | *.729* | *.729* | *.712* |
| SEN | .570 | .604 | .650 | — | *.738* | *.765* |
| MZ | .604 | .641 | .677 | .671 | — | *.799* |
| RC | .612 | .628 | .619 | .651 | .709 | — |

*Note.* WRDC = word recognition and decoding; VOC = vocabulary; MOR = morphology; SEN = sentence processing; MZ = (MAZE) efficiency of basic reading; RC = reading comprehension. Values in lower triangle are observed; in upper (italics) are corrected for attenuation.

validity of scores includes strong, but not statistically indistinguishable, intercorrelations among the subtests (see also Mislevy & Sabatini, 2012; O'Reilly, Sabatini, Bruce, Pillarisetti, & McCormick, 2012; Sabatini, 2009; Sabatini, Bruce, & Pillarisetti, 2010; Sabatini, Bruce, Pillarisetti, & McCormick, 2010; Sabatini, Bruce, & Sinharay, 2009). The subtest means and percentiles demonstrate how the relative difficulty and variability of the subtest distributions vary within and across grades.

The adequacy of the measurement properties of the RISE assessment provides the basis for school administrators and teachers to interpret test scores as part of the evidence available for making educational decisions. For example, school administrators might use prevalence estimates of how many students are scoring at low levels on subtests of decoding or word recognition to determine how to plan and allocate resources for interventions targeting those basic subskills (which are usually implemented as supplements to subject-area classes). Classroom teachers can look at evidence of relative strengths and weaknesses across a range of their students to make adjustments to their instructional emphasis in teaching vocabulary, morphological patterns, or assigning reading practice to enhance reading fluency and efficiency. We continue to work with pilot schools and districts to develop professional development packages to assist teachers and administrators in using score evidence to make sound decisions aligned with their instructional knowledge and practices (for other applications, see Mislevy & Sabatini, 2012; O'Reilly et al., 2012; Sabatini, 2009).

**Table 16** Summary of Differential Item Functioning (DIF) Categorizations by Gender and Race Across Grades and Forms

| Subtest | *k* | C+ | B+ | A | B− | C− |
|---|---|---|---|---|---|---|
| Gender | | | | | | |
| WRDC | 200 | 0.0–0.0 | 0.0–2.0 | 96.0–98.0 | 0.0–4.0 | 0.0–0.0 |
| VOC | 152 | 0.0–2.6 | 2.6–5.3 | 78.9–94.7 | 0.0–10.5 | 0.0–2.6 |
| MOR | 128 | 0.0–0.0 | 0.0–0.0 | 96.9–100.0 | 0.0–3.1 | 0.0–0.0 |
| SEN | 104 | 0.0–0.0 | 0.0–3.8 | 92.3–100.0 | 0.0–3.8 | 0.0–0.0 |
| MZ | 170 | 0.0–0.0 | 0.0–2.0 | 97.8–100.0 | 0.0–2.2 | 0.0–0.0 |
| RC | 80 | 0.0–0.0 | 0.0–5.6 | 94.4–100.0 | 0.0–0.0 | 0.0–0.0 |
| Race/ethnicity | | | | | | |
| WRDC | 200 | 0.0–2.0 | 2.0–4.0 | 92.0–98.0 | 0.0–4.0 | 0.0–0.0 |
| VOC | 152 | 0.0–2.6 | 2.6–10.5 | 84.2–97.4 | 0.0–5.3 | 0.0–0.0 |
| MOR | 128 | 0.0–3.1 | 0.0–6.3 | 90.6–100.0 | 0.0–3.1 | 0.0–0.0 |
| SEN | 104 | 0.0–0.0 | 0.0–11.5 | 80.8–100.0 | 0.0–7.7 | 0.0–0.0 |
| MZ | 170 | 0.0–0.0 | 0.0–4.3 | 91.3–100.0 | 0.0–4.3 | 0.0–0.0 |
| RC | 80 | 0.0–0.0 | 0.0–0.0 | 95.0–100.0 | 0.0–5.0 | 0.0–0.0 |

*Note.* WRDC = word recognition and decoding; VOC = vocabulary; MOR = morphology; SEN = sentence processing; MZ = (MAZE) efficiency of basic reading; RC = reading comprehension.

The IRT scaling of parallel forms that is reported here enables users to make comparisons across grades with respect to growth or change in skills. Thus, the scores can be used to gather evidence of the effectiveness of different instructional programs in helping students progress or accelerate their reading skill growth. The battery can also be used for benchmarking and summative purposes such as for tracking student progress within and across school years.

The next steps, now under way, are conducting research that takes the RISE and SARA system in new directions. First, we are designing a wider range of items for each form. This broader item pool should enhance the discrimination across a wide grade and ability range. Second, we are continuing to expand the range of the RISE assessment by building and piloting forms for use in elementary, secondary, and adult literacy settings. Third, we are continuing to evaluate the properties of the tests with special populations, such as English-language learners. Fourth, we are expanding and elaborating on the item types within each of the componential constructs. Fifth, we are expanding our research on providing interpretative guidance for using results to inform decision making at the teacher and school levels, for which the development of proficiency levels and profiles will be useful. Finally, we are working on versions of the RISE and SARA system that can be used in more formative contexts for students and teachers.

In conclusion, the RISE forms of the SARA fill an important gap in assessment of reading difficulties in the middle grades. The RISE forms are a proof of concept that theory-based instruments can be designed to be practically implemented, scored, and interpreted in middle-grades contexts. We are hopeful that the information the RISE assessment provides is of practical utility to educators above and beyond scores obtained on state exams and traditional reading comprehension tests. The ongoing research agenda is to design items and collect evidence to enhance and improve the utility and validity of the RISE and SARA system in a wide range of contexts.

## Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305G04065 and Grant R305F100005 to the Educational Testing Service (ETS) as part of the Reading for Understanding Research Initiative. The opinions expressed are those of the authors and do not represent views of the institute or the U.S. Department of Education. We are extremely grateful to the Institute of Education Sciences and ETS for sponsoring and supporting this research.

## Endnotes

1 This report can be retrieved online (http://dx.doi.org/10.1002/j.2333-8504.2013.tb02315.x).
2 To find out more about the RISE, please visit http://rise.serpmedia.org/ or send an e-mail to RISE_Info@ets.org.
3 Of course, higher level reading comprehension includes even more complex skills that might include interpreting and evaluating texts with respect to author intentions or one's own purposes, critical thinking, making inferences across multiple texts, and so forth.

4 Correct answers are underlined and placed in the first position in the following examples for this and for all subsequent subtest examples.

5 White, Black/African American, and Hispanic/Latino student groups had sufficient sample sizes to be included in the DIF analyses.

## References

Abadzi, H. (2003). *Improving adult literacy outcomes: Lessons from cognitive research for developing countries*. Washington, DC: The World Bank.

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

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.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77 – 117). Newark, DE: International Reading Association.

Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society of Research in Child Development, 58*(10), 1 – 186.

Beck, I. L., & McKeown, M. G. (1991). Social studies texts are hard to understand: Mediating some of the difficulties. *Language Arts, 68,* 482 – 490.

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

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

Bennett, R. E. (2011). *CBAL: Results from piloting innovative K – 12 assessments* (Research Report No. RR-11-23). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02259.x

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

Berninger, V. W., Abbott, R. D., Nagy, W., & Carlisle, J. (2010). Growth in phonological, orthographic, and morphological awareness in Grades 1 to 6. *Journal of Psycholinguistic Research, 39,* 141 – 163.

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. 395 – 479). Reading, MA: Addison-Wesley.

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

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

Carlisle, J. F., & Rice, M. S. (2002). *Improving reading comprehension: Research-based principles and practices.* Baltimore, MD: York Press.

Carlisle, J. F., & Stone, C. A. (2003). The effects of morphological structure on children's reading of derived words. In E. Assink & D. Sandra (Eds.), *Reading complex words: Cross-language studies* (pp. 27 – 52). Amsterdam, The Netherlands: Kluwer.

Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippmann, D. N., . . . White, C. E. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39,* 188 – 215.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies.* New York, NY: Cambridge University Press.

Chall, J. S. (1967). *Stages of reading development*. New York, NY: McGraw-Hill.

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

Council of Chief State School Officers & National Governors Association. (2010). *Common Core State Standards for English language arts*. Washington, DC: Authors. Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

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

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33,* 934 – 945.

Daneman, M. (1988). Word knowledge and reading skill. In M. Daneman, G. E. Mackinnon, & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. *6*, pp. 145 – 175). San Diego, CA: Academic Press.

Deacon, S. H., & Kirby, J. (2004). Morphological awareness: Just "more phonological"? The roles of morphological and phonological awareness in reading development. *Applied Psycholinguistics, 25,* 223 – 238.

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

Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the MMSE: An application of the Mantel–Haenzel and standardization procedures. *Medical Care, 44*(S3), S107– S114.

Duke, N. K., & Carlisle, J. F. (2011). The development of comprehension. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. Afflerbach (Eds.), *Handbook of reading research* (Vol. *4*, pp. 199–228). London, England: Routledge.

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9,* 167–188.

Espin, C. A., Deno, S. L., Maruyama, G., & Cohen, C. (1989, March). *The Basic Academic Skills Samples (BASS): An instrument for the screening and identification of children at risk for failure in regular education classrooms.* Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Fowler, A. E., & Lieberman, I. Y. (1995). The role of phonology and orthography in morphological awareness. In L. Feldman (Ed.), *Morphological aspects of language processing* (pp. 189–209). Hillsdale, NJ: Erlbaum.

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

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1982). *Stanford Achievement Test*. Iowa City, IA: Harcourt Brace Jovanovich.

Gernsbacher, M. A., & Faust, M. (1991). The role of suppression in sentence compression. In G. B. Simpson (Ed.), *Comprehending word and sentence* (pp. 97–128). Amsterdam, The Netherlands: North-Holland.

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

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*, 6–10.

Graves, M. F., Brunetti, G. J., & Slater, W. H. (1982). The reading vocabularies of primary-grade children of varying geographic and social backgrounds. In J. A. Niles & L. A. Harris (Eds.), *New inquiries in reading research and instruction: Thirty-first yearbook of the National Reading Conference* (pp. 99–104). Rochester, NY: National Reading Conference.

Graves, M., & Slater, W. (1987, April). *The development of reading vocabularies in rural disadvantaged students, inner-city disadvantaged students, and middle-class suburban students.* Paper presented at the meeting of the American Educational Research Association, Washington, DC.

Haberman, S. J. (2005). *When can subscores have value?* (Research Report No. RR-05-08). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2005.tb01985.x

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Paul H. Brookes.

Hirsch, E. D. (2003). Reading comprehension requires knowledge of words and the world. *American Educator*, *27*, 10–31.

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

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

Kang, H.-W. (1993). How can a mess be fine? Polysemy and reading in a foreign language. *Mid-Atlantic Journal of Foreign Language Pedagogy, 1,* 35–49.

Kieffer, M. J., & Lesaux, N. K. (2007). Breaking down words to build meaning: Morphology, vocabulary, and reading comprehension in the urban classroom. *Reading Teacher, 61*, 134–144.

Kieffer, M. J., & Lesaux, N. K. (2008). The role of derivational morphological awareness in the reading comprehension of Spanish-speaking English language learners. *Reading and Writing: An Interdisciplinary Journal, 21,* 783–804.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163–182.

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

Kuo, L., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist, 41*, 161–180.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 293–323.

Lesaux, N. K., Kieffer, M. J., Faller, S. E., & Kelley, J. (2010). The effectiveness and ease of implementation of an academic vocabulary intervention for linguistically diverse students in urban middle schools. *Reading Research Quarterly, 45*, 198–230.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological review, 95*(4), 492–527.

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

Mann, V. A., Shankweiler, D., & Smith, S. T. (1984). The association between comprehension of spoken sentences and early reading ability: The role of phonetic representation. *Journal of Child Language, 11,* 627–643.

McClure, E., & Steffensen, M. (1985). A study of the use of conjunctions across grades and ethnic groups. *Research in the Teaching of English, 19*, 217–236.

McCormick, C., Sabatini, J., Bruce, K., Sinharay, S., & O'Reilly, T. (2012, July). *Subscore evaluation for a test of reading skills*. Paper presented at the meeting of the Psychometric Society, Lincoln, NE.

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

McNamara, D. S., & McDaniel, M. A. (2004). Suppressing irrelevant information: Knowledge activation or inhibition? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30,* 465–482.

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. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 119–134). Lanham, MD: Rowman and Littlefield.

Nagy, W., & Anderson, R. C. (1984). The number of words in printed school English. *Reading Research Quarterly, 19*, 304–330.

Nagy, W., Berninger, V. W., & Abbott, R. D. (2006). Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. *Journal of Educational Psychology, 98*, 134–147.

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

National Center for Education Statistics. (2012). *Vocabulary results from the 2009 and 2011 NAEP reading assessments* (NCES Report No. 2013-452). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Olson, R. (2007). Introduction to the special issue on genes, environment, and reading. *Reading and Writing, 20,* 1–11.

O'Reilly, T., & Sabatini, J. (2013). *Reading for understanding: How performance moderators and scenarios impact assessment design* (Research Report No. RR-13-31). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02338.x

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

O'Reilly, T., & Sheehan, K. M. (2009). *Cognitively based assessment of, for and as learning: A framework for assessing reading competency* (Research Report No. RR-09-26). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02183.x

Ouellet, 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.

Papamihiel, N. E., Lake, V., & Rice, D. (2005). Adapting a social studies lesson to include English language learners. *Social Studies and the Young Learner, 17,* 4–7.

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

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

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

Perfetti, C. A. (1985). *Reading ability*. New York, NY: Oxford University Press.

Perfetti, C. A. (1993). Why inferences might be restricted. *Discourse Processes, 16,* 181–192.

Perfetti, C. A. (1994). Psycholinguistics and reading ability. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 849–894). San Diego, CA: Academic Press.

Perfetti, C. A. (1995). Cognitive research can inform reading education. *Journal of Research in Reading, 18,* 106–115.

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

Perfetti, C. A., & Hart, L. (2001). The lexical bases of comprehension skill. In D. S. Gorfien (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 67–86). Washington, DC: American Psychological Association.

Proctor, C. P., Dalton, B., Uccelli, P., Biancarosa, G., Mo, E., Snow, C., & Neugebauer, S. (2011). Improving comprehension online: Effects of deep vocabulary instruction with bilingual and monolingual fifth graders. *Reading and Writing, 24,* 517–544.

Rayner, K. (1997). Understanding eye movements in reading. *Scientific Studies of Reading, 1,* 317–339.

Reynolds, R. E. (2000). Attentional resource emancipation: Toward understanding the interaction of word identification. *Scientific Studies of Reading, 4,* 169–195.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. The context enhancement effect and some tests and extensions of the model. *Psychological Review, 89,* 60–94.

Sabatini, J. P. (2009). From health/medical analogies to helping struggling middle school readers: Issues in applying research to practice. In S. Rosenfield & V. Berninger (Eds.), *Translating science-supported instruction into evidence-based practices: Understanding and applying the implementation process* (pp. 285–316). New York, NY: Oxford University Press.

Sabatini, J., Bruce, K., & Pillarisetti, S. (2010, May). *Designing and implementing school level assessments with district input.* Paper presented at the meeting of the American Educational Research Association, Denver, CO.

Sabatini, J., Bruce, K., Pillarisetti, S., & McCormick, C. (2010, July). *Investigating the range and variability in reading subskills of middle school students: Relationships between reading subskills and reading comprehension for non-proficient and proficient readers*. Paper presented at the meeting of the Society for the Scientific Study of Reading, Berlin, Germany.

Sabatini, J. P., Bruce, K., & Sinharay, S. (2009, June). *Heterogeneity in the skill profiles of adolescent readers.* Paper presented at the meeting of the Society for the Scientific Study of Reading, Boston, MA.

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

Sabatini, J., O'Reilly, T., & Deane, P. (2013). *Preliminary reading literacy assessment framework: Foundation and rationale for assessment and system design* (Research Report No. RR-13-30). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02337.x

Share, D. L. (1997). Understanding the significance of phonological deficits in dyslexia. *English Teacher's Journal*, *51*, 50–54.

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.

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.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Washington, DC: Rand Corporation.

Stahl, S. A., & Nagy, W. E. (2006). *Teaching word meanings.* Mahwah, NJ: Erlbaum.

Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading, 10*, 381–398.

Tong, X., Deacon, S. H., Kirby, J. R., Cain, K., & Parrila, R. (2011). Morphological awareness: A key to understanding poor reading comprehension in English. *Journal of Educational Psychology, 103,* 523–534.

Vellutino, F. R., Tunmer, W. E., Jaccard, J. 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.

Venezky, R. L. (1995). How English is read: Grapheme–phoneme regularity and orthographic structure in word recognition. In I. Taylor & D. R. Olson (Eds.), *Scripts and literacy: Reading and learning to read alphabets, syllabaries, and characters* (pp. 111–129). Dordrecht, The Netherlands: Kluwer Academic.

Venezky, R. L. (1999). *The American way of spelling: The structure and origins of American English orthography*. New York, NY: Guilford Press.

Verhoeven, L., & Perfetti, C. A. (2011). Morphological processing in reading acquisition: A cross-linguistic perspective. *Applied Psycholinguistics, 32,* 457–466.

Walczyk, J., Marsiglia, C. S., Bryan, K. S., & Naquin, P. J. (2001). Overcoming inefficient reading skills. *Journal of Educational Psychology, 93,* 750–757.

Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41*, 85–120.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III. Tests of Achievement.* Itasca, IL: Riverside.

## Suggested citation:

Sabatini, J., Bruce, K., Steinberg, J., & Weeks, J. (2015). *SARA reading components tests, RISE forms: Technical adequacy and test design, 2nd edition* (Research Report No. RR-15-32). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12076

**Action Editor:** James Carlson

**Reviewers:** Paul Deane and Gary Feng