A Validation Study of a Middle Grades Reading Comprehension Assessment

Lori Severino  
School of Education  
Drexel University  
Philadelphia, PA, USA  
las492@drexel.edu

Mary Jean Tecce DeCarlo and Toni Sondergeld  
School of Education  
Drexel University  
Philadelphia, PA, USA

Meltem Izzetoglu  
Villanova University  
Philadelphia, PA, USA

Alia Ammar  
Drexel University  
Philadelphia, PA, USA

Abstract

A student’s reading skill is essential to learning. Assessing reading skills, specifically comprehension, is difficult. In the middle grades, students read to learn; and their teachers need a quick, easy assessment that provides immediate data on reading comprehension skill. This study explores the holistic validation approach of one eighth-grade informational text with comprehension questions currently included in the Adolescent Comprehension Evaluation (ACE). Thirty-three eighth-grade students from four different schools participated in the study. Multiple forms of validity evidence were used including test content, response process, internal structure, relationship to other variables, and consequences of testing. These multiple forms of validity evidence provided the researchers with insights into comprehension questions that would not have been uncovered using psychometric means of validation alone. The results of this study support ACE as a direct measure of middle grades students’ reading comprehension.

Keywords: adolescent, assessment, comprehension, validation process
Young adolescent readers need support with meaning making. By the time students are in the middle grades, teachers focus their instruction on reading complex texts and thinking critically about those texts. By the middle grades, most readers who struggle “can read words accurately, but they do not comprehend what they read, for a variety of reasons” (Biancarosa & Snow, 2004, p. 8). To support struggling readers in the middle grades, teachers need a way to assess comprehension that is quick and easy to administer to a whole group and can provide information to inform instruction. If assessment results can help identify when and where students’ skills break down, teachers can develop more effective interventions (Morsy, Kieffer, & Snow, 2010).

Adolescent Comprehension Evaluation (ACE) is a web-based application that a teacher can use on a regular basis to monitor reading comprehension in 5–7 min with an entire class of students. The questions for each narrative or information passage are based on the Common Core State Standards (CCSS) for the particular grade level. Once a class or student has read the passage and answered the 10 questions, the teacher receives immediate data regarding total score, time spent reading and answering the questions, and which questions are giving which students difficulty, as well as much more specific data per student.

The purpose of this study was to test the validity evidence in relation to the ACE assessment with the intent of producing student results that allow classroom teachers to use the information to direct their instruction. This research was designed with the mindset that assessment development is an iterative process that requires multiple rounds of data collection from numerous sources in order to create an assessment that elicits valid and reliable results for an intended purpose. With this in mind, the specific overarching research question was: To what extent do ACE Assessment data provide validity evidence that meets criteria for being considered a valid and reliable assessment of student reading comprehension? The validity evidence considered in this study included test content, response processes, internal structure, relationship to other variables, and consequences of testing.

**Literature Review**

The adolescent reading model (Deshler & Hock, 2006), which encompasses the simple view of reading (Gough & Tunmer, 1986; Hoover & Gough, 1990) and construction integration theory (Kintsch, 1994), was the framework used to support the development of ACE. ACE is a web-based application for middle grades students designed to assess reading comprehension. The adolescent reading model (Figure 1) is comprised of three interdependent components that contribute to reading comprehension. First, five key skills for efficient word recognition are highlighted. The second component of the model describes the language comprehension skills necessary to make meaning, such as background knowledge and text structure. Finally, executive processes are included as part of the reading comprehension process. Word recognition, language comprehension, and executive processes.

![ Adolescent Reading Model](image)

**Figure 1.** Adolescent reading model (based on Deshler & Hock, 2006; adolescent reading model; Gough & Tunmer, 1986; simple view of reading; and Kintsch, 1994; construction integration theory). Permission granted by D. Deshler.
processes work together to allow a reader to understand a text. An effective assessment of reading comprehension for students in the middle grades would take all three components of the adolescent reading model into consideration in both its design and its validation.

**The Need for Middle Grades Assessments**

Middle grades teachers use standardized, diagnostic, or curriculum-based measurements to assess student reading comprehension. Commonly used standardized group measures include state assessments designed to address Every Student Succeeds Act requirements or online computer-adapted measures, such as the Group Reading Assessment and Diagnostic Evaluation from Pearson, Measures of Academic Progress from NWEA, or STAR 360 from Renaissance. These group-administered tests include direct measures of student reading comprehension, but they are designed to collect data only once, twice, or thrice a year. Further, most commercial test providers acknowledge that their products fail to connect to classroom pedagogy (Mandinach & Gummer, 2013).

According to a report by the Carnegie Foundation, screening and diagnostic reading comprehension assessments validated for the middle grades are “scant” (Morsy et al., 2010). Diagnostic measures can provide insights into an individual student’s strengths and weaknesses in reading comprehension (Sharpe, 2012). These include tests like the Gates MacGinitie and informal reading inventory (IRIs), such as the Burns/Roe information-reading inventory (IRI) (2011). The Gates MacGinitie is a group-administered test that takes approximately 55 min to complete and has two forms; therefore, it can only be administered twice a year (MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2000). IRIs can only be administered to one student at a time and can take anywhere from 15 to 60 min to complete. Diagnostic measures are usually reserved for select students who are either identified by group screening measures or whose performance in class causes concern (Sharpe, 2012).

Curriculum-based assessments include oral reading fluency measures, retellings, IRIs, and teacher-created tests to assess student reading achievement. These can be administered more frequently than standardized tests. However, written retellings and teacher-created tests lack reliability and validity and cannot offer grade-level norms. Oral reading fluency assessments, which are often used for screening and progress monitoring (Baker et al., 2015), have established grade-level benchmarks, but they are indirect measures of reading comprehension. Oral reading fluency is correlated with reading comprehension performance (Fuchs, Fuchs, Hosp, & Jenkins, 2001), but these measures cannot provide teachers with direct evidence of students’ reading comprehension strategies or their individual reading comprehension needs. CBM (Curriculum Based Measurement) reading is an online curriculum-based measure that directly measures reading comprehension and has grade-level benchmarks, but this assessment does not assess beyond sixth grade (EasyCBM Reading, 2018).

**Validating Educational Measures**

With the lack of assessments focusing on adolescent reading comprehension, it is critically important that new, valid assessments be developed to fill this need. Typically, assessment validation studies do not investigate beyond content and internal structure (Beckman, Cook, & Mandrekar, 2005). Relationship to other variables, response processes, and consequences of testing are far less represented in the literature (for exceptions, see Bostic & Sondergeld, 2015; Bostic, Sondergeld, Folger, & Kruse, 2017). To create educational assessments that produce valid and reliable outcomes, it is imperative to use multiple forms of validity evidence (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Gall, Gall, & Borg, 2007). While there are many possible types of validity evidence, the Standards for Educational and Psychological Testing recommend five types of validity evidence: test content, response processes, internal structure, relationship to other variables, and consequences of testing (AERA et al., 2014).

**Test Content.** Test content validity evidence examines the degree to which assessment items (test content) align with the construct (theoretical trait) being measured. Evidence supporting test content validity can be logical or empirical and often comes from subject matter experts evaluating for item to domain alignment (Sireci & Faulkner-Bond, 2014)

**Response Process.** Response process validity evidence assesses the alignment between participant responses or performance and test construct. Cognitive interviews (think-aloud tasks) or focus group interviews with potential typical test takers are
often used to collect information on response process validity to ensure test takers are understanding items and responding to them in ways researchers/test developers intended (Padilla & Benitez, 2014).

**Internal Structure.** Internal structure validity evidence focuses on three main aspects: dimensionality (Is the measure unidimensional or multidimensional?), measurement invariance (Is the test fair and free from systematic bias?), and reliability (Does the measure produce internally consistent or replicable outcomes?) (Rios & Wells, 2014). Traditional classical test theory (CTT) and/or more modern measurement methods, such as item response theory, can be used to assess a test’s internal structure.

**Relationship to Other Variables.** Relationship to other variables validity evidence refers to test outcome association with variables hypothesized to be related (either positively or negatively) (Beckman et al., 2005). Traditional statistical analyses can be conducted to assess strength of relationships (e.g., correlations) or differences in test outcomes by various factors (e.g., independent-samples t-tests or ANOVAs) that are postulated to impact assessment results (e.g., gender, school, race/ethnicity, and special education status).

**Consequences of Testing.** Last, consequences of testing, or consequential validity evidence, examines how test takers are impacted by taking a test or by the results of a test.

**Functional Near-Infrared Spectroscopy**
Functional near-infrared spectroscopy (fNIRS) is an emerging brain-imaging tool that allows researchers to monitor brain activity in everyday environments in a safe, portable, and affordable way. It has been used in language and reading studies, but not for validating educational assessments. Reading is a complex cognitive process that requires the coordination, implementation, and integration of different cognitive abilities via the information processing system, such as working memory attention, perception, executive functions, and long-term memory within a short period of time (Landi, Frost, Mencel, Sandak, & Pugh, 2013; Perfetti & Hart, 2002; Stein, 2003; Stowe et al., 1999). If the origin of brain processing in reading can be unfolded, then the reading abilities and skill acquisition can be effectively monitored, the reasons for failures can be accurately identified, and reading evaluation and enhancement procedures can be efficiently developed.

With recent advances in neuroscience, the reading process has been studied by employing established brain-imaging technologies (e.g., functional magnetic resonance, elektroencephalography, positron emission tomography, and magnetoecephalography). These studies identified at least three main regions involved in reading, all primarily in the left hemisphere: inferior frontal, temporal, and posterior-parietal regions (Landi et al., 2013; Pugh et al., 2000; Shaywitz et al., 2004). Most of these existing neuroimaging results were based on word processing and not long, complex, connected text, whereas the ultimate goal of reading is comprehension of connected text. Fewer studies using the aforementioned neuroimaging modalities have examined the brain areas involved in comprehension of sentence level and longer connected texts as compared to word processing due to mainly technological limitations (Cutting et al., 2006; Gernsbacher & Kaschak, 2003; Robertson et al., 2000). These studies have found that the regions involved in sentence and connected text are similar to those in single word processing, but with greater activations and additional involvement of the right hemisphere and more of the prefrontal regions. This could possibly be due to the increased need for semantic processing and higher-level cognitive processing in maintaining text meaning and drawing inferences in connected text processing (Landi et al., 2013).

Researchers have used fNIRS to assess several types of brain function including motor and visual activation, auditory stimulation, and performance of various cognitive tasks targeting domains of attention, memory, and executive function (Hoshi, 2005, 2007; Irani, Platek, Bunce, Ruocco, & Chute, 2007; Izzetoglu, Bunce, Izzetoglu, Onaral, & Pourrezaei, 2007; Izzetoglu, Bunce, Onaral, Pourrezaei, & Chance, 2004; Izzetoglu et al., 2005; Rolfe, 2000; Strangman, Boas, & Sutton, 2002; Wolf et al., 2008). Several language-related studies used fNIRS technology to specifically focus on the involvement of the frontal lobe in different aspects of language mainly at the word processing level (Hofmann et al., 2014; Jasinska & Petitto, 2014; Minagawa-Kawai et al., 2009; Quaresima, Bisconti, & Ferrari, 2012; Sela, Izzetoglu, Izzetoglu, & Onaral, 2014; Yamamoto, Mashima, & Hiroyasu, 2018). However, to date, no study has taken full advantage of fNIRS for the evaluation of reading comprehension under natural conditions for the validation of educational assessments. fNIRS may prove to be an additional, effective tool in the process of validating an assessment that is quick, easy to use,
and robustly validated; which would be extremely effective in aiding teachers in making instructional decisions for students.

Method

Participants

A convenience sample of 33 eighth-grade students from four different schools participated in this study. School A was a private school in a suburban area \((n = 10, 30.3\%)\), School B was a private school in a suburban area \((n = 7, 21.2\%)\), School C was a public school in an urban area \((n = 12, 36.4\%)\), and School G \((n = 4, 12.1\%)\) was a public school in a suburban area. All eighth-grade students from the schools were invited to participate in the pilot study, but only those students whose parents consented in writing and signed an assent form were selected as participants. As depicted in Table 1, the sample included students with (Individualized Education Program) IEPs and a range of reading levels that roughly corresponded with National Assessment of Educational Progress (NAEP) results for eighth-grade students across the nation. Approximately 69% of the students in the sample scored at the proficient and basic levels \((69\%)\) compared to 72% at the proficient and basic levels in the 2017 NAEP study (McFarland et al., 2018). Of the 33 students in the sample, 5 participated in completing the think-aloud procedure for qualitative analysis. One student was from School A and the remaining four students were from School D. A total of seven students from School A and School C participated in the study using the fNIRS for quantitative analysis.

Table 1.
Demographic data on student participants

| Gender       | Male | Female |
|--------------|------|--------|
| School setting | 67%  | 33%    |
|              | Urban | 36%    |
|              | Suburban | 64%  |
| Reading level | 15%  |        |
|              | Advanced | 33%  |
|              | Proficient | 36%  |
|              | Basic | 3%     |
|              | Below basic | 12%  |
|              | Unknown |        |
| IEP          | 42%  |        |
|              | Without IEP |      |
|              | With IEP | 58%   |

Participants for the think-aloud procedure and the fNIRS procedure were selected based on assent and consent forms. The team needed specific consent and assent for the additional procedures.

Instrumentation

The assessment described in this study was an eighth-grade level passage at a 1070 Lexile level. Lexile measures incorporate two of the three components of the adolescent reading model – word recognition and language comprehension. The passage in this study was an original, informational text with a total of 11 questions. Questions were written to align with the eighth-grade CCSS ELA Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers [NGACB & CCSSO], 2010a). These multiple-choice questions included two literal questions, two inference questions, one vocabulary question, a summary question, a best evidence question, a key idea question, an author’s point of view question, a text structure question, and an author’s purpose question. Each question had four possible answers, which were assigned point values of three, two, one, or zero. The correct answer for each was worth three points. The three distractors for each of the comprehension questions were written to a specific set of constructs that reflected the common errors students make while reading passages or trying to answer questions about what they read. Teachers could then make instructional decisions based on patterns in the data regarding which types of questions students are answering incorrectly, when and how often the student went back to the passage to answer the question, and how much time the student is spending on each question.

Data Sources and Analysis for Validity Evidence

Five types of validity evidence were used in the study: test content, response process, internal structure, relationship to other variables, and consequences of testing. Table 2 summarizes the alignment between validity type, instrumentation, sample, and data analysis.

Expert Panel. An expert panel \((n = 3)\) evaluated the text complexity, Lexile level, and grade level of the passage. This particular passage was an eighth-grade-level informational text on a science topic that was descriptive in nature. The panel was trained by a psychometrician in developing test questions and responses. The panel hired a professional writer to develop this informational text and trained the writer on
vocabulary levels, question types for an eighth-grade level, and the developed heuristic responses for the distractors (incorrect answers). Once the writer completed the passage and questions, the panel reviewed the passage for content, vocabulary level, and grammatical errors. The writer used the Lexile Analyzer® to determine Lexile level and Flesch Kincaid in Microsoft Word to determine grade level (Metametrics, 2018). The expert panel also used the same procedures to check the reading levels. The iterative process included a review by one of the experts to include suggested changes and then a review by the panel for consensus on the changes. The expert panel also used the Text Complexity rubric from the 2010 CCSS to determine the complexity of the passage. Rubric elements included level of purpose, structure, language conventionality, and knowledge demands – content/discipline knowledge. Consensus was achieved through discussion and agreement by at least two of the panel members. For this passage, there was consensus on all items.

**Think Alouds.** Think alouds (n = 5) were used to understand the students’ metacognitive process while reading and answering questions and to collect response process validity data. In order to investigate the efficacy of the constructs, the researchers used a think-aloud protocol to obtain data on the ways in which the students selected their answers to the reading comprehension questions. In a think-aloud protocol, subjects verbally report on their thinking as they complete a task (Pressley & Aflerbach, 1995). Think-aloud protocol data are intended to help researchers achieve a better understanding of students’ thinking or reasoning while they are completing a task (Ying, 2009). Think alouds have long been used to access invisible processes such as reading comprehension (Israel, 2015). This research used a concurrent report protocol because the students reported their thinking as they answered questions on the passage. These were Level 1, or direct verbalizations of thinking, and Level 2, or encoded verbalizations, which include both current thinking and explanations of information in short-term and long-term memory (Ericsson & Simon, 1993).

Before the students read the passage, the researcher used a script to introduce the students to the think-aloud process and to provide rehearsal. Since “directions provided to subjects can color their self-reports” (Pressley & Aflerbach, 1995, p. 121), mathematics practice questions were used instead of reading comprehension questions. The script began with the researcher modeling the process using the question, “What is 10 squared?” and offering four multiple-choice answers. Then the students practiced with another math problem, “What is the average of 10, 15, and 5?” At the conclusion of this practice, the researcher read aloud the instructions adapted from Cordon and Day (1996).

Like I said earlier, today you will be taking a practice reading comprehension test. On this
As the students completed the protocol, the researcher encouraged them to explain their thinking by asking questions such as: How do you know that is the correct answer? How do you know that is not the right answer? Tell me more about how you decided that.

Each of the protocols was digitally recorded and professionally transcribed. The think-aloud protocols were analyzed using a deductive coding scheme. To validate the constructs, closed codes mirrored the previously developed informational text selection multiple-choice constructs, which had been used to write each of the questions, the keys, and the distractors. The researcher created a data collection sheet for each question type. The table for literal comprehension questions is shown in Table 3. The researcher read each transcript and looked for verbal evidence that the students’ reasons for selecting an answer or not selecting an answer aligned with the intended distractor. The first column contains the key and the numbered distractors. The researcher placed a check in the second column if the students’ verbal responses aligned with the expected response from column 3. Direct quotes from the transcripts were cut and pasted into the fourth column on the table to substantiate the check in column 2.

Rasch Psychometrics. Rasch (1980) measurement is considered by many to be a highly effective approach for assessment creation, refinement, and validation (Bond & Fox, 2007; Boone, Townsend, & Staver, 2010; Liu, 2010; Smith, Conrad, Chang, & Piazza, 2002; Waugh & Chapman, 2005; Wright, 1996). “Rasch models are mathematical models that require unidimensionality and result in additivity” (Smith et al., 2002, p. 190). Regardless of the measurement model or theory being used, the specification of unidimensionality is a strict theoretical underpinning with Rasch methods. When data fit the Rasch model specifications, raw scores are converted into logits (logarithm of odds), or equal interval units of measurement, and form a conjoint measurement scale between item difficulty and person ability. This allows for item difficulty and person ability to be “estimated together in such a way that they are freed from the distributional properties of the incidental parameter” (Waugh & Chapman, 2005, p. 81). Unlike CTT, Rasch indices are considered item and sample independent within standard error bands. This means that regardless of the items selected or sample chosen for an assessment, results are comparable across various assessment forms and samples (Bond & Fox, 2007). Further, missing data do not pose a problem when using Rasch measurement because of the probabilistic nature of the model. Well-constructed measurements distinguish between ability levels of a person who correctly answers only the five most difficult items on an assessment and a person who correctly answers only the five easiest items on the same assessment. Rasch measurement allows for the person answering the more difficult items correctly to

| Choice for literal questions | X if confirmed | X if confirmed | Answer construct | Evidence |
|-----------------------------|---------------|---------------|-----------------|----------|
| Key                         |               |               | Correct Answer  |          |
| Distractor 1                |               |               | Text-based literal fact, but with incomplete information/somewhat related to the question |
| Distractor 2                |               |               | Text-based literal fact not related to the question |
| Distractor 3                |               |               | Common background knowledge not in text |

Table 3.
Think-aloud protocol analysis
be rated higher than the person answering the easier items correctly by probabilistically estimating their measure in relation to the item difficulty of their correct responses. This is a great benefit over CTT that would instead assign both individuals a raw score of five because all items are weighted the same with this method.

For this study, Rasch psychometric analyses informed internal structure validity evidence. Dimensionality of the construct was assessed with item fit indices, Rasch principal components analysis (PCA), and mean item difficulty comparison with mean student ability. Reliability was evaluated with Rasch reliability and separation indices.

**fNIRS.** We used fNIRS to determine if questions in the assessment were text based. During pilot testing, seven students (n = 7) wore an fNIRS device while performing the assessment. The device is a flexible band that measures the amount of oxygen in the frontal lobe to monitor prefrontal cortex activity responsible for attention, working memory, decision-making, and executive functioning. fNIRS measures changes in blood oxygenation continuously as a response to cognitive activity similar to functional magnetic resonance imaging (fMRI), but in a portable, affordable, easy to apply and use manner that is less prone to movement and muscle artifacts. Technology is based on optical methods where certain wavelengths of light are shone on the skin to measure the amount of light higher than the levels that can be handled by the detectors) and dark current levels (not enough light is received by the detectors), and those levels were removed from the analysis. Next, the intensity measurements were filtered with a finite impulse response low-pass filter having a cut-off frequency of 0.1 Hz to remove any high-frequency noise from them (see Izzetoglu et al., 2007). Once the preprocessed intensity measurements were obtained, they were converted into changes in oxygenated- and deoxygenated-hemoglobin relative to the pre-task global resting baseline region using modified Beer–Lambert law (see Izzetoglu et al., 2007). In the ACE evaluation, the focus was on oxygenated-hemoglobin that is more directly related to oxygenation changes in the brain due to cognitive activity. For each of the 16 channels covering the forehead, data epochs were extracted from oxygenated-hemoglobin measurements from the start of the question until the last answer was received for each question. Each data epoch was normalized (baseline corrected) according to the 5 s of data immediately prior to the question. Average value of the data epochs was found and used for comparison purposes between question (e.g., factual, inference, and main idea) and answer (correct vs. incorrect) types.

**Traditional Inferential Statistics.** Traditional statistical techniques were implemented to better understand how quantitative data from students completing the ACE assessment informed relationships to other variables validity evidence. Three variables were used to compare students in terms of their ACE assessment total score: school of attendance (categorical – A, B, C, D), total read time (continuous – measured in seconds), and times back to passage (continuous – measured by frequency of times a student looked back at passage). A one-way ANOVA was used to compare students’ ACE assessment total scores by school attended. A one-way ANOVA was an appropriate test to run in this situation because the dependent variable was continuous and the independent variable was categorical with three levels. Pearson correlation analysis was used to examine the relationships between ACE assessment total scores and total read time and times back to passage. Pearson correlations were the appropriate statistical tests to use in this case because all variables being examined were continuous.

**Field Notes**

Members of the research team were present with the students using the ACE. Field notes were taken during group administration of the ACE, the
think alouds, and fNIRS procedures. The field notes were used to identify comments related to what students did and did not like, responses based on passage content, and difficulty or ease of use of ACE.

**Iterative Process of Analysis**

While data analysis methods are presented in Table 1 and explained above in an independent and somewhat linear fashion, the research team engaged in an iterative research process. This process consisted of collecting various types of data, analyzing each data source independently, and then comparing multiple sources of evidence to revise and refine assessment items based on a more holistic understanding of the validity evidence rather than one piece alone. For example, Rasch psychometric analysis findings were compared with fNIRS oxygenation outcomes and student think-aloud results to better inform item acceptability or modification.

**Results**

**Test Content**

An expert panel reviewed the passage to be sure it aligned with an eighth-grade level. The Lexile level for the passage was 1070, which is in the new Lexile bands (955–1155) for sixth through eighth grade (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010a), and a Flesch Kincaid analysis placed the passage at an eighth grade level. A panel of three trained professionals reviewed the passage and questions using the CCSS “Standards Approach to Text Complexity” (NGACBP & CCSSO, 2010b). They determined the passage had a relatively clear, dual purpose that described the hawksbill sea turtle and the recent discovery that it is biofluorescent. The text structure was moderately complex, reading similarly to a narrative; however, it changed from explicitly describing the hawksbill sea turtle to explaining how it was discovered to be biofluorescent. The language used in the passage is mostly literal, but includes domain-specific vocabulary and possibly unfamiliar terms. The panel qualitatively rated the passage as more complex. The panel also reviewed the questions and responses to ensure that the questions and distractors followed the constructs developed specifically for ACE on information texts. Two of the experts worked together to review the questions and distractors, and some questions and distractors were changed prior to student use.

**Think Alouds**

Overall, the students’ responses aligned with the intended answer constructs 41% of the time. For the two literal questions, the students provided verbal data that supported the intended constructs 58% of the time. For the three inferential questions, the students’ statements supported the constructs 48% of the time. Table 4 displays the results of this analysis for each question type. Based on these data, the constructs for the vocabulary question and the key idea questions were revised. For example, the vocabulary distractors no longer utilize morphology and syntax. Instead, the new distractors use semantics, such as weak synonyms and homonyms. In light of triangulation with the statistical data, the main idea construct was not revised. Students who incorrectly answered that main idea question most often selected the author’s purpose instead, which is a common error and one that teachers can address through direct instruction.

**Internal Structure**

Three aspects of the assessment were investigated to assess internal structure: dimensionality, measurement invariance, and reliability (Rios & Wells, 2014).

**Dimensionality.** Dimensionality analysis is an essential requirement in the validation of educational assessments because it helps to express that the outcomes resulting from the administration of an
instrument are effective measures that represent the single, desired construct. Determining an assessment to be unidimensional or multidimensional, however, is not an either/or question. Rather, constructs are seen as being more or less unidimensional and, thus, need to be evaluated on a continuum. Because there is no singular most appropriate method for evaluating dimensionality, the researchers used a variety of Rasch-based methods (i.e., item fit, Rasch PCA, and mean item difficulty comparison with mean student ability). All items performed within acceptable psychometric ranges to suggest a unidimensional construct. Rasch in fit and outfit statistics were appropriate (ZSTD between −2.0 and 2.0 and mean squares between 0.5 and 1.5 logits), and no items possessed a negative point-biserial statistic.

A Rasch Principal Components Analysis (PCA) was also conducted. The strongest evidence of unidimensionality is uncovered when the items on an instrument predict more than 60% of the score variance associated with the instrument. In the case of the current assessment, 41.2% of the variance was predicted. This suggests that there may be additional concepts not covered in the current instrument that are influencing the student reading comprehension ability being assessed. Furthermore, the amount of unaccounted variance present in the first contrast (12.5%) may be indicative of a secondary underlying dimension (Linacre, 2006), but this is only accounted for by less than three items, making it difficult to determine the potential second dimension’s meaning.

A comparison of mean item difficulty ($M = 0.0$ logits, $SEM = 0.49$ logits) to mean student ability ($M = 0.15$, $SEM = 0.81$) shows that the assessment was appropriate for the students completing the test. Any mean student ability value within ±2 SEM of the mean item difficulty is considered an appropriate value. Overall, combined psychometric findings suggest that the ACE assessment functioned reasonably well in terms of Rasch unidimensionality requirements.

**Reliability.** Rasch item reliability and separation were both examined to determine the internal consistency of measures. Rasch reliability and separation of 0.90 and 3.00, respectively, are excellent; 0.80 and 2.00, respectively, are good; 0.70 and 1.50, respectively, are acceptable (Duncan, Bode, Lai, & Perera, 2003). The ACE assessment had excellent item reliability (0.91) and separation (3.23).

**Functional Near-Infrared Spectroscopy.** The fNIRS data were used to ensure that the questions for the passage were text dependent because the CCSS focuses on text-based questions that require students to read and comprehend in order to answer. Other than the inference questions, which do rely on both text and background knowledge, the fNIRS provided data that helped determine for which questions students used the frontal lobe (attending to text) as indicated by higher oxygenation. The researchers reexamined the questions students answered correctly and did not have average or increased oxygenation. If most students answered the question without increased oxygenation, the team examined the think-aloud and psychometric data on each question and determined if the question needed to be rewritten, as it might contain bias or rely completely on background knowledge to answer. Based on the fNIRS data, the vocabulary question showed low oxygenation levels for students that answered correctly and incorrectly, thus representing the possibility that students did not have to read the passage to be able to answer the question.

**Relationship to Other Variables**

Multiple variables of interest (school attending, total read time, times back to passage) were assessed for their relationship to the ACE assessment total score. There were no significant differences in total score regardless of the school the students attended: $F(2, 28) = 0.476$, $p = 0.627$. This was an interesting result as two of the schools were private schools for students with language-based learning disabilities and all students in the study from these schools did have an IEP. The other two schools were public schools in the suburbs. The researchers expected there might be a difference in total score based on school due to the number of students with IEPs at two of the schools; however, it was not determined to be an issue for this study as mean scores for students without an IEP (24.3 or 74%) and students with an IEP (23.2 or 70%) were similar.

Further, there were no significant relationships between total read time and total score or times back to passage and total score (see Table 5 for statistics). Speed-based, indirect measures of reading comprehension, such as oral reading fluency, correlate reading speed and fluency with better reading comprehension (Munger & Blachman, 2013; Salvador, Schoeneberger, Tingle, & Algozzine, 2012), and proficient word reading may free up space in the mind to concentrate on meaning of text (Perfetti, 1985). Thus, the researchers expected that students who read the passage in less time would
have a higher score. The results did not support that assumption. However, this study did not track exact word reading fluency, only the amount of time it took a student to read the passage.

**Consequences of Testing**

Data on the consequences of testing were collected through researcher field notes and observations. The ACE assessment was easy to administer. Students quickly learned to log in and identify the passage to read, and they found it easy to move between the passage and the questions. The average time for students to read and answer questions on one passage was 7 min, which is significantly less than the 30–50 that it is estimated to give a Roe & Burns Informal Reading Inventory (Burns & Roe, 2011) or the 30 min that it is estimated it would take to complete the Easy CBM Reading (EasyCBM Reading: UO DIBELS Data System, 2018). No student expressed any anxiety or stress related to taking the ACE assessment, and field notes contained multiple references to the ease of use of the application. Moreover, this was a low-stakes assessment for students because scores did not impact their grades or standing in class.

The results of this ACE informational passage met many of the aspects of the adolescent reading model. It included the components of language comprehension, and, with the use of the fNIRS during validation, ACE also examined the executive process. For the majority of the students who participated in this study, word recognition was not an issue. For the two students that did have word recognition issues, they would have benefited from listening to the passage, which is a future function of the ACE.

**Discussion**

Current research supports the need for a quick, easy, middle grades reading comprehension assessment tool that can identify students’ strengths or weaknesses (Morsy et al., 2010). The current study found that this ACE passage addressed this need by producing a valid assessment of eighth-grade student reading comprehension. This is the only one passage in ACE, which currently contains 10 narrative and 10 informational texts and question sets for sixth, seventh, and eighth grades. This study procedure will be followed for each of the remaining passages in ACE in order to create a valid and reliable assessment tool.

The adolescent reading model holds that language and cognitive processes – such as word recognition, language comprehension, and executive processes – work together to support reading comprehension (Deshler & Hock, 2006). Assessments that only look at one of these variables will not be able to capture a true measure of a student’s reading comprehension skills. ACE is designed to address word identification through leveled passages and language comprehension through multiple-choice reading comprehension questions, and it does this effectively as demonstrated through Rasch measurement findings (internal structure validity evidence). Further, by using fNIRS to assess student executive processing while taking the assessment, this research has shown that the ACE reading comprehension questions are text dependent and require students to engage their short-term memory while answering the questions (internal structure validity evidence).

ACE also aims to provide teachers with actionable, instructional information about the kinds of errors a student makes when he or she does not get a reading comprehension question correct, which few reading comprehension assessments can do (Mandinach & Gummer, 2013). It does this by providing data by student and class. This would include, as an example, which students are having difficulty answering main idea questions. ACE also provides progress-monitoring data if the student is improving his ability to get closer to the correct response. By using the fNIRS in the validation process, developers are able to remove questions that are answered without using the prefrontal cortex and ensure that the questions on the assessment are text based. Therefore, it is assumed that students are not relying on background

|                      | n   | M (SD)          | Correlation with total score |
|----------------------|-----|-----------------|------------------------------|
| Total read time      | 33  | 711.09 (415.87) | 0.305                        |
| Times back to passage| 33  | 4.15 (3.78)     | 0.260                        |
knowledge to answer the questions. ACE assesses comprehension in a non-threatening way for students as noted by their ease in completing this low-stakes assessment without noted stress or anxiety (consequences of testing validity evidence). This is largely accomplished through appropriate text complexity and constructs that were used to build each of the incorrect answers for the multiple-choice questions on the assessment (content validity evidence). The think-aloud protocols conducted as part of the validation of the ACE assessment suggest that, for most of the question types, the distractors function as intended within the assessment (response process validity evidence). Additionally, there did not appear to be any biases in ACE results based on some commonly hypothesized indicators of potential bias (school type, total reading type, times back to passage, relationship to other variables, and validity evidence). This type of information is important for middle grades teachers who are responsible for ensuring that all students can read with accuracy and, more importantly, with understanding.

**Holistic Validation Approach**

The validation process outlined in this research is holistic, robust, and aligns well with the Standards for Educational and Psychological Testing set forth by AERA, APA, and NCME (2014). Multiple forms of validity evidence have been investigated, and the research team found that these different forms of validity evidence actually provided the researchers with insights into specific ACE questions that would not have been uncovered using psychometric means of validation alone. For example, Question 6 was a vocabulary question that Rasch measurement methods regarded as fitting well. It met appropriate psychometric parameters regarding its overall functioning with the other items on the assessment, and its distractors were working as would be expected. However, when the think-aloud protocols were analyzed, it became apparent that student responses did not align with the answer constructs. This meant that teachers would not be able to use the constructs to analyze student errors to plan instruction or form flexible groups. Instead, the researchers used the results of the deductive coding to analyze the students’ responses during the think alouds for that particular question and rewrote the constructs in order to reflect the decision-making process and error patterns that they engaged in when deciding which answer choice best defined the vocabulary word. In addition, the oxygenation-level data from the fNIRS supported the data from the think alouds. Using psychometric analysis (internal structure validity evidence) alone would not have illuminated this important discrepancy, which required response processing validity evidence to uncover.

There were also incidents in which multiple types of validity evidence aligned well in terms of assessing specific ACE items and facilitating actions for improving ACE items. For instance, psychometric analysis indicated that Question 7 worked well with the overall construct of ACE items (internal structure validity evidence), but it was performing poorly in terms of differentiating who knew the content and who did not. The item was noted as too difficult because only one student (the most capable) was able to answer the item correctly with 44% of students who were considered “more able” selecting distractor “B,” and 52% of students who were considered “less able” selecting distractor “C.” When looking at the item more closely, the research team saw that the item asked students to identity the main idea of the passage. Think-aloud data showed that student descriptions of their decision-making process only supported the construct 35% of the time, and this was largely due to the students’ inability to distinguish between the main idea and author’s purpose. fNIRS data showed increased oxygenation for Question 7, which demonstrated students attended to the text; however, no student wearing the fNIRS while taking the assessment answered Question 7 correctly. While students were attending to the text, they were not able to answer it correctly, supporting both psychometric and think-aloud data.

As a result of the multiple sets of data, the vocabulary question and the heuristics for the distractors for the vocabulary question were changed. The team decided to maintain the main idea question, as it was more confusion between the main idea and author’s purpose rather than an inappropriate question. This information should be shared with teachers in order to affect change in classroom instruction.

The Standards for Educational and Psychological Testing (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014) defined expectations related to educational assessment design, implementation, scoring, and reporting with a focus on the critical importance of implementing multiple forms of validity evidence when developing educational assessments. Through this study, the researchers demonstrated that evaluating multiple
types of assessment validity evidence allowed for better informed conclusions to be drawn about how the ACE functions as a whole and specific item along with their distractors. This holistic approach resulted in more robust and comprehensive conclusions about ACE in terms of its validity and reliability for producing appropriate reading comprehension assessment results for middle grades students. However, it is imperative to note that conducting this type of validation research is an extremely time-intensive process comprised of iterative phases. A single researcher could not undertake this type of research alone. Rather, it requires a team of researchers with expertise in content, psychometrics, theory, and data analysis. Perhaps this is, in part, why such extensive validation studies focusing on all recommended validity evidence components are not regularly conducted.

**Limitations**

As with all research, this study is subject to limitations. Although the sample size was relatively small, minimum sample size requirements were met to conduct all of the quantitative and qualitative analyses and interpret results with appropriate confidence in this validation study. Additionally, while the sample of students for this study was drawn from diverse socioeconomic, racial, gender, educational, and special education eligibility groups, participants all came from the northeast, limiting geographical generalizability. Finally, though the ACE has multiple passages developed at varying levels of complexity, this study only focused on results from one of these passages. Thus, generalizability of findings across ACE passages cannot be assumed, and other passages have been or are being assessed using a similar process.

**Conclusion**

ACE provides a direct measure of reading comprehension by using original passages leveled by both qualitative and quantitative measures and multiple-choice questions aligned with current standards for English Language Art (NGACBP & CCSSO, 2010b). The robust, holistic validation process described in this paper demonstrated that this passage from the ACE assessment is reliable and valid as a direct measure of reading comprehension. Statistical analysis showed that this passage from ACE provides validity evidence for measuring reading comprehension. fNIRS data demonstrate that the reading comprehension questions in ACE are text dependent. Students who correctly answered questions were using their prefrontal cortex to do so, which is, where new information is temporarily stored, as opposed to background information, which would not engage the prefrontal cortex to the same degree that was evidenced in this analysis.

The ACE assessment is currently being piloted in sixth-through eighth-grade classrooms, and the research team will also begin developing the ACE to include ninth-through twelfth-grade passages. In addition, a teacher dashboard will provide immediate data to teachers, but the researchers have not yet piloted the teacher dashboard with classroom teachers. This is an area that needs to be developed. Also, in working with the eighth-grade students in this pilot validation study, some of the students offered suggestions to improve ACE. Students would like the ability to change the color of the background and the color of the text. There has been some research to support white text on a darker background, and the researchers intend to explore this research. Students also requested to have audio to accompany the text. These requests came from students who continue to struggle with decoding, but not with comprehension. The researchers are considering including a listening comprehension component to the assessment.

**Acknowledgments**

The ACE assessment in this study is patent pending.

**Funding**

This study was made possible through a grant provided by Drexel Ventures.

**References**

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

Baker, D. L., Biancarosa, G., Park, B. J., Busselot, T., Smith, J. L., Baker, S. K., . . . Tindal, G. (2015). Validity of CBM measures of oral reading fluency and reading comprehension on high-stakes reading assessments in grades 7 and 8. *Reading and Writing, 28*(1), 57–104. doi:10.1007/s11145-014-9505-4
Beckman, C., Cook, Mandrekar, D. A., & Mandrekar, J. N. (2005). What is the validity evidence for assessments of clinical teaching? *Journal of General Internal Medicine, 20*(2), 1159–1164. doi:10.1111/j.1525-1497.2005.0258.x

Biancarosa, G., & Snow, C. E. (2004). Reading next: A vision for action and research in middle and high school literacy: A report to Carnegie Corporation of New York. Washington, DC: Alliance for Excellence in Education.

Bond, T., & Fox, C. (2007). *Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Boone, W. J., Townsend, J. S., & Staver, J. (2010). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education, 95*(2), 258–280. doi:10.1002/sec.20413

Bostic, J. D., & Sondergeld, T. A. (2015). Measuring sixth-grade students’ problem-solving: Validating an instrument addressing the mathematics common core. *School Science and Mathematics Journal, 115*(6), 281–291. doi:10.1111/ssm.2015.115.issue-6

Bostic, J. D., Sondergeld, T. A., Folger, T., & Kruse, L. (2017). PSM7 and PSM8: Validating two problem-solving measures. *Journal of Applied Measurement, 18*(2), 1–12.

Burns, P. C., & Roe, B. D. (2011). *Burns/Roe informal reading inventory: Preprimer to twelfth grade*. Boston, MA: Houghton Mifflin.

Cordon, L. A., & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology, 88*(2), 288. doi:10.1037/0022-0663.88.2.288

Cutting, L. E., Clements, A. M., Courtney, S., Rimrodt, S. I., Schafer, J. G., Bisesi, J., & Pugh, K. R. (2006). Differential components of sentence comprehension: Beyond single word reading and memory. *NeuroImage, 29*(2), 429–438. doi:10.1016/j.neuroimage.2005.07.057

Deshler, D. D., & Hock, M. F. (2006). *Shaping literacy achievement*. New York, NY: Guilford Press.

Duncan, P., Bode, R., Lai, S., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation, 84*, 950–963.

EasyCBM Reading: UO DIBELS Data System. (2018). Retrieved from http://dibels.uoregon.edu/assessment/reading.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis*. Cambridge, MA: MIT press.

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*(3), 239–256. doi:10.1207/S1532799XSSR0503_3

Gall, M., Gall, J., & Borg, W. (2007). *Educational research: An introduction* (8th ed.). Boston, MA: Pearson.

Gernsbacher, M. A., & Kaschak, M. P. (2003). Neuroimaging studies of language production and comprehension. *Annual Review of Psychology, 54*(1), 91–114. doi:10.1146/annurev.psych.54.101601.145128

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

Hofmann, M. J., Dambacher, M., Jacobs, A. M., Kliegl, R., Radach, R., Kuchinke, L., & Herrmann, M. J. (2014). Occipital and orbitofrontal hemodynamics during naturally paced reading: An fNIRS study. *NeuroImage, 94*, 193–202. doi:10.1016/j.neuroimage.2014.03.014

Hoover, W., & Gough, P. (1990). The simple view of reading. *Reading and Writing: an Interdisciplinary Journal, 2*, 127–160. doi:10.1007/BF00401799

Hoshi, Y. (2005). Functional near-infrared spectroscopy: Potential and limitations in neuroimaging studies. *International Review of Neurobiology Neuroimaging. Part A*, 237–266.

Hoshi, Y. (2007). Functional near-infrared spectroscopy: Current status and future prospects. *Journal of Biomedical Optics, 12*(6), 062106. doi:10.1117/1.2804911

Irani, F., Platek, S. M., Bunce, S., Ruocco, A. C., & Chute, D. (2007). Functional near infrared spectroscopy (fNIRS): An emerging neuroimaging technology with important applications for the study of brain disorders. *The Clinical Neuropsychologist, 21*(1), 9–37. doi:10.1080/13854040600611392

Israel, S. E. (2015). *Verbal protocols in literacy research: Nature of global reading development*. New York, NY: Routledge.

Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., & Chance, B. (2004). Functional optical brain imaging using near-infrared during cognitive tasks. *International Journal of Human-Computer Interaction, 17*(2), 211–227. doi:10.1207/s15327590ijhci1702_6
Izzetoglu, M., Bunce, S. C., Izzetoglu, K., Onaral, B., & Pourrezaei, K. (2007). Functional brain imaging using near-infrared technology. *IEEE Engineering in Medicine and Biology Magazine*, 26(4), 38. doi:10.1109/MEMB.2007.384094

Izzetoglu, M., Izzetoglu, K., Bunce, S., Ayaz, H., Devaraj, A., Onaral, B., & Pourrezaei, K. (2005). Functional near-infrared neuroimaging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(2), 153–159. doi:10.1109/TNSRE.2005.847377

Jasińska, K. K., & Petitto, L. A. (2014). Development of neural systems for reading in the monolingual and bilingual brain: New insights from functional near infrared spectroscopy neuroimaging. *Developmental Neuropsychology*, 39(6), 421–439. doi:10.1080/87565641.2014.939180

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, 49(4), 294–303. doi:10.1037/0003-066X.49.4.294

Landi, N., Frost, S. J., Mencel, W. E., Sandak, R., & Pugh, K. R. (2013). Neurobiological bases of reading comprehension: Insights from neuroimaging studies of word-level and text-level processing in skilled and impaired readers. *Reading & Writing Quarterly*, 29(2), 145–167. doi:10.1080/10573569.2013.758566

Linacre, J. M. (2006). Data variance explained by measures. *Rasch Measurement Transactions*, 20, 1045–1047.

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch modeling approach*. Charlotte, NC: Information Age.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2000). *Gates-MacGinitie reading tests, fourth edition* (GRMT-4). Itasca, IL: Riverside.

Mandinach, E. B., & Guummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher*, 42(1), 30–37. doi:10.3102/0013189X12459803

McFarland, J., Hussar, B., Wang, X., Zhang, J., Wang, K., Rathbun, A., . . . Bullock Mann, F. (2018). *The condition of education 2018*. National Center for Educational Statistics, Washington, DC.

Metametrics, I. (2018). *The Lexile Analyzer® (computer software)*. The Lexile Framework for Reading.

Minagawa-Kawai, Y., Naoi, N., Kikuchi, N., Yamamoto, J., Nakamura, K., & Kojima, S. (2009). Cerebral laterality for phonemic and prosodic cue decoding in children with autism. *NeuroReport*, 20(13), 1219–1224. doi:10.1097/WNR.0b013e32832ba65f

Morsy, L., Kieffer, M., & Snow, C. E. (2010). *Measure for measure: A critical consumers’ guide to reading comprehension assessments for adolescents*. New York, NY: Carnegie Corporation of New York.

Munger, K. A., & Blachman, B. A. (2013). Taking a ‘simple view’ of the dynamic indicators of basic early literacy skills as a predictor of multiple measures of third-grade reading comprehension. *Psychology in the Schools*, 50(7), 722–737. doi:10.1002/pits.2013.50.issue-7

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010a). *Common core state standards for English language arts*. Washington, DC: Authors.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010b). *Appendix A*. Washington, DC: Authors.

Padilla, J. L., & Benitez, I. (2014). Validity evidence based on response process. *Psicothema*, 26(1), 136–144.

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

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). Amsterdam, NL: John Benjamins.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. New York, NY: Routledge.

Pugh, K. R., Mencel, W. E., Jenner, A. R., Katz, L., Frost, S. J., Lee, J. R., & Shaywitz, B. A. (2000). Functional neuroimaging studies of reading and reading disability (developmental dyslexia). *Mental Retardation and Developmental Disabilities Research Reviews*, 6(3), 207–213. doi:10.1002/1098-2779(2000)6:3<207::AID-MRDD8>3.0.CO;2-P

Quaresima, V., Bisconti, S., & Ferrari, M. (2012). A brief review on the use of functional near-infrared spectroscopy (fNIRS) for language imaging studies in human newborns and adults. *Brain and Language*, 121(2), 79–89. doi:10.1016/j.bandl.2011.03.009

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, DK: Danmarks Paedagoiske Institut.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, 26(1), 108–116.
Robertson, D. A., Gernsbacher, M. A., Guidotti, S. J., Robertson, R. R., Irwin, W., Mock, B. J., & Campana, M. E. (2000). Functional neuroanatomy of the cognitive process of mapping during discourse comprehension. *Psychological Science, 11*(3), 255–260. doi:10.1111/1467-9280.00251

Rolfe, P. (2000). In vivo near-infrared spectroscopy. *Annual Review of Biomedical Engineering, 2*(1), 715–754. doi:10.1146/annurev.bioeng.2.1.715

Salvador, S. K., Schoeneberger, J., Tingle, L., & Algozzine, B. (2012). Relationship between second grade oral reading fluency and third grade reading. *Assessment in Education: Principles, Policy & Practice, 19*(3), 341–356. doi:10.1080/0969594X.2011.613368

Sato, H., Yahata, N., Funane, T., Takizawa, R., Katura, T., Atsumori, H., & Fukuda, M. (2013). A NIRS–FMRI investigation of prefrontal cortex activity during a working memory task. *Neuroimage, 83*, 158–173. doi:10.1016/j.neuroimage.2013.06.043

Sela, I., Izzetoglu, M., Izzetoglu, K., & Onaral, B. (2014). A functional near-infrared spectroscopy study of lexical decision task supports the dual route model and the phonological deficit theory of dyslexia. *Journal of Learning Disabilities, 47*(3), 279–288. doi:10.1177/0022219412451998

Sharpe, C. (2012). *Secondary assessments: Universal screening, diagnostic, & progress monitoring*. Middletown, CT: SERC.

Shaywitz, B. A., Shaywitz, S. E., Blachman, B. A., Pugh, K. R., Fulbright, R. K., Skudlarski, P., & Gore, J. C. (2004). Development of left occipito-temporal systems for skilled reading in children after a phonologically-based intervention. *Biological Psychiatry, 55*(9), 926–933. doi:10.1016/j.biopsych.2003.12.019

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *PsycLethema, 26*(1), 100–107.

Smith, E., Conrad, K., Chang, K., & Piazza, J. (2002). An introduction to Rasch measurement for scale development and person assessment. *Journal of Nursing Measurement, 10*, 189–206. doi:10.1891/jnum.10.3.189.52562

Stein, J. (2003). Visual motion sensitivity and reading. *Neuropsychologia, 41*(13), 1785–1793. doi:10.1016/S0028-3932(03)00179-9

Stowe, L. A., Paans, A. M., Wijers, A. A., Zwarts, F., Mulder, G., & Vaalburg, W. (1999). Sentence comprehension and word repetition: A positron emission tomography investigation. *Psychophysiology, 36*(6), 786–801. doi:10.1111/1469-8986.3660786

Strangman, G., Boas, D. A., & Sutton, J. P. (2002). Non-invasive neuroimaging using near-infrared light. *Biological Psychiatry, 52*(7), 679–693.

Waugh, R., & Chapman, E. (2005). An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: What is the difference? Which method is better? *Journal of Applied Measurement, 6*, 80–99.

Wijekumar, S., Huppert, T. J., Magnotta, V. A., Buss, A. T., & Spencer, J. P. (2017). Validating an image-based fNIRS approach with fMRI and a working memory task. *NeuroImage, 147*, 204–218. doi:10.1016/j.neuroimage.2016.12.007

Wolf, M., Morren, G., Haensse, D., Karen, T., Wolf, U., Fauchère, J., & Bucher, H. (2008). Near infrared spectroscopy to study the brain: An overview. *Opto-Electronics Review, 16*(4), 412–414. doi:10.2478/s11772-008-0042-z

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3*, 3–24. doi:10.1080/10705519609540026

Yamamoto, U., Mashima, N., & Hiroyasu, T. (2018). Evaluating working memory capacity with functional near-infrared spectroscopy measurement of brain activity. *Journal of Cognitive Enhancement, 2*(3), 217–224. doi:10.1007/s41465-017-0063-y

Ying, Z. (2009). Protocol analysis in the validation of language tests: Potential of the method, state of the evidence. *International Journal of Pedagogies & Learning, 5*(1), 124–137. doi:10.5172/ijpl.5.1.124