

Academic Achievement Battery:

Development of a Novel Approach to
Assessing Reading Comprehension

Melissa A. Messer, MHS

Jennifer A. Greene, PhD

Executive Summary

The Academic Achievement Battery (AAB) assesses basic academic skills, such as reading, spelling, and math.

A subtest within the AAB, Reading Comprehension: Passages (RC: P), Uses the sentence identification method to assess strengths and weaknesses in reading comprehension across a wide age and grade range.

The RC: P subtest represents an improvement to most traditional reading assessments, which are susceptible to “passageless” comprehension, or the likelihood that an examinee could respond correctly without ever reading the corresponding passage. The evidence presented in this white paper demonstrates the RC: P subtest is a reliable and valid assessment of reading comprehension and, as a result, provides professionals with a valid tool to examine reading comprehension skills.

Introduction

In 1997, a National Reading Panel was established with a goal to evaluate existing research and evidence to find the best ways to teach children to read. Panel members determined five essential components of reading instruction and identified the following necessary steps to reading proficiency: phonemic awareness, phonics, fluency, vocabulary, and reading comprehension (National Reading Panel, 2000; see Figure 1). Most educators agree that although the concept of reading comprehension may seem simple, it is not necessarily easy to teach, learn, or practice. Despite the National Reading Panel's research (concluded in 2000) and following years of research by others on teaching and assessing reading comprehension, "understanding and measurement of this ability has proven elusive" (McGrew, Moen, & Thurlow, 2010, p. 1).

There are various definitions of what constitutes reading comprehension, and as a result, there are a wide variety of methods that have been developed to assess it (Morsy, Kieffer, & Snow, 2010; Pearson & Hamm, 2005). Although multiple definitions exist (with little consensus), even the earliest definitions focused on thinking about text (Thorndike, 1917). Overall, there tends to be agreement regarding the basic building blocks required to master reading comprehension (see Figure 1). More recently, the focus has shifted to the interactive process of reading (National Center for Education Statistics, 2005).

Accurate assessment of reading comprehension is necessary not only to identify reading comprehension difficulties but also to plan and monitor interventions aimed at improving reading comprehension.

A large assortment of formats have been developed and utilized for the purpose of assessing reading comprehension (see Morsy et al., 2010; Spear-Swerling, 2006). In general, these various measures correlate significantly, and quite substantially, with each other. However, there is evidence that the differing formats may tap abilities that underlie reading comprehension (e.g., decoding, vocabulary, listening comprehension, working memory, reading rate, fluency). Each of these methods may be criticized for introducing additional constructs, and these confounds could have a significant impact on what is actually being assessed (e.g., knowledge of the question, vocabulary knowledge, ability to articulate orally to express response [Pearson & Hamm, 2005]). The following presents a brief review of the most common reading comprehension assessment formats. For a more detailed historical review of the foundations of reading comprehension assessment, refer to Paris and Stahl (2005).

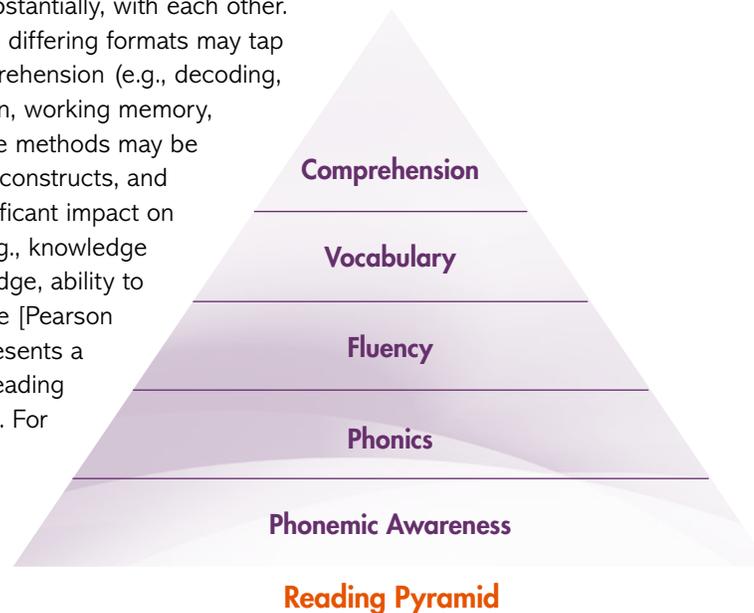


Figure 1. National Reading Panel's steps to reading proficiency.

Accurate assessment of reading comprehension is necessary not only to identify reading comprehension difficulties but also to plan and monitor interventions aimed at improving reading comprehension.

Cloze Procedure

The cloze procedure refers to “reading closure,” which requires the reader to fill in a missing word or words within a sentence or possibly longer text. The word *cloze* is derived from closure in Gestalt theory (Taylor, 1953). The cloze procedure was originally introduced in 1953 as a tool for readability. Initial studies found the method to correlate highly with Flesch (1948) and Dale and Chall (1948) techniques for estimating readability (Taylor, 1953). Following Taylor’s work, Chatel (2001) expanded on the purpose and use of the procedure by indicating it could be used as a way to determine how a reader uses the context of a sentence or passage to get meaning from the text. However, Chatel (2001) indicated some concern for using this with students as a diagnostic tool and believed that test takers tended to focus on the “blank” and used only the immediate context as opposed to attending to the entire passage or text.

Radice (1978) identified several benefits for utilizing this procedure, including ease of administration, ease of interpreting results, ability to provide feedback to a teacher easily, and flexibility. Several modifications of the cloze procedure have been implemented since its original development, including varying the length of sentences or passages, deletion frequency, fixed interval deletion vs. random deletion, allowing of synonyms vs. exact replacement of missing words, and multiple choice options for missing words.

The cloze procedure has been criticized for the ambiguity between whether it assesses individual difference in reading comprehension or if it assesses the “linguistic predictability of the passage to which [the cloze procedure is] applied” (Pearson & Hamm, 2005, p. 24).

Even more concerning is research that has established that cloze tests are typically not sensitive to comprehension that spans a passage. For example, Shanahan, Kamil, and Tobin (1982) used several passage variations to assess correct completion rates in cloze procedures. This included randomizing sentence order within a passage and across passages and using isolated sentences from different passages to form a passage. There were no differences found in the completion rate for the blanks across the various research conditions, indicating that an individual’s ability to fill in the blank was not dependent on the passage context. This reflects Chatel’s (2001) concerns noted earlier—that individuals “completing the blanks” were not integrating text across the passage to complete the task.

Multiple-Choice Questions

One of the most common formats utilized in reading comprehension assessments is the use of multiple-choice questions. In this format, the examinee is required to answer questions based on a passage that he or she reads. One reason for the popularity of this format is its ease of development. It’s also a familiar format for most test takers because it is often used in classroom settings. However, a common concern raised by a number of researchers is in regard to passage independence, or what has been called “passageless” comprehension (Coleman, J. Lindstrom, Nelson, W. Lindstrom, & Gregg, 2010; Ready, Chaudhry, Schatz, & Strazzullo, 2012). This refers to the likelihood that an examinee could respond correctly to multiple-choice questions (typically based on prior knowledge) without ever reading the corresponding passage. Although it can be argued that “prior knowledge” is a part of

“Passageless” comprehension refers to the likelihood that an examinee could respond correctly to multiple-choice questions without ever reading the corresponding passage.

successful reading comprehension, “passage-independent items are recognized as major threats to content validity” (Coleman, et al., 2010, p. 244).

There are several examples of research uncovering passage-independence on standardized reading comprehension measures—including the Minnesota Scholastic Aptitude Test (Fowler & Kroll, 1978), the Stanford Achievement Test (Lifson, Scruggs, & Bennion, 1984), the Scholastic Achievement Test (SAT; e.g., Daneman & Hannon, 2001; Katz, Lautenschlager, Blackburn, & Harris, 1990), the Gray Oral Reading Test (GORT; Keenan & Betjemann, 2006), the Nelson-Denny Reading Comprehension Test (Coleman et al., 2010), the Canadian Adult Achievement Test (CAAT; Roy-Charland, Colangelo, Foglia, & Reguigui, 2017), and the Wechsler Individual Achievement Test, Third Edition (WIAT-III; Roy-Charland et al., 2017).

Development of a New Reading Comprehension Assessment

The various forms of the [Academic Achievement Battery](#) (AAB; Messer, 2014a, 2014b, 2017) were designed to measure aspects of academic achievement in children and adults ages 4 to 85 years (see Figure 2 for an overview of each of the three AAB forms). The AAB was designed to assess basic academic skills including letter and word reading, spelling, reading comprehension, and mathematical calculation. See Table 1 for a description of each subtest included on the AAB Standard Form. The AAB is intended for use by professionals who need a quick and easy-to-administer assessment of the basic areas of achievement with a focus on reading comprehension.

The AAB was designed to assess basic academic skills including letter and word reading, spelling, reading comprehension, and mathematical calculation.

Development of the Reading Comprehension: Passages Subtest

Assessment of basic reading and reading comprehension is often a part of academic testing and has historically been included in many comprehensive measures of achievement, such as the Kaufman Test of Educational Achievement, Second Edition (KTEA-II; Kaufman & Kaufman, 2004); the Kaufman Test of Educational Achievement, Third Edition (KTEA-III; Kaufman & Kaufman, 2014); the Wechsler Individual Achievement Test, Third Edition (WIAT-III; Wechsler, 2009); the Woodcock-Johnson Tests of Achievement, Third Edition (WJ-III; Woodcock, McGrew, & Mather, 2007); and the Wide Range Achievement Test 4 (WRAT4; Wilkinson & Robertson, 2006). In general, most tests of reading comprehension (especially those that are part of a larger academic battery) tend to be broad measures that by themselves do not pinpoint specific component abilities or specific comprehension processes (Spear-Swerling, 2006).

Several considerations were made when developing the Reading Comprehension: Passages (RC: P) subtest for the AAB. First, it needed to be applicable to a wide age range (ages 5 to 85+ years). Next, it needed to be administered in a relatively short timeframe given that it is one of several other subtests included in a comprehensive academic achievement battery. Finally, (and most importantly), the format and function of the subtest needed to be established prior to development.

	AAB	AAB Comprehensive	AAB Screening
What it does	Delivers a quick measure of basic academic skills, including a reading comprehension subtest	Provides a complete assessment of an individual's overall performance on seven disparate aspects of achievement	Offers a snapshot of an individual's performance in four areas of achievement, including a measure of writing
Administration and scoring time	15-30 minutes to administer; 5-10 minutes to score	90 minutes to administer; 15 minutes to score	15-30 minutes to administer; 5-10 minutes to score
When to use it	To obtain a quick and accurate measure of an individual's performance that includes a reading comprehension subtest	To conduct an in-depth and complete assessment of academic achievement	To perform a fast and reliable screening of academic achievement that offers an optional writing subtest
Areas assessed	<p>Subtests: Letter/Word Reading Spelling Mathematical Calculation Reading Comprehension: Passages</p> <p>Composites: Reading Total AAB</p>	<p>Subtests: Reading Foundational Skills Letter/Word Reading Reading Fluency Reading Comprehension: Words and Sentences Reading Comprehension: Passages Listening Comprehension: Words and Sentences Listening Comprehension: Passages Oral Fluency Oral Expression Oral Production Pre-Writing Skills Spelling Written Comprehension Mathematical Calculation Mathematical Reasoning</p> <p>Composites: Basic Reading Mathematical Calculation Mathematical Reasoning Listening Comprehension Expressive Communication Written Expression Reading Comprehension AAB Total Comprehensive</p>	<p>Subtests: Letter/Word Reading Spelling Mathematical Calculation Written Composition (optional)</p> <p>Composite: Screening AAB Total</p>
How it helps clinicians	Offers a quick, efficient measure of academic achievement that includes a Reading Composite score, which provides more data to understand an individual's reading skills; IQ discrepancy data are available	Provides a complete assessment of an individual's academic skills that is suitable for use in eligibility decisions or intervention planning; IQ discrepancy data are available	Delivers a fundamental evaluation of academic skills for those referred for learning or vocational concerns; IQ discrepancy data are available

Figure 2. Overview of the three forms of the Academic Achievement Battery (AAB).

Table 1
Description of AAB Standard Form Subtests

Subtest	Acronym	Description
Letter/Word Reading	LWR	Letter Reading requires the examinee to identify lowercase and uppercase letters. Word Reading requires the examinee to pronounce words of increasing difficulty.
Spelling	SP	Letter Writing requires the examinee to write lowercase and uppercase letters. Word Writing requires the examinee to correctly spell words of increasing difficulty.
Reading Comprehension: Passages	RC: P	RC: P requires the examinee to read passages of increasing difficulty and draw a line after each sentence.
Mathematical Calculation	MC	Part 1 requires the examinee to provide oral and written responses to math problems. Part 2 requires the examinee to complete increasingly difficult math calculations in a timed task.

Development of this subtest was based on the definition provided by the RAND Reading Study Group (Snow, 2002), which indicates that reading comprehension “is the process of simultaneously extracting and constructing meaning through interaction and involvement with written language.” Moreover, it is universally agreed that reading comprehension is a multidimensional construct that can be easily influenced by many external sources; the RAND Reading Study Group (Snow, 2002) has identified four factors: the reader (e.g., his or her current skills, knowledge, and preferences), the text being read (e.g., vocabulary, structure, knowledge assumed, format, and reading level), the reading activity (e.g., reading a Web site versus a novel), and reading over time (e.g., comprehension is highly influenced by cognitive development).

These factors were considered when developing the RC: P subtest:

- 1 **The reader:** Because of the wide age range being assessed with the AAB, it was important the paradigm worked at all age and grade levels.
- 2 **The text:** Multiple indexes were analyzed during development to determine the grade appropriateness of the text.
- 3 **The reading activity:** Both nonfiction and fiction passages were used to generalize real-world reading done by both students and adults.
- 4 **Reading over time:** Again, because of the large age span of the test, both the passages used and the format chosen needed to be appropriate for all ages. As an example, it has been found that working memory has more of an impact on reading comprehension in younger children (ages 8-11 years) but has less of an impact as individuals’ age, while knowledge and vocabulary begin to account for more of the variance as the reader progresses through adolescent years (Siegel, 1994).

Moreover, as part of the development of the AAB, the author examined a range of methodologies utilized to assess reading comprehension (see Morsy et al., 2010; Paris & Stahl, 2005; Spear-Swerling, 2006). In reviewing this literature, the method of sentence identification was selected as a possible format.

Sentence identification has been used previously to represent reading, including work by Guilford (1967, 1988) in his Structure of Intellect model, as well as Brown, Wiederholt, and Hammill (2008) in the Contextual Fluency subtest of the Test of Reading Comprehension-Fourth Edition (TORC-4). The TORC-4 requires examinees to identify individual words within a passage, with each passage printed in uppercase letters without punctuation or spaces between words. The paradigm is also based on the research summarized by Scott (2009), which illustrates that sentence comprehension (or more specifically, general sentence-level syntactic/semantic abilities) is a requirement of reading comprehension. Scott (2009)

points out that it is still important to ensure that the task is not decontextualized, and more specifically, “the syntax of complex sentences poses challenges that are not accounted for by text-level processes such as relating sentences or reading beneath the lines to draw inferences” (p. 189).

To develop the passages for the AAB RC: P subtest, a list of categories and topics was first created. Next, specific grade and reading levels were specified as a target for each topic. Editorial and quality assurance staff then reviewed the passages. Several common reading indexes were used to determine readability. The Flesch Reading Ease and Flesch-Kincaid Grade Level (Flesch, 1948) were determined for each passage using the readability statistics function in Microsoft Word. These readability indexes are based on research by Kincaid, Fishburne, Rogers, and Chissom (1975). The Lexile measure for each passage was also determined (MetaMetrics, 2013). For each passage, word count, sentence count, percentage of passive sentences, and mean sentence length were calculated. Also calculated was mean log word frequency, which is the logarithm of the number of times a word appears in each 5 million words of the MetaMetrics research corpus of 571 million words. The mean log word frequency is the average of all such values for words that appear in the analyzed text. Thirteen fiction and nonfiction passages were initially developed for this task. Following the first phase of development (pilot phase), one passage was replaced and one was revised to be more consistent with the other passages in terms of the total number of sentences. The final version of the subtest

Table 2
Item Characteristics and Skills: AAB Reading Comprehension: Passages Subtest

Passage no.	Item characteristic/skill								
	Word count	Sentence count	Passive sentences (%)	Flesch Reading Ease	Flesch-Kincaid Grade Level	Lexile (L)	Mean sentence length (words)	Mean log word frequency	
1	21	4	0	100.0	0.0	80	5.3	3.8	
2	44	7	0	100.0	0.2	300	6.3	3.6	
3	109	13	0	95.0	2.0	460	8.4	3.8	
4	152	16	0	90.3	3.0	560	9.5	3.7	
5	113	10	0	89.0	3.6	700	11.3	3.7	
6	115	10	10	89.9	3.5	730	11.5	3.7	
7	154	10	20	68.8	7.5	960	15.4	3.6	
8	164	13	46	58.8	8.1	970	12.6	3.2	
9	202	13	15	49.2	10.2	1,080	15.5	3.3	
10	176	10	20	49.0	10.7	990	14.6	3.4	
11	235	12	41	46.2	11.3	1,220	19.6	3.4	
12	192	12	25	31.2	12.8	1,210	15.9	3.0	
13	235	10	20	10.5	17.6	1,460	23.5	3.1	

Note. Flesch Reading Ease and Flesch-Kincaid Grade Level derived from J. P. Kincaid, R. P. Fishburne, R. L., Rogers, & B. S. Chissom, (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. Chief of Naval Technical Training: Naval Air Station Memphis.

consists of 13 passages, with each examinee reading three passages. Table 2 provides characteristics of each of the 13 passages.

During the first phase (pilot phase) of development, multiple-choice questions (both literal and inferential) of varying difficulty were created. Each item contained four response options. At the same time, examinees were instructed to draw a line at the end of each sentence. In the second phase (refinement phase) of development, the same questions were used, but they were modified to be in an open-ended response format and were scored as incorrect (0), partially correct (1), or completely correct (2). Again, the sentence identification procedure was used.

Throughout the development process, an expert panel and a bias panel were consulted. The expert panel was composed of curriculum experts in reading, mathematics, and writing and a school psychologist. All members of the expert panel had experience in test construction at both the state and national level. The expert panel reviewed items from all AAB subtests, including the RC: P subtest, to ensure that content reflected the construct

intended, to provide feedback on the face validity of each subtest and the quality of the items, to review and make suggestions for the scoring rubrics, and to ensure tasks were appropriate across the age or grade range.

Pilot and Refinement Samples

The AAB went through three phases of data collection. The first phase, pilot testing, was conducted during the winter of 2011 with 133 participants ages 4 to 83 years (see Table 3). Examinees were selected based on demographic characteristics (i.e., age, gender, education level, and ethnicity). All cases were checked for accuracy and scored by trained scorers. The refinement version of the AAB was administered to a sample of 280 individuals ages 4 to 70 years (see Table 4) during the spring of 2011. Again, examinees were selected based on demographic characteristics (i.e., age, gender,

The expert panel was composed of curriculum experts in reading, mathematics, and writing and a school psychologist. All members of the expert panel had experience in test construction at both the state and national level.

Table 3
Demographic Characteristics of the AAB Pilot Sample

Characteristic	Total
<i>n</i>	133
Gender (%)	
Male	40.6
Female	59.4
Age (years)	
<i>M</i>	26.4
<i>SD</i>	20.8
Range	4-83
Race/ethnicity (%)	
Caucasian	65.4
African American	10.5
Hispanic	20.3
Other	3.8
Education level^a (%)	
<12 years	9.0
12 years	29.3
13-15 years	26.3
16+ years	35.3

^aParent education level was used for individuals ages 4 to 21 years. Percentages may not sum to 100% due to rounding.

Table 4
Demographic Characteristics of the AAB Refinement Sample

Characteristic	Total
<i>n</i>	280
Gender (%)	
Male	50.4
Female	49.6
Age (years)	
<i>M</i>	21.1
<i>SD</i>	18.0
Range	4-70
Race/ethnicity (%)	
Caucasian	58.2
African American	19.6
Hispanic	13.9
Other	8.2
Education level^a (%)	
<12 years	13.2
12 years	23.2
13-15 years	21.4
16+ years	42.1

^aParent education level was used for individuals ages 4 to 21 years. Percentages may not sum to 100% due to rounding.

education level, and ethnicity). All cases were checked for accuracy and scored by trained research assistants.

Standardization Sample

From January 2013 through March 2014, the standardization sample was collected from individuals in 30 states. Two samples were created for standardization (see Table 5): an age-based sample, which was based on 1,274 individuals between the ages of 4 and 83 years, and a grade-based sample, which was based on 1,447 individuals (737 from fall; 710 from spring) between the ages of 4 and 19 years. Both samples are representative of the 2012 U.S. Census in terms of age, gender, ethnicity, and education level. For individuals ages 4 to 21 years, the consenting parent's highest level of education was used to determine education level.

A subset of the age-based sample was administered various academic achievement and reading diagnostic measures. The demographic characteristics of each of these validity samples can be found in Table 6. In addition, a subset of the age-based sample was administered the AAB a second time. The interval between the two test administrations ranged from 7 to 49 days, with a median test-retest interval of 18 days. See Table 7 for the demographic characteristics of the test-retest sample. Participants in both subsamples were systematically

Table 5
Demographic Characteristics
of the AAB Standardization Sample

Characteristic	Sample		
	Age-based	Grade-based	
		Fall	Spring
<i>n</i>	1,274	737	710
Gender (%)			
Male	49.0	49.2	49.3
Female	51.0	50.8	50.7
Age (years)			
<i>M</i>	21.75	11.1	11.0
<i>SD</i>	16.64	4.1	4.3
Range	4-83	4-18	4-19
Race/ethnicity (%)			
Caucasian	61.9	59.4	59.5
African American	12.0	12.0	12.0
Hispanic	19.7	19.7	19.7
Other	8.9	8.9	8.9
Education level^a (%)			
<12 years	10.9	11.0	10.9
12 years	26.6	26.6	26.6
13-15 years	27.9	27.8	27.9
16+ years	34.6	34.6	34.6

^aParent education level was used for individuals ages 4 to 21 years. For individuals ages 22 years and older, actual obtained education level was used.

Table 6
Demographic Characteristics of the AAB Construct Validity Samples

Characteristic	WIAT-III	KTEA-II	WJ-III	WRAT4	FAR
<i>N</i>	18	34	48	47	85
Gender (%)					
Male	58.3	44.1	52.1	46.8	56.5
Female	41.7	55.9	47.9	53.2	43.5
Age (years)					
<i>M</i>	12.78	16.91	27.96	30.23	13.43
<i>SD</i>	2.90	4.79	19.42	18.21	3.58
Race/ethnicity (%)					
Caucasian	72.2	67.6	50	66	50.6
African American	5.6	2.9	20.8	0.0	15.3
Hispanic	22.2	26.5	27.1	27.7	23.5
Other	0.0	2.9	2.1	6.4	10.6
Education level (%)					
<12	38.9	26.5	25	19.1	10.6
12	22.2	29.4	35.4	42.6	28.2
13-15	11.1	26.5	14.6	19.1	23.5
16+	27.8	17.6	25	19.1	37.6

Note. WIAT-III = Wechsler Individual Achievement Test, Third Ed. (Wechsler, 2009); KTEA-II = Kaufman Test of Educational Achievement, Second Ed. (Kaufman & Kaufman, 2004); WJ-III = Woodcock-Johnson Tests of Achievement, Third Ed. (Woodcock, McGraw, & Mather, 2007); WRAT4 = Wide Range Achievement Test, Fourth Ed. (Wilkinson & Robertson, 2006); FAR = Feifer Assessment of Reading (Feifer & Gerhardstein Nader, 2015).

^aParent education level is reported for individuals ages 4 to 21 years.

Table 7
Demographic Characteristics of
the AAB Test–Retest Sample

Characteristic	Total
<i>n</i>	142
Gender (%)	
Male	47.2
Female	52.8
Age (years)	
<i>M</i>	30.25
<i>SD</i>	18.41
Range	5-74
Race/ethnicity (%)	
Caucasian	66.2
African American	11.3
Hispanic	17.6
Other	4.9
Education level (%)	
<12 years	10.6
12 years	31
13-15 years	31.7
16+ years	26.8

Note. Parent education level is reported for individuals ages 4 to 21 years.

recruited, so the resulting samples had approximately the same proportion of males and females and racial/ethnic proportions similar to the U.S. population. Moreover, a wide variety of education levels were obtained.

Procedures

During each phase of data collection, data collectors were selected based on appropriate experience administering performance-based assessments and access to needed populations. Data collectors were responsible for obtaining informed consent and administering the AAB to examinees and scoring their responses. Each data collector was asked to review the administration and scoring guidelines prior to administering his or her first protocol. Each protocol was thoroughly reviewed for completeness and accuracy by trained research assistants after submission. Each data collector received feedback, and examiners who had difficulty with administration or scoring were asked to submit a second protocol for detailed review before they were approved to collect data for the standardization sample. Throughout standardization, each incoming protocol was checked

for completeness and accuracy, and it was then entered into SPSS (Version 18) by trained research assistants. The data were cleaned and checked for missing data before analyses were conducted.

Measures

In addition to the AAB, participants completed a variety of academic achievement and reading diagnostic measures. The KTEA-II (Kaufman & Kaufman, 2004) is a comprehensive achievement test for individuals ages 4 years, 6 months to 25 years. It includes 14 subtests that make up eight composite measures of achievement: reading, math, written language, oral language, sound-symbol, oral fluency, decoding, and reading fluency. The WIAT-III (Wechsler, 2009) is a comprehensive achievement test for individuals ages 4 to 50 years. It includes 15 subtests that make up seven composite measures of achievement: oral language, basic reading, total reading, reading comprehension and fluency, written expression, mathematics, and math fluency. The WJ-III (Woodcock, McGrew, & Mather, 2007) is a comprehensive achievement test for individuals ages 2 to 90+ years. The core test includes 13 subtests that address 13 clusters of achievement: broad reading, oral language, broad math, math calculation skills, broad written language, written expression, academic skills, academic fluency, academic applications, brief reading, brief math, brief writing, and brief achievement. The WRAT4 (Wilkinson & Robertson, 2006) is a screening achievement test for individuals ages 5 to 94 years. It includes four subtests (reading, sentence comprehension, spelling, and math computation) and one composite measure of reading. The Feifer Assessment of Reading (FAR; Feifer & Gerhardstein Nader, 2015) is a comprehensive reading test designed to assess the underlying cognitive and linguistic processes that support proficient reading skills. It includes 15 individual subtests that make up five indexes: the Phonological Index, Fluency Index, Comprehension Index, Mixed Index, and Total Index.

Results

Pilot Data Analysis

Data were analyzed separately for pilot and refinement phases. First, during the pilot phase, the relationship between the multiple-choice format and the sentence-identification format was investigated by examining the correlation between correct answers on the multiple-choice questions for each passage and the correct number of sentences identified for each passage. Correlations were significant across all passages and ranged from .51-.73 across individual passages (Messer, 2014a).

Refinement Data Analysis

Next, the relationship between the open-ended responses (scored 0, 1, or 2 points) and the sentence-identification technique was investigated by examining the correlations. For each passage, the total number of correct sentences identified was significantly correlated with the total points awarded for the multiple-choice comprehension items (.54-.87). The range in these correlations is a result of examining each individual passage across a large age range (ages 5-85 years). Among the questions, several items were rarely missed, even with the 2-point response option (Messer, 2014a).

Reliability and Validity

In addition to examining the relationship sentence identification had with other approaches for assessing reading comprehension, reliability and validity analysis were conducted with two larger samples—the age-based standardization sample and the grade-based standardization sample. See Table 5 for demographic information for these samples.

Reliability refers to a test's stability, consistency, and accuracy. When used to measure a stable construct, scores that are highly reliable will yield consistent, accurate results across factors such as time and examiner. The following section considers several indicators of reliability, including internal consistency, standard error of measurement (*SEM*), and stability of test scores over time.

Internal consistency of the RC: P subtest of the AAB was high in the standardization samples, with a Cronbach's alpha of .88 in the age-based sample and .81 for both the fall and spring grade-based samples. Similarly, the subtest had small *SEMs*, ranging from 3.00 to 7.65 in the age-based sample and from 4.50 to 6.18 in both the fall and spring grade-based samples (Messer, 2014a).

Test score stability refers to the extent to which an individual's test performance remains constant over time. The stability of the AAB subtest and composite scores over time was evaluated by retesting a subset of individuals from the standardization sample. The correlation between Time 1 scores and Time 2 RC: P subtest scores was .81, which indicates a high degree of temporal stability (Messer, 2014a).

A valid test is one that accurately measures the psychological construct for which it is intended. Test validity is multidimensional in nature and should be evaluated using a variety of different sources and methodologies, each providing unique evidence that supports the validity of the test.

Construct validity with other omnibus measures of achievement and reading is evidence of the validity of the RC: P subtest. The construct validity of the RC: P subtest of the AAB was demonstrated by evaluating the correlations found between the RC: P subtest and related subtests on other tests of passage comprehension (e.g., WIAT-III, KTEA-II, WJ-III, WRAT4, and FAR). The RC: P subtest correlates well with the WIAT-III Reading Comprehension subtest ($r = .40$), the KTEA-II Reading Comprehension subtest ($r = .45, p < .01$), the WJ-III Passage Comprehension subtest ($r = .32, p < .05$), the WRAT4 Sentence Comprehension subtest ($r = .63, p < .01$), and the FAR Silent Reading Fluency: Comprehension subtest ($r = .44, p < .01$). It is worth noting that although the measures

correlate, this does not indicate they are measuring identical functions. As a result, additional analyses were conducted with related constructs such as fluency (Messer, 2014a).

In addition to examining the relationship between the RC: P subtest and tests that purport to measure reading comprehension, the author followed the research by Fuchs et al. (Fuchs, Fuchs, Hosp, & Jenkins, 2001), which cited that the strongest correlation between "true" reading comprehension should be between reading and oral fluency. On the AAB, the Reading Fluency and RC: P subtests were found to have significant correlations in both children (ages 4 to 18 years) and adults (ages 19 years and older; $r = .34, p < .01$; $r = .29, p < .01$, respectively). When examining this relationship with achievement tests that utilize more traditional methods, similar correlations have been reported (WIAT-III Oral Fluency and Reading Comprehension subtests, $r = .47$ for the grade-based sample; WJ-III Oral Reading and Passage Comprehension subtests, $r = .51$; FAR Oral Reading Fluency and Silent Reading Fluency-Comprehension subtests, $r = .28$) (Messer, 2014a).

Conclusion

The author of the AAB examined a novel approach to assess reading comprehension that is not typically utilized in omnibus measures of achievement. The AAB Reading Comprehension: Passages subtest was developed to identify strengths and weaknesses in reading comprehension across a wide age and grade range. The evidence presented here demonstrates the AAB Reading Comprehension: Passages subtest is a reliable and valid assessment of reading comprehension and, as a result, it provides professionals with a valid tool to examine reading comprehension skills. As mentioned earlier, most tests of reading comprehension (especially those that are part of a larger academic battery) tend to be broad measures that by themselves do not pinpoint specific component abilities or specific comprehension processes (Spear-Swerling, 2006). The AAB, like other subtests included in academic achievement batteries, does not attempt to pinpoint specific comprehension processes. Instead, it assists in identifying strengths and weaknesses in reading comprehension through a brief assessment (6 minutes to administer) that is easy to administer and offers unbiased scoring.

The AAB Reading Comprehension: Passages subtest is a reliable and valid assessment of reading comprehension.

References

- Brown, V. L., Wiederholt, J. L., & Hammill, D. D. (2008). *Test of Reading Comprehension—Fourth Edition*. Austin, Texas: PRO-ED.
- Chatel, R. G. (2001). Diagnostic and instructional uses of the cloze procedure. *New England Reading Association Journal*, 37, 3-6.
- Coleman, C., Lindstrom, J., Nelson, J., Lindstrom, W., & Gregg, K. N. (2010). Passageless comprehension on the Nelson-Denny Reading Test: Well above chance for university students. *Journal of Learning Disabilities*, 43, 244-249.
- Dale, E., & Chall, J. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27, 11-20.
- Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology: General*, 130, 208-223.
- Feifer, S., & Gerhardstein Nader, R. (2015). *Feifer Assessment of Reading*. Lutz, FL: PAR.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32, 221-233.
- 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.
- Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.
- Guilford, J. P. (1988). Some changes in the structure-of-intellect model. *Educational and Psychological Measurement*, 48, 1-4.
- Fowler, B., & Kroll, B. M. (1978). Verbal skills as factors in the passageless validation of reading comprehension tests. *Perceptual and Motor Skills*, 47, 335-338.
- Katz, S., Lautenschlager, G. J., Blackburn, A. B., & Harris, F. H. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Science*, 1, 122-127.
- Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Test of Educational Achievement, Second Edition*. San Antonio, TX: Pearson.
- Kaufman, A. S., & Kaufman, N. L. (2014). *Kaufman Test of Educational Achievement, Third Edition*. San Antonio, TX: Pearson.
- Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading*, 10, 363-380.
- 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.
- Lifson, S., Scruggs, T. E., & Bennion, K. (1984). Passage independence in reading achievement tests: A follow-up. *Perceptual and Motor Skills*, 58, 945-946.
- McGrew, K. S., Moen, R. E., & Thurlow, M. L. (2010). Cognitive and achievement differences between students with divergent reading and oral comprehension skills: Implications for accessible reading assessment research. Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.
- Messer, M. A. (2014a). *Academic Achievement Battery Comprehensive Form*. Lutz, FL: PAR.
- Messer, M. A. (2014b). *Academic Achievement Battery Screening Form*. Lutz, FL: PAR.
- Messer, M. A. (2017). *Academic Achievement Battery Standard Form*. Lutz, FL: PAR.
- MetaMetrics. (2013). Matching readers with texts. Retrieved from <http://www.metametricsinc.com/lexileframework-reading/>
- Morsy, L., Kieffer, M., & Snow, C. (2010). *Measure for measure: A critical consumers' guide to reading comprehension assessments for adolescents*. Final Report from Carnegie Corporation of New York's Council on Advancing Adolescent Literacy. Carnegie Corporation of New York.
- National Center for Education Statistics. (2005). *2009 NAEP reading framework*. Washington, DC: Author.
- National Reading Panel, National Institute of Child Health & Human Development. (2000). *Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.
- Paris, S. G., & Stahl, S. A. (Eds.). (2005). *Children's reading comprehension and assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices—Past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment*, 13-69.
- Radice, F. W. (1978). Using the cloze procedure as a teaching technique. *English Language Teaching Journal*, 32, 201-204.
- Ready, R. E., Chaudhry, M. F., Schatz, K. C., & Strazzullo, S. (2012). "Passageless" administration of the Nelson-Denny Reading Comprehension Test: Associations with IQ and reading skills. *Journal of Learning Disabilities*, 46, 377-384.
- Roy-Charland, A., Colangelo, G., Foglia, V., & Reguigui, L. (2017). Passage independence within standardized reading comprehension tests. *Reading and Writing*, 1-16.

- Scott, C. M. (2009). A case for the sentence in reading comprehension. *Language, Speech, and Hearing Services In Schools, 40*, 184-191.
- Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*, 229-255.
- Siegel, L. S. (1994). Working memory and reading: A life-span perspective. *International Journal of Behavioral Development, 17*, 109-124.
- Snow, C. (2002). *Reading for understanding: Toward a R&D program in reading comprehension*. Santa Monica, CA: RAND.
- Spear-Swerling, L. (2006). Assessment of reading comprehension. Retrieved from http://www.ldonline.org/spearswerling/Assessment_of_Reading_Comprehension
- SPSS Inc. Released 2009. PASW Statistics for Windows, Version 18.0. Chicago, IL: SPSS Inc.
- Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*, 415-433.
- Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational psychology, 8*, 323.
- Wechsler, D. (2009). *Wechsler Individual Achievement Test, Third Edition*. San Antonio, TX: Pearson.
- Wilkinson, G. S., & Robertson, G. J. (2006). *Wide Range Achievement Test 4*. Lutz, FL: PAR.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). *Woodcock-Johnson III Normative Update*. Rolling Meadows, IL: Riverside.

PAR • 16204 N. Florida Ave. • Lutz, FL 33549 • 1.800.331.8378 • www.parinc.com

Copyright © 2014, 2017 by PAR. All rights reserved. May not be reproduced in whole or in part in any form or by any means without written permission of PAR.

To cite this document, use:

Messer, M. A., & Greene, J. A. (2018). *Academic Achievement Battery: Development of a novel approach to assessing reading comprehension* [white paper]. PAR.



PAR
Melissa A. Messer, MHS

Director of
Product Development
mmesser@parinc.com
1.800.331.8378

Melissa A. Messer, MHS, is the Director of Product Development at PAR. She oversees all new product acquisitions and development and manages a team of project directors and quality control specialists. Melissa received her bachelor's and master's degrees from the University of Florida.

Melissa joined PAR in 2002. During her time at PAR, she has worked on the development of more than 30 products. Melissa is the author of several PAR products, including the Academic Achievement Battery (AAB) series, the Working Styles Assessment (WSA), and the Work Values Inventory (WVI). In addition, she has coauthored multiple editions of John Holland's Self-Directed Search (SDS).



PAR
Jennifer A. Greene, PhD

Coordinator—Digital Product
Design & Psychometrics
jgreene@parinc.com
1.800.331.8378

Jennifer A. Greene, PhD, works in Research and Development at PAR. She holds a PhD in Measurement and Evaluation from the University of South Florida. In her role at PAR, she is responsible for conducting statistical analyses and norming assessments, as well as developing digital assessments.

PAR[®]